TÉCNICO LISBOA

Planning, Learning and Decision Making

MSc in Computer Science and Engineering

First test – November 7, 2019

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 your number in all others).
- The test has a total of 7 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.)

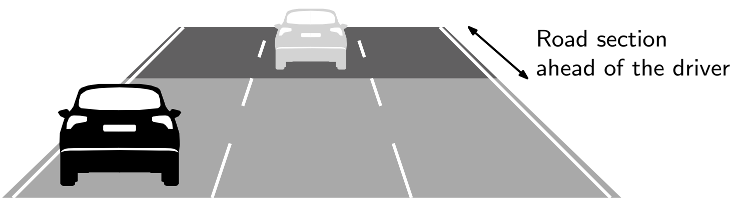


Figure 1: A driver speeds through a 3-line highway. The driver is moving faster than other cars, so it must avoid cars appearing ahead.

Consider the following problem. A driver is speeding through a 3-lane highway. The driver's vehicle moves faster than the other cars, so it must be careful to avoid crashing into them.

At any moment, the car may be in any of the three lanes. Similarly, in the road segment immediately ahead of the driver's, there may be (at most) one car in any of the three lanes. We assume that cars further ahead do not matter. Figure 1 illustrates one possible situation, where the car in the left-most lane, and there is a car ahead in the central lane.

If there is a car in any of the three lanes ahead of the driver at some time step, all lanes of the road segment ahead will be clear at the next time step (both if the driver safely passed by such car or if there was a crash). Conversely, if all lanes of the road segment ahead of the driver are clear at some time step, in the next time step a car will appear in one of the three lanes uniformly at random.

At any moment, the driver may decide to stay in the same lane, move to the lane immediately to the right (if there is one) or to the lane immediately to the left (if there is one). The movement succeeds with 0.9 probability, but with a 0.1 probability, the movement fails and the driver remains in the same lane.

If, at any moment, the driver selects an action that brings it (or leaves it) behind another car it will cause an accident, something the driver wishes to avoid. On the other hand, the driver also wants to adopt a smooth driving style, so it avoids changing lanes as much as possible.

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 the action corresponding to moving to the lane immediately to the right;
- The observation probabilities (if relevant);
- The immediate cost function. Note that your cost depends on the *outcome* of the action.

Solution 1.

We describe the problem using a Markov decision problem, since the relevant information (position of the car and whether there are cars ahead) is fully observable to the agent. The MDP is thus a tuple $(\mathcal{X}, \mathcal{A}, \{\mathbf{P}_a\}, c, \gamma)$ where

- The state space, as stated above, includes the position (lane) of the driver and whether or not there is a car in any of the lanes ahead. This leads to

  $$\mathcal{X} = \{(L, \emptyset), (M, \emptyset), (R, \emptyset), (L, L), (M, L), (R, L), (L, M), (M, M), (R, M), (L, R), (M, R), (R, R)\},$$

  where each state is a pair (c, o), with c the lane of the driver and o the lane of cars ahead. c can be either L, M, or R (left, middle, and right lane, respectively); o can also take any of the three previous values and can also be $\emptyset$, indicating that there is no car in any of the three lanes ahead.
- There are three possible actions: stay in the same lane (S), move to the lane immediately to the right (R) or move to the lane immediately to the left (L). Therefore, $\mathcal{A} = \{S, R, L\}$.
- There are no observations or observation probabilities, as the model is an MDP.
- The transition probabilities for the action R are

  $$\mathbf{P}_R = \begin{bmatrix} 0 & 0 & 0 & \frac{1}{30} & \frac{3}{10} & 0 & \frac{1}{30} & \frac{3}{10} & 0 & \frac{1}{30} & \frac{3}{10} & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{30} & \frac{3}{10} & 0 & \frac{1}{30} & \frac{3}{10} & 0 & \frac{1}{30} & \frac{3}{10} \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{3} & 0 & 0 & \frac{1}{3} & 0 & 0 & \frac{1}{3} \\ \frac{1}{10} & \frac{9}{10} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{10} & \frac{9}{10} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{10} & \frac{9}{10} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{10} & \frac{9}{10} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{10} & \frac{9}{10} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{10} & \frac{9}{10} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$
- Finally, the costs depend on the actions, and penalize both crashes and changing lanes. Therefore, we assign maximum cost (1.0) to an accident, a medium cost (0.1) to changing lanes, and minimum cost to safely staying in the same lane. This leads to the cost function

  $$\mathbf{C} = \begin{bmatrix} 0.0 & \boxed{0.09} & 0.0 \\ 0.0 & 0.09 & 0.09 \\ 0.0 & 0.0 & 0.09 \\ 1.0 & \boxed{0.19} & 1.0 \\ 0.0 & 0.09 & \boxed{0.99} \\ 0.0 & 0.0 & 0.09 \\ 0.0 & 0.99 & 0.0 \\ 1.0 & 0.19 & 0.19 \\ 0.0 & 0.0 & 0.99 \\ 0.0 & 0.09 & 0.0 \\ 0.0 & 0.99 & 0.09 \\ 1.0 & 1.0 & 0.19 \end{bmatrix}.$$

  We provide a quick description regarding the computation of the boxed numbers. The first (0.09) corresponds to the cost of moving to a contiguous lane where there is no car ahead. The action fails with probability 0.1, in which case the driver does not change lane and thus pays no cost, but succeeds with 0.9 probability, in which case the driver pays a cost of 0.1. The resulting cost is, therefore, $0.1 \times 0.9 = 0.09$.

The second value corresponds to the situation where the driver wishes to deviate from a car approaching ahead. The action fails with probability 0.1, in which case there is an accident (with a cost of 1). If the action succeeds, which happens with probability 0.9, the agent changes lane and avoids the accident, but pays a cost of 0.1. This leads to $0.1 \times 0.9 + 0.1 \times 1 = 0.19$.

Finally, the third value corresponds to the situation where the agent changes lane behind an approaching car. If the action fails, the car stays in the same lane and the driver pays 0. However, if the action succeeds—which happens with a probability 0.9—the agent changes lane (a cost of 0.1) and causes an accident (a cost of 1.0), leading to a total cost of $0.9 \times (0.1 + 1.0) = 0.99$.

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

Consider an arbitrary MDP $(\mathcal{X}, \mathcal{A}, \{\mathbf{P}_a\}, c, \gamma)$, where $\mathcal{X}$ and $\mathcal{A}$ are assumed finite. Show that, by adding an arbitrary constant $k \in \mathbb{R}$ to the cost function c, the optimal policy does not change.

Solution 2.

For any policy π and any state $x \in \mathcal{X}$,

$$J^\pi(x) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t \mathrm{c}_t \mid \mathrm{x}_0 = x\right].$$

If we add k to the cost function, the cost to go for the same policy and state is, with the new cost $\hat{c}$,

$$\hat{J}^\pi(x) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t \hat{\mathrm{c}}_t \mid \mathrm{x}_0 = x\right] = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t \mathrm{c}_t + k \mid \mathrm{x}_0 = x\right] = J^\pi(x) + \frac{k}{1-\gamma},$$

where the last equality follows from the fact that k does not depend on any of the random variables considered in the expectation. Thus, with the new cost,

$$\hat{J}^*(x) = \min_\pi \hat{J}^\pi(x) = \min_\pi \left\{J^\pi(x) + \frac{k}{1-\gamma}\right\} = \min_\pi \{J^\pi(x)\} + \frac{k}{1-\gamma} = J^*(x) + \frac{k}{1-\gamma}.$$

It follows that the policy that minimizes $\hat{J}^\pi$ is the same that minimizes J^π, and the conclusion follows. An alternative (and more precise) way of establishing the desired result departs from the recursive expression for $\hat{Q}^*$. We have

$$Q^*(x,a) = c(x,a) + \gamma \sum_{y \in \mathcal{X}} \mathbf{P}(y \mid x,a) \min_{a' \in \mathcal{A}} Q^*(y,a').$$

Adding $k/(1-\gamma)$ to both sides yields

$$Q^*(x,a) + \frac{k}{1-\gamma} = c(x,a) + \gamma \sum_{y \in \mathcal{X}} \mathbf{P}(y \mid x,a) \min_{a' \in \mathcal{A}} Q^*(y,a') + \frac{k}{1-\gamma}.$$

We now manipulate the expression to get the term $k/(1-\gamma)$ inside the summation:

$$\begin{aligned}
Q^*(x,a) + \frac{k}{1-\gamma} &= c(x,a) + \gamma \sum_{y \in \mathcal{X}} \mathbf{P}(y \mid x,a) \min_{a' \in \mathcal{A}} Q^*(y,a') + \frac{k + \gamma k - \gamma k}{1-\gamma} \\
&= c(x,a) + \gamma \sum_{y \in \mathcal{X}} \mathbf{P}(y \mid x,a) \min_{a' \in \mathcal{A}} \left[Q^*(y,a') + \frac{k}{1-\gamma}\right] + k \\
&= [c(x,a) + k] + \gamma \sum_{y \in \mathcal{X}} \mathbf{P}(y \mid x,a) \min_{a' \in \mathcal{A}} \left[Q^*(y,a') + \frac{k}{1-\gamma}\right].
\end{aligned}$$

The expression above is the recursive expression for $\hat{Q}^*$, which implies that

$$\hat{Q}^*(x,a) = Q^*(x,a) + \frac{k}{1-\gamma}$$

and hence

$$\hat{\pi}^*(x) = \operatorname*{argmin}_{a\in\mathcal{A}} \hat{Q}^*(x,a) = \operatorname*{argmin}_{a\in\mathcal{A}} Q^*(x,a).$$

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

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

$$\mathbf{P}_a = \begin{bmatrix} 0 & 0.5 & 0.5 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}; \qquad \mathbf{P}_b = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

- The observation probabilities are

$$\mathbf{O}_a = \mathbf{O}_b = \begin{bmatrix} 0.5 & 0.5 \\ 0.0 & 1.0 \\ 1.0 & 0.0 \end{bmatrix}.$$

- The cost function c is given by $c(x) = 1 - \mathbb{I}(x=2)$.
- Finally, the discount is given by $\gamma = 0.9$.

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

Consider the MDP obtained from $\mathcal{M}$ by ignoring partial observability, and suppose that the optimal Q-function for that MDP is

$$\mathbf{Q}^* = \begin{bmatrix} 1.45 & 1.9 \\ 1.31 & 0.0 \\ 1.0 & 1.9 \end{bmatrix}.$$

Compute the optimal policy for the MDP.

Solution 3.

The optimal policy is given by

$$\pi^*(x) = \underset{a\in\mathcal{A}}{\operatorname{argmin}}\, Q^*(x,a) = \begin{bmatrix} 1.0 & 0.0 \\ 0.0 & 1.0 \\ 1.0 & 0.0 \end{bmatrix}.$$

Question 4. (6 pts.)

Suppose that the initial belief for the POMDP $\mathcal{M}$ was

$$\boldsymbol{b}_0 = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix}.$$

(a) **(1.5 pts.)** Suppose that, at time step $t = 0$, the agent selects the action according to the Q-MDP heuristic (you can use the Q-function from Question 3). Compute the action selected by the agent.

(b) **(1.5 pts.)** Suppose that, as a consequence of the action selected in (a), the agent observes $z_1 = u$. Compute the resulting belief at time step $t = 1$.

Note: If you haven't solved (a), just pick one of the two actions a or b, indicating in your solution that you selected an action at random.

(c) **(1.5 pts.)** Suppose that the optimal cost-to-go for the POMDP can be represented using the α-vectors

$$\Gamma = \left\{ \begin{bmatrix} 1.93 & 0.03 & 1.93 \end{bmatrix}^\top, \begin{bmatrix} 1.48 & 1.54 & 1.03 \end{bmatrix}^\top \right\}$$

where the first corresponds to action b and the second to a. Compute the optimal action at $t = 0$.

(d) **(1.5 pts.)** Indicate one advantage of Q-MDP over the MLS and AV heuristic.

Solution 4.

(a) The Q-MDP heuristic prescribes the action

$$a_0 = \underset{a\in\mathcal{A}}{\operatorname{argmin}} \sum_{x\in\mathcal{X}} b_0(x) Q^*(x,a),$$

where Q^* is the Q-function from Question 3. We have that

$$\boldsymbol{b}_0 \mathbf{Q}^* = \begin{bmatrix} 1.25 & 1.27 \end{bmatrix},$$

and, therefore, $a_0 = a$.

(b) We use the belief update rule to get

$$\boldsymbol{b}_1 = \rho\, \boldsymbol{b}_0 \mathbf{P}_a \operatorname{diag}(\mathbf{O}_{a,u}),$$

where ρ is a normalization constant. We thus have

$$\boldsymbol{b}_1 = \rho \begin{bmatrix} \frac{1}{6} & 0 & \frac{1}{6} \end{bmatrix} = \begin{bmatrix} 0.5 & 0.0 & 0.5 \end{bmatrix}.$$

(c) We compute, for the belief $\boldsymbol{b}_0$, the value of the cost-to-go function as

$$J^*(\boldsymbol{b}_0) = \min_{\boldsymbol{\alpha}\in\Gamma} \boldsymbol{b}_0 \cdot \boldsymbol{\alpha}.$$

In our case, this yields

$$J^*(\boldsymbol{b}_0) = \min\{1.29, 1.35\} = 1.30,$$

and the corresponding action is action b.

(d) Both the MLS and AV heuristics use directly the optimal MDP policy and, therefore, prescribe only actions that are optimal for the underlying MDP. Q-MDP, on the other hand, may prescribe actions that the underlying MDP policy never selects. This was observed, for example, in the Tiger problem, where the Q-MDP did select the action "Listen", while neither MLS nor AV were able to select such action.

Question 5. (3 pts.)

Consider once again the POMDP $\mathcal{M}$ with initial belief

$$\boldsymbol{b}_0 = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix}.$$

(a) **(1.5 pts.)** Suppose that the agent performs the actions $\boldsymbol{a}_{0:1} = \{a, b\}$. Compute the distribution over states at time step $t = 2$, assuming that the agent makes no observations.

(b) **(1.5 pts.)** Suppose now that, as a consequence of the actions in (a), the agent observes $\boldsymbol{z}_{1:2} = \{u, u\}$. Compute the most likely state at time step $t = 2$.

Solution 5.

(a) To get the distribution of states at time step $t = 2$, we simply compute

$$\boldsymbol{\mu}_2 = \boldsymbol{b}_0\mathbf{P}_a\mathbf{P}_b = \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix}.$$

(b) The first step of the forward algorithm has been computed in Question (b), yielding

$$\boldsymbol{\mu}_{1|0:1} = \begin{bmatrix} \frac{1}{2} & 0 & \frac{1}{2} \end{bmatrix}.$$

We now perform another step of the forward algorithm to get

$$\boldsymbol{\mu}_{2|0:2} = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}.$$

We can thus conclude that the most likely state at time step $t = 2$ is 3.

Question 6. (3 pts.)

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

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

(a) **(1.5 pt.)** Show that

$$\boldsymbol{b}_0 = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix}.$$

is a stationary distribution for the chain defined by the policy π above.

(b) **(1.5 pt.)** Is the chain ergodic? Explain your reasoning.

Solution 6.

(a) The transition probabilities for the chain are given by

$$\mathbf{P} = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}.$$

Therefore, it suffices to show that

$$\boldsymbol{b}_0\mathbf{P} = \boldsymbol{b}_0.$$

We have that

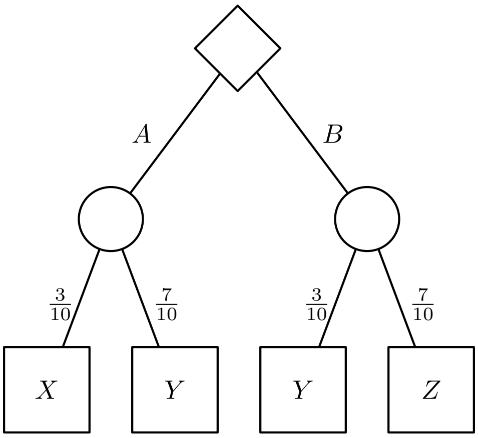
$$\boldsymbol{b}_0\mathbf{P} = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix} \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix},$$

and the conclusion follows.

(b) The chain is periodic with period 3. Therefore, for any initial distribution $\boldsymbol{b} \neq \boldsymbol{b}_0$, the state distribution will oscillate with a period of 3 and will not converge to $\boldsymbol{b}_0$. It follows that the chain is not ergodic.

Question 7. (**2 pts.**)

Consider the decision problem described by the following decision tree, where $u(X) = u(Z) > u(Y)$.



Compute the optimal action.

Solution 7.

We have that

$$\begin{aligned} Q(A) &= 0.3u(X) + 0.7u(Y) \\ Q(B) &= 0.3u(Y) + 0.7u(Z) = 0.7u(X) + 0.3u(Y). \end{aligned}$$

Since $u(X) > u(Y)$, it follows that $Q(B) > Q(A)$ and the optimal action is B.