TÉCNICO LISBOA

Learning and Decision Making 2017-2018

MSc in Computer Science and Engineering

First test – April 3, 2018

Instructions

- You have 90 minutes to complete the test.
- Make sure that your test has a total of 9 pages and is not missing any sheets, then write your full name and student n. on this page (and all others if you want to be safe).
- The test has a total of 6 questions, with a maximum score of 20 points. The questions have different levels of difficulty. The point value of each question is provided next to the question number.
- *If you get stuck in a question, move on.* You should start with the easier questions to secure those points, before moving on to the harder questions.
- *No interaction with the faculty is allowed during the exam.* If you are unclear about a question, clearly indicate it and answer to the best of your ability.
- Please provide your answer in the space below each question. If you make a mess, clearly indicate your answer.
- The exam is open book and open notes. You may use a calculator, but any other type of electronic or communication equipment is not allowed.
- Good luck.

Question 1. (**3 pts.**)

Consider the following problem. Back in the 16th century, a caravel routinely traveled back and forth between two ports, A and B, transporting products for sale in each port. To go from one port to the other required the caravel to traverse a foggy strait with troubled waters, during which the boat would sometimes lose its bearing and get disoriented; this occurs with probability 0.4.

When disoriented, if the captain chooses to keep moving, the boat ends up returning to the port it just left with a probability 0.4; with a probability 0.4 it will successfully reach the destination port; and with a 0.2 probability, will remain lost in the strait. On the other hand, if the boat is properly oriented and the captain chooses to keep moving, it will reach the destination port with probability 0.7; it will remain oriented in the strait with probability 0.2; with a 0.1 probability it will lose its bearing and get disoriented.

There is, however, a complicated maneuver that, if successful, will put the caravel back in track. If the captain decides to perform such maneuver at any of the ports or if the boat is oriented, then the maneuver has no effect. However, if the boat is disoriented, performing the maneuver will succeed in orienting the boat with a probability 0.7 (and leave it disoriented with probability 0.3).

Finally, because of the fog, when navigating the strait the captain can never be completely certain whether the boat is oriented or not. In particular, if the boat remains oriented after having moved, the captain will perceive the boat as oriented with a probability 0.6, but will feel lost with a probability 0.4. Conversely, if a movement causes the caravel to lose its bearing, the captain will perceive the boat as oriented with a probability 0.4 and will feel lost with a probability 0.6. On the other hand, if the captain decides to perform the complicated maneuver instead of moving, the probability that he correctly assesses whether oriented or not afterwards is 0.8.

The goal of the captain is, of course, to successfully reach the destination port as quickly as possible.

Describe the decision problem of the captain as a POMDP. In particular, indicate the state, action and observation spaces, the transition probabilities, the observation probabilities and the immediate cost function.

Solution 1.

We describe the problem as POMDP $(\mathcal{X}, \mathcal{A}, \mathcal{Z}, \{\mathbf{P}_a\}, \{\mathbf{O}_a\}, c, \gamma)$. We have

- The state space should consider the situation of the caravel. We may be in either one of the two ports, lost in sea, or oriented in sea. As for the ports, we don't really care if the boat is in A or B. We only need to know, with respect to the current trip, whether it is at the *origin* or at the *destination* of the trip. Hence, we consider the states (O)rigin, (D)estination, (L)ost and (G)ood, i.e., $\mathcal{X} = \{O, D, L, G\}$.
- The actions are just to m(o)ve on or to execute the m(a)neuver. As such, we have $\mathcal{A} = \{o, a\}$.
- The observations should consider what the captain is able to perceive. From the description of the problem, it follows that $\mathcal{Z} = \mathcal{X}$, since the captain knows when it is at either port, and is able to perceive (although inaccurately) whether the caravel is lost or oriented. Hence, $\mathcal{Z} = \{O, D, L, G\}$.
- The transition probabilities should express how the situation of the caravel evolves depending on the

actions of the captain and can be inferred from the descriptive text. We have, for action o (move),

$$\mathbf{P}_o = \begin{bmatrix} 0.0 & 0.0 & 0.4 & 0.6 \\ 1.0 & 0.0 & 0.0 & 0.0 \\ 0.4 & 0.4 & 0.2 & 0.0 \\ 0.0 & 0.7 & 0.1 & 0.2 \end{bmatrix},$$

and for action a (maneuver),

$$\mathbf{P}_a = \begin{bmatrix} 1.0 & 0.0 & 0.0 & 0.0 \\ 1.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.3 & 0.7 \\ 0.0 & 0.0 & 0.0 & 1.0 \end{bmatrix}.$$

Note that, after reaching the destination port, the other port (which was previously the origin port) becomes the new destination, and the current port becomes the new origin port. We model this by setting $\mathbf{P}(O|D, \cdot) = 1$, for all actions. Alternatively, for the move action, the following transition probability matrix is also admissible:

$$\mathbf{P}_o = \begin{bmatrix} 0.0 & 0.0 & 0.4 & 0.6 \\ 0.0 & 0.0 & 0.4 & 0.6 \\ 0.4 & 0.4 & 0.2 & 0.0 \\ 0.0 & 0.7 & 0.1 & 0.2 \end{bmatrix}.$$

- The observation probabilities also follow from the description as

$$\mathbf{O}_o = \begin{bmatrix} 1.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 1.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.6 & 0.4 \\ 0.0 & 0.0 & 0.4 & 0.6 \end{bmatrix}, \qquad \mathbf{O}_a = \begin{bmatrix} 1.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 1.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.8 & 0.2 \\ 0.0 & 0.0 & 0.2 & 0.8 \end{bmatrix}.$$

- Finally, the cost should penalize the captain for every moment it takes to get to the destination port, yielding the cost function

$$\mathbf{C} = \begin{bmatrix} 1.0 & 1.0 \\ 0.0 & 0.0 \\ 1.0 & 1.0 \\ 1.0 & 1.0 \end{bmatrix}.$$

Finally, since no information is provided regarding γ, we leave this parameter unspecified.

In the remainder of the test, consider the POMDP $\mathcal{M} = (\mathcal{X}, \mathcal{A}, \mathcal{Z}, \{\mathbf{P}_a\}, \{\mathbf{O}_a\}, c, \gamma)$ where

- $\mathcal{X} = \{1, 2, 3\}$;
- $\mathcal{A} = \{A, B\}$;
- $\mathcal{Z} = \{1, 2, 3\}$;
- The transition probability matrices are:

$$\mathbf{P}_A = \begin{bmatrix} 0.0 & 0.5 & 0.5 \\ 0.0 & 0.0 & 1.0 \\ 1.0 & 0.0 & 0.0 \end{bmatrix}; \qquad \mathbf{P}_B = \begin{bmatrix} 1.0 & 0.0 & 0.0 \\ 0.0 & 1.0 & 0.0 \\ 0.0 & 0.0 & 1.0 \end{bmatrix};$$

- The observation probability matrices are:

$$\mathbf{O}_A = \begin{bmatrix} 1.0 & 0.0 & 0.0 \\ 0.0 & 0.5 & 0.5 \\ 0.0 & 0.5 & 0.5 \end{bmatrix}; \qquad \mathbf{O}_B = \begin{bmatrix} 1.0 & 0.0 & 0.0 \\ 0.0 & 1.0 & 0.0 \\ 0.0 & 0.0 & 1.0 \end{bmatrix};$$

- The cost is given by $c(x, a) = 1 - \mathbb{I}[x = 3]$;
- $\gamma = 0.9$.

Question 2. (**4 pts.**)
Suppose that the POMDP $\mathcal{M}$ departs from the initial state $\mathrm{x}_0 = 1$ and the agent takes the actions $\boldsymbol{a}_{0:2} = \{A, A, A\}$ and makes the observations $\boldsymbol{z}_{1:3} = \{2, 1, 2\}$.

(a) **(2 pts.)** Compute the belief of the agent at time step $t = 3$ using the forward algorithm. Note that there is no initial observation.

(b) **(2 pt.)** What is the most likely sequence of states given the sequence of actions and observations?

Solution 2.

(a) We follow the forward algorithm to compute the belief of the agent at time step $t = 3$, noting that the belief corresponds to the distribution $\boldsymbol{\mu}_{3|0:3}$. Since we depart from $x_0 = 1$, we have

$$\boldsymbol{\mu}_0 = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} = \boldsymbol{\alpha}_0^\top.$$

Then, following the forward algorithm yields

$$\boldsymbol{\alpha}_1^\top = \boldsymbol{\alpha}_0^\top \mathbf{P}_A \operatorname{diag}(\mathbf{O}_A(z_1 \mid \cdot) = \begin{bmatrix} 0 & 0.25 & 0.25 \end{bmatrix},$$

$$\boldsymbol{\alpha}_2^\top = \boldsymbol{\alpha}_1^\top \mathbf{P}_A \operatorname{diag}(\mathbf{O}_A(z_2 \mid \cdot) = \begin{bmatrix} 0.25 & 0 & 0 \end{bmatrix},$$

$$\boldsymbol{\alpha}_3^\top = \boldsymbol{\alpha}_2^\top \mathbf{P}_A \operatorname{diag}(\mathbf{O}_A(z_3 \mid \cdot) = \begin{bmatrix} 0 & 0.063 & 0.063 \end{bmatrix}.$$

Finally, we get

$$\boldsymbol{\mu}_{3|0:3} = \frac{\boldsymbol{\alpha}_3^\top}{\boldsymbol{\alpha}_3^\top \mathbf{1}} = \begin{bmatrix} 0 & 0.5 & 0.5 \end{bmatrix}.$$

(b) To compute the most likely sequence of states, we use the Viterbi algorithm on the provided sequence of observations. In the forward pass, we get

$$\boldsymbol{m}_0^\top = \boldsymbol{\mu}_0 = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix},$$

$$\max_x \{\operatorname{diag}(\boldsymbol{m}_0)\mathbf{P}_A\} = \begin{bmatrix} 0 & 0.5 & 0.5 \end{bmatrix},$$

$$\boldsymbol{i}_1 = \operatorname*{argmax}_x \{\operatorname{diag}(\boldsymbol{m}_0)\mathbf{P}_A\} = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix},$$

$$\boldsymbol{m}_1^\top = \max_x \{\operatorname{diag}(\boldsymbol{m}_0)\mathbf{P}_A\} \operatorname{diag}(\mathbf{O}_A(z_1 \mid \cdot)) = \begin{bmatrix} 0 & 0.25 & 0.25 \end{bmatrix},$$

$$\max_x \{\operatorname{diag}(\boldsymbol{m}_1)\mathbf{P}_A\} = \begin{bmatrix} 0.25 & 0 & 0.25 \end{bmatrix},$$

$$\boldsymbol{i}_2 = \operatorname*{argmax}_x \{\operatorname{diag}(\boldsymbol{m}_1)\mathbf{P}_A\} = \begin{bmatrix} 3 & 1 & 2 \end{bmatrix},$$

$$\boldsymbol{m}_2^\top = \max_x \{\operatorname{diag}(\boldsymbol{m}_1)\mathbf{P}_A\} \operatorname{diag}(\mathbf{O}_A(z_2 \mid \cdot)) = \begin{bmatrix} 0.25 & 0 & 0 \end{bmatrix},$$

$$\max_x \{\operatorname{diag}(\boldsymbol{m}_2)\mathbf{P}_A\} = \begin{bmatrix} 0 & 0.125 & 0.125 \end{bmatrix},$$

$$\boldsymbol{i}_3 = \operatorname*{argmax}_x \{\operatorname{diag}(\boldsymbol{m}_2)\mathbf{P}_A\} = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix},$$

$$\boldsymbol{m}_3^\top = \max_x \{\operatorname{diag}(\boldsymbol{m}_2)\mathbf{P}_A\} \operatorname{diag}(\mathbf{O}_A(z_3 \mid \cdot)) = \begin{bmatrix} 0 & 0.063 & 0.063 \end{bmatrix}.$$

We thus have $x_3^* = \operatorname{argmax}_x m_3(x)$. Since $m_3(2) = m_3(3)$, both states are equally probable, so we can select any of the two. In both cases, $x_2^* = i_3(x_3^*) = 1$, $x_1^* = i_2(x_2^*) = 3$ and $x^*(0) = i_1(x_1^*) = 1$. Therefore, the two most likely sequences are $1 \to 3 \to 1 \to 2$ or $1 \to 3 \to 1 \to 3$.

Question 3. (**3 pts.**)

Consider the MDP obtained from $\mathcal{M}$ by ignoring partial observability, and the policy

$$\boldsymbol{\pi} = \begin{bmatrix} 1.0 & 0.0 \\ 1.0 & 0.0 \\ 0.5 & 0.5 \end{bmatrix}.$$

(a) **(1.5 pts.)** Draw the transition diagram for the Markov chain describing the evolution of the MDP when the agent follows policy π.

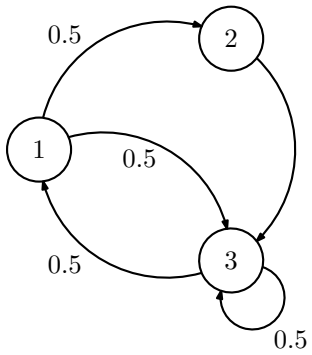
(b) **(0.5 pt.)** Is the chain irreducible? Why?

(c) **(0.5 pt.)** Is the chain aperiodic? Why?

(d) **(0.5 pt.)** Is the chain ergodic? Why?

Solution 3.

(a) The transition diagram for the resulting chain is



(b) We start by noting, from the transition diagram and using the Chapman-Kolmogorov where needed, that

$$\begin{array}{lll} \mathbf{P}^2(1 \mid 1) > \mathbf{P}(3 \mid 1)\mathbf{P}(1 \mid 3) > 0 & \mathbf{P}(2 \mid 1) > 0 & \mathbf{P}(3 \mid 1) > 0 \\ \mathbf{P}^2(1 \mid 2) > \mathbf{P}(3 \mid 2)\mathbf{P}(1 \mid 3) > 0 & \mathbf{P}^3(2 \mid 2) > \mathbf{P}(3 \mid 2)\mathbf{P}(1 \mid 3)\mathbf{P}(2 \mid 1) > 0 & \mathbf{P}(3 \mid 2) > 0 \\ \mathbf{P}(1 \mid 3) > 0 & \mathbf{P}(2 \mid 3) > \mathbf{P}(1 \mid 3)\mathbf{P}(2 \mid 1) > 0 & \mathbf{P}(3 \mid 3) > 0. \end{array}$$

Therefore, all states communicate, there is a single recurrence class and the chain is irreducible.

(c) We note that $\mathbf{P}(3 \mid 3) > 0$, so the period of state 3 is $d_3 = 1$. Since the chain is irreducible, all states have the same period, so the chain is aperiodic.

(d) Since the chain is irreducible and aperiodic, it is ergodic.

Question 4. (**6 pts.**)

Consider once again the MDP obtained from $\mathcal{M}$ by ignoring partial observability, and the policy π_A that always selects action A. Suppose that the optimal cost-to-go associated with that policy is

$$\boldsymbol{J}^{\pi_A} = \begin{bmatrix} 6.29 \\ 6.09 \\ 5.66 \end{bmatrix}.$$

(a) **(3 pt.)** Show that the policy π_A is *not* optimal.

(b) **(3 pts.)** Show that the greedy policy with respect to J^{π_A} is optimal.

Useful fact:

$$\begin{bmatrix} a & b & c \\ 0 & d & e \\ 0 & 0 & f \end{bmatrix}^{-1} = \begin{bmatrix} \frac{1}{a} & -\frac{b}{ad} & \frac{be-cd}{afd} \\ 0 & \frac{1}{d} & -\frac{e}{fd} \\ 0 & 0 & \frac{1}{f} \end{bmatrix}.$$

Solution 4.

(a) If π_A is optimal, the corresponding cost-to-go is the fixed point of the operator T and should verify $\boldsymbol{J}^{\pi_A} = \mathsf{T}\boldsymbol{J}^{\pi_A}$. Let us then apply T to $\boldsymbol{J}^{\pi_A}$. We have, for action A,

$$\boldsymbol{C}_{:,A} + \gamma \mathbf{P}_A \boldsymbol{J}^{\pi_A} = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} + 0.9 \begin{bmatrix} 0.0 & 0.5 & 0.5 \\ 0.0 & 0.0 & 1.0 \\ 1.0 & 0.0 & 0.0 \end{bmatrix} \begin{bmatrix} 6.29 \\ 6.09 \\ 5.66 \end{bmatrix} = \begin{bmatrix} 6.29 \\ 6.09 \\ 5.66 \end{bmatrix},$$

and for action B,

$$\boldsymbol{C}_{:,B} + \gamma \mathbf{P}_B \boldsymbol{J}^{\pi_A} = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} + 0.9 \begin{bmatrix} 1.0 & 0.0 & 0.0 \\ 0.0 & 1.0 & 0.0 \\ 0.0 & 0.0 & 1.0 \end{bmatrix} \begin{bmatrix} 6.29 \\ 6.09 \\ 5.66 \end{bmatrix} = \begin{bmatrix} 6.66 \\ 6.48 \\ 5.09 \end{bmatrix}.$$

This means that

$$\mathsf{T}\boldsymbol{J}^{\pi_A} = \min \left\{ \begin{bmatrix} 6.29 \\ 6.09 \\ 5.66 \end{bmatrix}, \begin{bmatrix} 6.66 \\ 6.48 \\ 5.09 \end{bmatrix} \right\} = \begin{bmatrix} 6.29 \\ 6.09 \\ 5.09 \end{bmatrix} \neq \boldsymbol{J}^{\pi_A}.$$

Therefore, π_A is not optimal.

(b) From Question 4(a), we have that

$$\boldsymbol{Q}^{\pi_A} = \begin{bmatrix} 6.29 & 6.66 \\ 6.09 & 6.48 \\ 5.66 & 5.09 \end{bmatrix},$$

and the corresponding greedy policy comes

$$\boldsymbol{\pi}' = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

To show that it is optimal, we show that it remains unchanged after one step of policy iteration. We have that

$$\boldsymbol{C}_{\pi'} = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} \qquad \mathbf{P}_{\pi'} = \begin{bmatrix} 0.0 & 0.5 & 0.5 \\ 0.0 & 0.0 & 1.0 \\ 0.0 & 0.0 & 1.0 \end{bmatrix},$$

leading to

$$(\boldsymbol{I} - \gamma \mathbf{P}_{\pi'})^{-1} = \begin{bmatrix} 1.0 & -0.45 & -0.45 \\ 0.0 & 1.0 & -0.9 \\ 0.0 & 0.0 & 0.1 \end{bmatrix}^{-1} = \begin{bmatrix} 1.0 & 0.45 & 8.55 \\ 0.0 & 1.0 & 9.0 \\ 0.0 & 0.0 & 10.0 \end{bmatrix},$$

using the provided expression. We can now compute $\boldsymbol{J}^{\pi'}$ as

$$\boldsymbol{J}^{\pi'} = (\boldsymbol{I} - \gamma \mathbf{P}_{\pi'})^{-1} \boldsymbol{C}_{\pi'} = \begin{bmatrix} 1.45 \\ 1.0 \\ 0.0 \end{bmatrix}.$$

The greedy policy with respect to $J^{\pi'}$ can be computed from the corresponding Q-function. We have

$$\boldsymbol{Q}^{\pi'}(:, A) = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} + 0.9 \begin{bmatrix} 0.0 & 0.5 & 0.5 \\ 0.0 & 0.0 & 1.0 \\ 1.0 & 0.0 & 0.0 \end{bmatrix} \begin{bmatrix} 1.45 \\ 1.0 \\ 0 \end{bmatrix} = \begin{bmatrix} 1.45 \\ 1.0 \\ 1.31 \end{bmatrix},$$

and

$$\boldsymbol{Q}^{\pi'}(:, B) = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} + 0.9 \begin{bmatrix} 1.0 & 0.0 & 0.0 \\ 0.0 & 1.0 & 0.0 \\ 0.0 & 0.0 & 1.0 \end{bmatrix} \begin{bmatrix} 1.45 \\ 1.0 \\ 0 \end{bmatrix} = \begin{bmatrix} 2.305 \\ 1.9 \\ 0 \end{bmatrix}.$$

The conclusion follows.

Question 5. (**2 pts.**)

Consider once again the POMDP in the gray box of page 4. Using the optimal Q-values computed in Question 4(b), compute the action prescribed by the Q-MDP heuristic in the belief computed in Question 2(a).

Note: If you did not answer Question 2(a), use the belief $\boldsymbol{b} = \begin{bmatrix} 0 & 0.5 & 0.5 \end{bmatrix}$. If you did not answer Question 4(b), use the Q-function

$$Q = \begin{bmatrix} 2.9 & 4.6 \\ 2.0 & 3.8 \\ 2.6 & 0.0 \end{bmatrix}.$$

Solution 5.

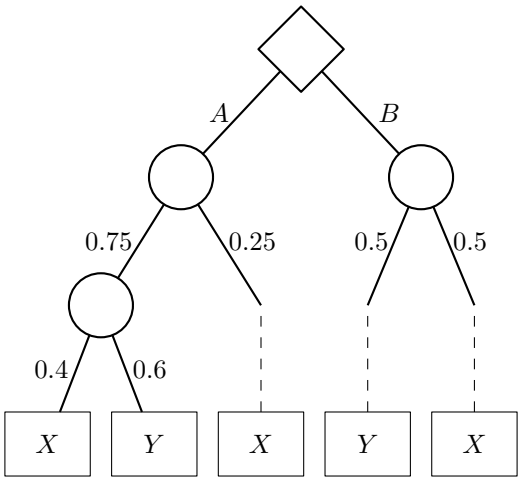
To compute the action prescribed by the Q-MDP heuristic, we compute

$$\underset{a \in \mathcal{A}}{\operatorname{argmin}}\, \boldsymbol{b}^\top \boldsymbol{Q}^* = \operatorname{argmin} \begin{bmatrix} 0 & 0.5 & 0.5 \end{bmatrix} \begin{bmatrix} 1.45 & 2.305 \\ 1.0 & 1.9 \\ 1.31 & 0.0 \end{bmatrix} = \operatorname{argmin} \begin{bmatrix} 1.15 & 0.95 \end{bmatrix},$$

yielding the action $a = B$.

Question 6. (**2 pts.**)

Consider the decision problem described by the following decision tree.



(a) **(1 pts.)** Knowing that the decision maker selected action A, indicate a *utility function* that may explain the agent's action.

(b) **(1 pts.)** Knowing that the decision maker selected action B, indicate a *utility function* that may explain the agent's action.

Solution 6.

(a) We compute the expected utility associated with each action as a function of the utility of outcomes X and Y. We have

$$Q(A) = 0.75 \times 0.4 \times u(X) + 0.75 \times 0.6 \times u(Y) + 0.25 \times u(X) = 0.55u(X) + 0.45u(Y)$$
$$Q(B) = 0.5u(Y) + 0.5u(X).$$

The decision maker will prefer action A if $Q(A) > Q(B)$, i.e., if $0.55u(X) + 0.45u(Y) > 0.5u(Y) + 0.5u(X)$. Equivalently, the decision maker will prefer action A if $u(X) > u(Y)$. Then, the utility $u(x) = \mathbb{I}[x = X]$ explains the observed behavior.

(b) Following on the previous question, the decision maker will prefer action B if $u(X) < u(Y)$. Then, the utility $u(x) = \mathbb{I}[x = Y]$ explains the observed behavior.