TÉCNICO LISBOA

**Planning, Learning and Decision Making**

MSc in Computer Science and Engineering

First test – November 8 , 2018

# Instructions

- You have 90 minutes to complete the test.
- Make sure that your test has a total of 7 pages and is not missing any sheets, then write your full name and student n. on this page (and your number in all others).
- 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.**)


Figure 1: A taxi moves around in a $2 \times 2$ grid. It must pick up the passenger and drop it in the cell marked as "Goal". **Note:** The positions and orientations of the taxi and passenger in this diagram are just indicative and can be ignored in your model.

Consider the following problem, depicted in Fig. 1. A taxi moving in $2 \times 2$ grid must pick up a passenger and drop her in the cell marked with "Goal". Both the taxi and the passenger can occupy any position in the grid.

The taxi has 6 actions available: moving up, down, left and right, as well as a "pick-up" action and a "drop" action. When the taxi and the passenger occupy the same cell, the "pick-up" action lets the passenger into the taxi. On the other hand, when the passenger and the taxi occupy different cells, the "pick-up" action has no effect. Finally, when the passenger is in the taxi, the "pick-up" action also has no effect.

The movement actions are deterministic and move the taxi to the adjacent position in the corresponding direction (if there is one). When the passenger is in the taxi, she moves with the taxi. Otherwise, the position of the passenger does not change.

The "drop" action lets the passenger out of the taxi in the taxi's current position. If the position is other than the goal, the passenger remains in that position; if the position is the goal, the passenger goes away and a new passenger appears in one of the other 3 cells at random. If the passenger is not in the taxi, the "drop" action has no effect.

The goal of the taxi driver is to drop the passenger in the goal cell.

Describe the decision problem faced by the taxi driver using the adequate type of model. In particular, you should indicate:

- The type of model needed to describe the decision problem of the taxi;
- The state space;
- The action space;
- The observation space (if relevant);
- The transition probabilities for actions "pick-up" and "drop"—you can provide the transition probability matrices for these actions, but other representations are also admissible;
- The immediate cost function.

**Solution 1.**

- The state should include the position of the taxi, the position of the passenger, and assess whether or not the passenger is in the taxi. This leads to the following state space:

$$\mathcal{X} = \big\{(u, v) \mid u \in \{1, 2, 3, 4\}, v \in \{1, 2, 3, 4, T\}\big\},$$

  where $u$ is the position of the taxi (in one of the cells 1, 2, 3, or 4), $v$ is the position of the passenger (in one of the cells 1, 2, 3, or 4, or in the taxi).

- $\mathcal{A} = \{U, D, L, R, Pick, Drop\}$.

- The action "Pickup" only succeeds if the position of the passenger and the taxi coincide, in which case the passenger goes into the taxi. Otherwise, the state should remain unchanged. This yields

$$\mathbf{P}_{Pickup}\big((u', v') \mid (u, v)\big) = \begin{cases} 1 & \text{if } u = v, u' = u \text{ and } v' = T \\ 1 & \text{if } u \neq v \text{ and } (u, v) = (u', v') \\ 0 & \text{otherwise.} \end{cases}$$

  Conversely, the action "Drop" only succeeds if the passenger is in the taxi, in which case the passenger leaves at the taxi's current position—or completely, leading to the arrival of a new passenger. Otherwise, the state should remain unchanged. This yields

$$\mathbf{P}_{Drop}\big((u', v') \mid (u, v)\big) = \begin{cases} 1 & \text{if } u \neq 2, v = T, u' = u, \text{ and } v' = u \\ \frac{1}{3} & \text{if } u = 2, v = T, u' = u, \text{ and } v' \in \{1, 3, 4\} \\ 1 & \text{if } v \neq T \text{ and } (u, v) = (u', v') \\ 0 & \text{otherwise.} \end{cases}$$

- Finally, the cost function should reward the taxi driver for dropping the passenger at the destination, yielding

$$c_{Drop}\big((u', v'), a\big) = \begin{cases} 0 & \text{if } (u, v) = (2, T) \text{ and } a = Drop \\ 1 & \text{otherwise.} \end{cases}$$

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

- $\mathcal{X} = \{1, 2, 3\}$;
- $\mathcal{A} = \{a, b, c\}$;
- $\mathcal{Z} = \{u, v\}$;
- The transition probabilities are

$$\mathbf{P}_a = \begin{bmatrix} 0.5 & 0.5 & 0.0 \\ 0.75 & 0.25 & 0.0 \\ 0.2 & 0.8 & 0.0 \end{bmatrix}; \qquad \mathbf{P}_b = \begin{bmatrix} 0.5 & 0.5 & 0.0 \\ 0.2 & 0.8 & 0.0 \\ 0.8 & 0.2 & 0.0 \end{bmatrix}; \qquad \mathbf{P}_c = \begin{bmatrix} 0.5 & 0.5 & 0.0 \\ 0.0 & 0.2 & 0.8 \\ 0.0 & 0.0 & 1.0 \end{bmatrix}.$$

- The observation probabilities are

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

- The cost function $c$ is given by $c(x) = 1 - \mathbb{I}(x = 3)$.
- Finally, the discount is given by $\gamma = 0.9$.

**Question 2.** (**4 pts.**)
Suppose that the POMDP $\mathcal{M}$ departs from the initial state $\mathrm{x}_0 = 1$.

(a) **(2 pts.)** What would be the most likely state at time step $t = 2$ if the agent took actions $\boldsymbol{a}_{0:1} = \{a, b\}$? Indicate the relevant computations.

(b) **(2 pts.)** Suppose that the agent actually takes the actions in Question (a) and makes the observations $\boldsymbol{z}_{1:2} = \{u, v\}$. What is the most likely sequence of states given the sequence of actions and observations? Indicate the relevant computations.

**Solution 2.**

(a) Considering only the transition information, we have that

$$\begin{aligned}
\mu_3 &= \mu_0 \mathbf{P}_a \mathbf{P}_b \\
&= \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0.5 & 0.5 & 0.0 \\ 0.75 & 0.25 & 0.0 \\ 0.2 & 0.8 & 0.0 \end{bmatrix} \begin{bmatrix} 0.5 & 0.5 & 0.0 \\ 0.2 & 0.8 & 0.0 \\ 0.8 & 0.2 & 0.0 \end{bmatrix} \\
&= \begin{bmatrix} 0.35 & 0.65 & 0.0 \end{bmatrix}.
\end{aligned}$$

The most likely state is, therefore, state 2.

(b) We use the Viterbi algorithm. We have

$$\begin{aligned}
m_0 &= \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} \\
m_1 &= \big( \max \operatorname{diag}(\boldsymbol{m}_0) \mathbf{P}_a \big) \operatorname{diag}(\mathbf{O}_{a,u}) = \begin{bmatrix} 0.0 & 0.35 & 0.0 \end{bmatrix} \\
i_0 &= \operatorname{argmax} \operatorname{diag}(\boldsymbol{m}_0) \mathbf{P}_a \big) = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix} \\
m_2 &= \big( \max \operatorname{diag}(\boldsymbol{m}_1) \mathbf{P}_b \big) \operatorname{diag}(\mathbf{O}_{b,v}) = \begin{bmatrix} 0.07 & 0.084 & 0.0 \end{bmatrix} \\
i_1 &= \operatorname{argmax} \operatorname{diag}(\boldsymbol{m}_1) \mathbf{P}_b \big) = \begin{bmatrix} 2 & 2 & 1 \end{bmatrix}
\end{aligned}$$

The most likely state at $t = 2$ is, therefore, $x_2 = 2$, and we get the most likely sequence as $\boldsymbol{x}_{0:2} = \{1, 2, 2\}$.

**Question 3.** (**6 pts.**)

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

$$\boldsymbol{\pi} = \begin{bmatrix} 0.2 & 0.7 & 0.1 \\ 0.0 & 0.0 & 1.0 \\ 0.0 & 0.0 & 1.0 \end{bmatrix}.$$

(a) **(3 pts.)** Compute the cost-to-go $J^\pi$ associated with the policy $\pi$ above.

(b) **(3 pts.)** Is policy $\pi$ optimal? Support your answer with the adequate computations.

**Solution 3.**

(a) We have that

$$\boldsymbol{J}^\pi = (\mathbf{I} - \gamma \mathbf{P}_\pi)^{-1} \boldsymbol{c}_\pi.$$

From the provided policy,

$$\boldsymbol{c}_\pi = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, \qquad \mathbf{P}_\pi = \begin{bmatrix} 0.5 & 0.5 & 0.0 \\ 0.0 & 0.2 & 0.8 \\ 0.0 & 0.0 & 1.0 \end{bmatrix},$$

yielding

$$\begin{aligned} \boldsymbol{J}^\pi &= \begin{bmatrix} 0.55 & -0.45 & 0.0 \\ 0.0 & 0.82 & -0.72 \\ 0.0 & 0.0 & 0.1 \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} \\ &= \begin{bmatrix} 1.82 & 1.0 & 7.18 \\ 0.0 & 1.22 & 8.78 \\ 0.0 & 0.0 & 10.0 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} \\ &= \begin{bmatrix} 2.82 \\ 1.22 \\ 0 \end{bmatrix}. \end{aligned}$$

(b) We perform one step of value iteration (or, equivalently, policy iteration). Using $J^\pi$ as computed in (a), we have:

$$\begin{aligned} \boldsymbol{Q}^\pi_{:,a} &= \mathbf{C}_{:,a} + \gamma \mathbf{P}_a \boldsymbol{J}^\pi \\ &= \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} + 0.9 \begin{bmatrix} 0.5 & 0.5 & 0.0 \\ 0.75 & 0.25 & 0.0 \\ 0.2 & 0.8 & 0.0 \end{bmatrix} \begin{bmatrix} 2.82 \\ 1.22 \\ 0.0 \end{bmatrix} = \begin{bmatrix} 2.82 \\ 3.18 \\ 1.39 \end{bmatrix}, \\ \boldsymbol{Q}^\pi_{:,b} &= \mathbf{C}_{:,b} + \gamma \mathbf{P}_b \boldsymbol{J}^\pi \\ &= \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} + 0.9 \begin{bmatrix} 0.5 & 0.5 & 0.0 \\ 0.2 & 0.8 & 0.0 \\ 0.8 & 0.2 & 0.0 \end{bmatrix} \begin{bmatrix} 2.82 \\ 1.22 \\ 0.0 \end{bmatrix} = \begin{bmatrix} 2.82 \\ 2.39 \\ 2.25 \end{bmatrix}, \\ \boldsymbol{Q}^\pi_{:,c} &= \mathbf{C}_{:,c} + \gamma \mathbf{P}_c \boldsymbol{J}^\pi \\ &= \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} + 0.9 \begin{bmatrix} 0.5 & 0.5 & 0.0 \\ 0.0 & 0.2 & 0.8 \\ 0.0 & 0.0 & 1.0 \end{bmatrix} \begin{bmatrix} 2.82 \\ 1.22 \\ 0.0 \end{bmatrix} = \begin{bmatrix} 2.82 \\ 1.22 \\ 0.0 \end{bmatrix}. \end{aligned}$$

Finally,

$$\boldsymbol{J}_{\text{new}} = \min_a \boldsymbol{Q}_{:,a}, = \begin{bmatrix} 2.82 \\ 1.22 \\ 0.0 \end{bmatrix},$$

where the minimum is taken row-wise. Since $J_{\text{new}} = J^\pi$, we can conclude that $J^\pi = J^*$ and thus $\pi = \pi^*$.

**Question 4.** (**2 pts.**)

Show that, in a POMDP, the belief is a sufficient statistic for the history of actions and observations.

**Solution 4.**

In a POMDP, the belief at time step $t$ is defined as a vector $\boldsymbol{b}_t$ with $x$th component

$$\boldsymbol{b}_t(x) = \mathbb{P}\left[\mathrm{x}_t = x \mid \mathcal{H}_t\right] = \mathbb{P}\left[\mathrm{x}_t = x \mid \mathrm{a}_{t-1} = a, \mathrm{z}_t = z, \mathcal{H}_{t-1}\right],$$

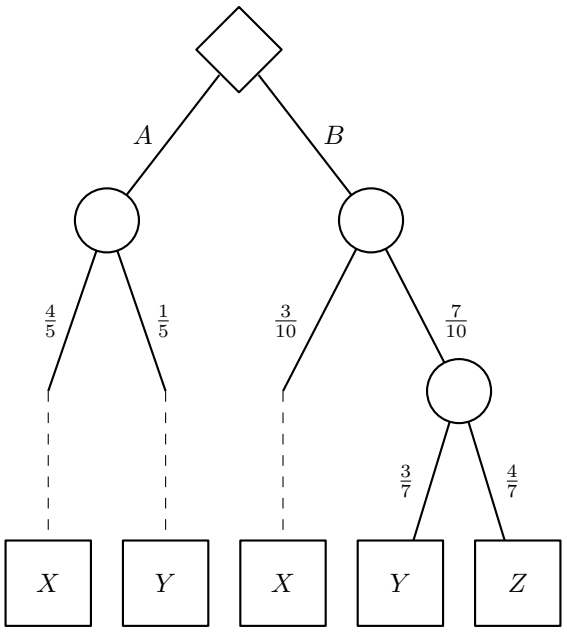
where $\mathcal{H}_t$ is the history up to time step $t$ and we made explicit the action at time-step $t-1$, $a$, and the observation at time step $t$, $z$. On the other hand, from the belief update equation, we now have that

$$\boldsymbol{b}_t(x) = \frac{\sum_{y \in \mathcal{X}} \boldsymbol{b}_{t-1}(y) \mathbf{P}_a(x \mid y) \mathbf{O}_a(z \mid x)}{\sum_{x', y \in \mathcal{X}} \boldsymbol{b}_{t-1}(y) \mathbf{P}_a(x' \mid y) \mathbf{O}_a(z \mid x')}. \tag{1}$$

Comparing (1), with the prior definition of belief, we note the common presence of the action at time-step $t-1$, $a$, and the observation at time step $t$, $z$. However, in (1) there is no explicit reference to the history $\mathcal{H}_{t-1}$ and, instead, the belief $\boldsymbol{b}_{t-1}$ is present. But this must mean that all the relevant information in $\mathcal{H}_{t-1}$ must be captured in $\boldsymbol{b}_{t-1}$, showing that the belief is a sufficient statistic for the history.

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

Consider the decision problem described by the following decision tree.



Compute the optimal action, assuming that $u(X) = 1$, $u(Y) = 0$ and $u(Z) = 1.5$.

**Solution 5.**

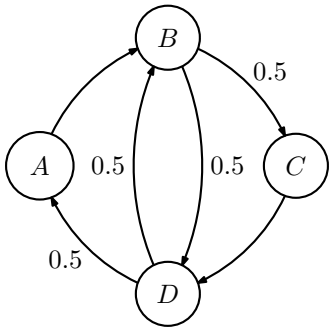
We have that

$$Q(A) = \frac{4}{5} \times 1 + \frac{1}{5} \times 0 = 0.8$$
$$Q(B) = \frac{3}{10} \times 1 + \frac{3}{10} \times 0 + \frac{4}{10} \times 1.5 = 0.9.$$

The optimal action is $a = B$.

**Question 6.** (**3 pts.**)

Consider the Markov chain described by the following transition diagram.




(a) **(1 pt.)** Indicate the state space and transition probability matrix for the chain.

(b) **(1 pt.)** Show that

$$\boldsymbol{\mu} = \begin{bmatrix} \frac{1}{6} & \frac{1}{3} & \frac{1}{6} & \frac{1}{3} \end{bmatrix}$$

is a stationary distribution for the chain.

(c) **(1 pt.)** Is the chain ergodic? Why?

**Solution 6.**

(a) From the transition diagram we get, immediately,

$$\mathcal{X} = \{A, B, C, D\} \qquad \mathbf{P} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0.5 & 0.5 \\ 0 & 0 & 0 & 1 \\ 0.5 & 0.5 & 0 & 0 \end{bmatrix}.$$

(b) To show that $\boldsymbol{\mu}$ is stationary, we note that

$$\boldsymbol{\mu}\mathbf{P} = \begin{bmatrix} \frac{1}{6} & \frac{1}{3} & \frac{1}{6} & \frac{1}{3} \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0.5 & 0.5 \\ 0 & 0 & 0 & 1 \\ 0.5 & 0.5 & 0 & 0 \end{bmatrix} = \begin{bmatrix} \frac{1}{6} & \frac{2}{6} & \frac{1}{6} & \frac{2}{6} \end{bmatrix} = \boldsymbol{\mu}.$$

(c) The chain is irreducible (it has a single communicating class) and aperiodic (all states have a period of 1), from which we can conclude that it is ergodic.