TÉCNICO LISBOA

Planning, Learning and Decision Making

MSc in Computer Science and Engineering

Repeat examination — July 1, 2024

Instructions

- You have 120 minutes to complete the examination.
- 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 4 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. (**2.5 pts.**)

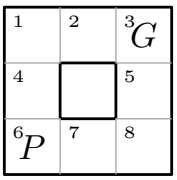


Figure 1: A taxi driver must pick up a passenger standing in the cell marked with "P" and drop it in the cell marked with "G".

Consider a taxi driver moving in the grid-world environment of Fig. 1. The driver must pick up a passenger (the passenger is in the cell marked with "P") and drop her at the cell marked with "G".

At each step, the agent has four actions available, corresponding to movement in the four directions: *move up, down, left*, or *right*. Movement across a gray cell division succeeds with a 0.8 probability, and fails with a 0.2 probability. Movements across black cell divisions have no effect. When a movement action fails, the taxi remains in the same cell.

The passenger is automatically picked up when the agent stands in the cell marked as P and drop off at the cell marked with G (the agent needs to take no explicit action). Upon dropping off the passenger in G, the taxi should transition to a final absorbing state with a cost of 0.

Describe the decision problem faced by the agent using the adequate type of model, indicating:

- The type of model needed to describe the decision problem of the agent;
- The state, action, and observation spaces (when relevant);
- The transition probabilities corresponding to the actions "Move Up";
- The immediate cost function. Make sure that the cost function is as simple as possible and verifies $c(x, a) \in [0, 1]$ for all states $x \in \mathcal{X}$ and actions $a \in \mathcal{A}$.

Solution 1.

At each time step, in order to decide where to go, the agent should keep track of its position in the maze, and whether or not it has picked the passenger in P. Therefore, the state consists of tuples (t, p), where $t \in \{1, 2, \ldots, 8\}$ and corresponds to the position of the taxi, and $p \in \{P, \neg P\}$, P indicating that the passenger has been picked, and $\neg P$ that it has not. There is one final state F, after the passenger is dropped. Since there is no information about the agent not being able to observe any of these two elements, we can model this problem as an MDP, with state space $\mathcal{X} = \{(1, \neg P), (2, \neg P), \ldots, (5, \neg P), (7, \neg P), (8, \neg P), (1, P), \ldots, (8, P), F\}$ (notice that there is no state $(6, \neg P)$).

The action space includes the 4 actions *move up, down, left*, and *right*, which we represent as $\mathcal{A} = \{U, D, L, R\}$.

The transition probabilities for the action U are given by

$$\mathbf{P}_U = \begin{bmatrix}
1.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\
0.0 & 1.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\
0.0 & 0.0 & 1.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\
0.8 & 0.0 & 0.0 & 0.2 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\
0.0 & 0.0 & 0.8 & 0.0 & 0.2 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\
0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 1.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\
0.0 & 0.0 & 0.0 & 0.0 & 0.8 & 0.0 & 0.2 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\
0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 1.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\
0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 1.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\
0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 1.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\
0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.8 & 0.0 & 0.0 & 0.2 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\
0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.8 & 0.0 & 0.2 & 0.0 & 0.0 & 0.0 & 0.0 \\
0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.8 & 0.0 & 0.2 & 0.0 & 0.0 & 0.0 \\
0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 1.0 & 0.0 & 0.0 \\
0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.8 & 0.0 & 0.0 & 0.2 & 0.0 \\
0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 1.0
\end{bmatrix}$$

and the cost function could be simply

$$c(x, a) = \begin{cases} 0 & \text{if } x = F, \\ 1 & \text{otherwise.} \end{cases}$$

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

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

- The observation probabilities are

$$\mathbf{O}_A = \mathbf{O}_B = \mathbf{O}_C = \begin{bmatrix} 1.0 & 0.0 \\ 0.0 & 1.0 \\ 0.0 & 1.0 \end{bmatrix}.$$

- The cost function c is given by

$$\mathbf{C} = \begin{bmatrix} 0.0 & 0.0 & 0.0 \\ 0.1 & 0.1 & 0.1 \\ 1.0 & 0.1 & 0.1 \end{bmatrix}.$$

- Finally, the discount is given by $\gamma = 0.9$.

You may also find useful the fact that, given a 3×3 matrix

$$\mathbf{A} = \begin{bmatrix} a & b & c \\ 0 & d & e \\ 0 & 0 & f \end{bmatrix},$$

it holds that

$$\mathbf{A}^{-1} = \begin{bmatrix} \frac{1}{a} & -\frac{b}{ad} & \frac{be-cd}{adf} \\ 0 & \frac{1}{d} & -\frac{e}{df} \\ 0 & 0 & \frac{1}{f} \end{bmatrix}.$$

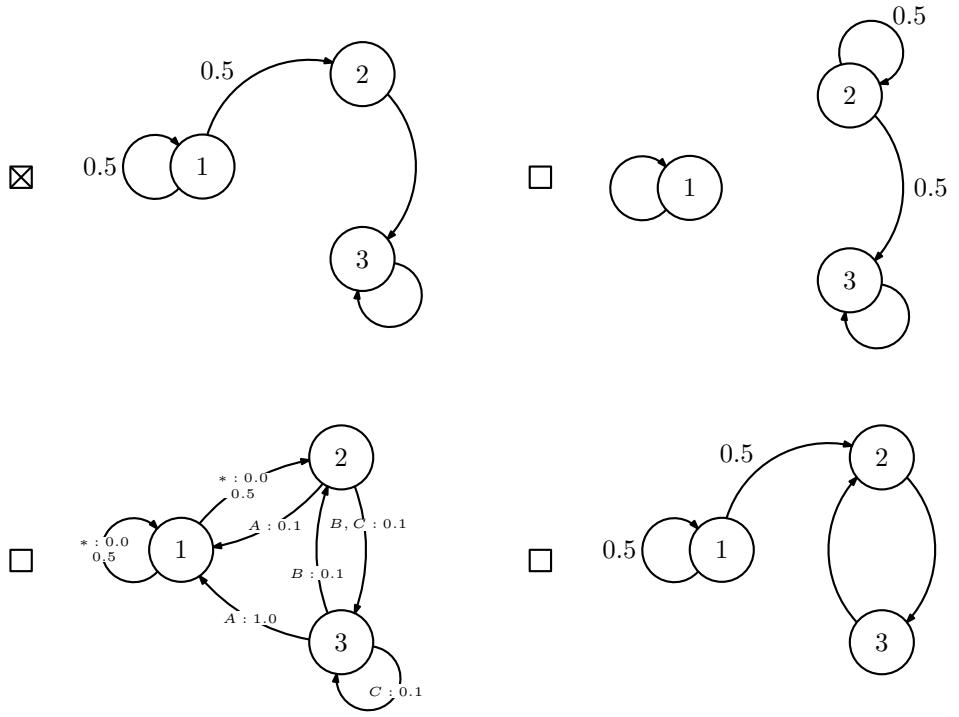
Question 2. (**8 pts.**)

For each of the following questions, indicate the *single most correct answer*.

(a) **(0.8 pts.)** Consider the MDP obtained from $\mathcal{M}$ by ignoring partial observability, which corresponds to the tuple $(\mathcal{X}, \mathcal{A}, \{\mathbf{P}_a\}, c, \gamma)$. Consider also the policy π given by

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

The Markov chain induced by policy π can be represented by the transition diagram...



(b) **(0.8 pts.)** Consider once again the Markov chain from (a).

- ☐ The Markov chain has a single communicating class.
- ☐ The Markov chain has two communicating classes.
- ☒ **The Markov chain has three communicating classes.**
- ☐ There is not enough information to determine the number of communicating classes.

(c) **(0.8 pts.)** Consider the POMDP $\mathcal{M}$ and suppose that the agent observes $\boldsymbol{z}_{1:2}$ after taking actions $\boldsymbol{a}_{0:1}$ from some initial distribution μ_0. Then, ...

- ☐ ... $\boldsymbol{b}_2 = \rho\mu_0\mathbf{P}_{a_0}\mathbf{P}_{a_1}\mathbf{O}_{a_0}\mathbf{O}_{a_1}^{\top}$, where ρ is a normalization constant.
- ☒ **... $\boldsymbol{b}_2 = \rho\mu_0\mathbf{P}_{a_0}\operatorname{diag}(\mathbf{O}_{a_0}(:, z_1))\mathbf{P}_{a_1}\operatorname{diag}(\mathbf{O}_{a_1}(:, z_2))$, where ρ is a normalization constant.**
- ☐ ... $\boldsymbol{b}_2 = \rho\mu_0$, where ρ is a normalization constant.
- ☐ None of the above.

(d) **(0.8 pts.)** Consider again the POMDP $\mathcal{M}$ with initial distribution $\mu_0 = \begin{bmatrix}0.5 & 0.5 & 0\end{bmatrix}$. Suppose that, at time step $t = 0$, the agent selects action $a_0 = A$; and for $t \geq 1$, it follows a memoryless policy that maps observation u to action A and observation v to action B. A possible sequence of observations for the agent is...

- ☐ ... $\boldsymbol{z}_{1:3} = (v, u, v)$.
- ☐ ... $\boldsymbol{z}_{1:3} = (v, v, u)$.
- ☐ ... $\boldsymbol{z}_{1:3} = (v, u, u)$.
- ☒ **... $\boldsymbol{z}_{1:3} = (u, v, v)$.**

(e) **(0.8 pts.)** Suppose that R is an arbitrary binary relation on some set $\mathcal{X}$ and we write $x \overset{R}{\to} y$ to denote that x and y are related according to R. Suppose that R is *negative transitive*. This means that...

- ☐ ... for arbitrary $x, y, z \in \mathcal{X}$, $x \overset{R}{\not\to} y$ and $y \overset{R}{\not\to} z$.
- ☒ **... for arbitrary $x, y, z \in \mathcal{X}$, $x \overset{R}{\not\to} y$ and $y \overset{R}{\not\to} z$ implies that $x \overset{R}{\not\to} z$.**
- ☐ ... for arbitrary $x, y, z \in \mathcal{X}$, $x \overset{R}{\to} y$ and $y \overset{R}{\to} z$.
- ☐ ... for arbitrary $x, y \in \mathcal{X}$, either $x \overset{R}{\to} y$ or $y \overset{R}{\to} x$, but not both.

(f) **(0.8 pts.)** Consider again the POMDP $\mathcal{M}$ with initial distribution $\mu_0 = \begin{bmatrix}0.5 & 0.5 & 0.0\end{bmatrix}$, and suppose that the agent is following the FIB heuristic, with Q-function

$$Q_{\text{FIB}} = \begin{bmatrix} 0.31 & 0.31 & 0.31 \\ 0.38 & 0.50 & 0.50 \\ 1.28 & 0.44 & 0.5 \end{bmatrix}.$$

Then, the action selected by the agent at time step $t = 0$ will be...

- ☒ **... action A.**
- ☐ ... action B.
- ☐ ... action C.
- ☐ There is not enough information to determine the action at time step $t = 0$.

(g) **(0.8 pts.)** In the MLS heuristic, the action chosen by the agent...

- ☐ ... is the most voted action prescribed by the underlying MDP optimal policy.
- ☐ ... is the action with smallest value, where the value is computed as the belief-weighted average of the optimal Q-values for the underlying MDP.
- ☒ **... is the action prescribed by the MDP optimal policy in the state with the largest belief value** .
- ☐ None of the above.

(h) **(0.8 pts.)** In a POMDP, the policy graph computed at each iteration of policy iteration...

- ☐ ... is always piecewise linear and concave.
- ☐ ... corresponds to a memoryless policy.
- ☒ **... has always a finite number of nodes.**
- ☐ None of the above.

(i) **(0.8 pts.)** The UCB algorithm...

- ☒ **... can be used in the context of Monte Carlo tree search to traverse the tree and select, in each iteration, the leaf node to expand.**
- ☐ ... has a regret that is linear in the number of actions.
- ☐ ... is most adequate for adversarial multi-armed bandits.
- ☐ ... cannot be used in stochastic multi-armed bandits.

(j) **(0.8 pts.)** The UCB algorithm...

- ☐ ... relies on the principle of Occam's razor.
- ☐ ... chooses an action with a probability that is proportional to its average cost.
- ☐ ... can be used even when the costs are set adversarially.
- ☒ **None of the above.**

Question 3. (**4 pts.**)

Consider the MDP $(\mathcal{X}, \mathcal{A}, \{\mathbf{P}_a\}, c, \gamma)$, obtained from $\mathcal{M}$ by ignoring partial observability and the policy π in Question 2.(a), and repeated here for convenience:

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

(a) **(2.0 pts.)** Compute the cost-to-go associated with π.

(b) **(2.0 pts.)** Is π optimal? Explain your answer with the adequate computations.

Solution 3.

(a) We can compute the cost-to-go function as

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

For the given policy, we have that

$$\mathbf{P}_{\pi} = \begin{bmatrix} 0.5 & 0.5 & 0.0 \\ 0.0 & 0.0 & 1.0 \\ 0.0 & 0.0 & 1.0 \end{bmatrix}, \qquad c_{\pi} = \begin{bmatrix} 0.0 \\ 0.1 \\ 0.1 \end{bmatrix}.$$

Hence,

$$\boldsymbol{J}^{\pi} = \begin{bmatrix} 0.55 & -0.45 & 0.0 \\ 0.0 & 1.0 & -0.9 \\ 0.0 & 0.0 & 0.1 \end{bmatrix}^{-1} \begin{bmatrix} 0.0 \\ 0.1 \\ 0.1 \end{bmatrix}$$
$$= \begin{bmatrix} 1.82 & 0.82 & 7.36 \\ 0.0 & 1.0 & 9.0 \\ 0.0 & 0.0 & 10.0 \end{bmatrix} \begin{bmatrix} 0.0 \\ 0.1 \\ 0.1 \end{bmatrix} = \begin{bmatrix} 0.82 \\ 1.0 \\ 1.0 \end{bmatrix}.$$

(b) To check whether the policy is optimal, we perform a step of value iteration from J^{π} computed in the previous question. We have that:

$$Q^{\pi}(x,a) = c(x,a) + \gamma \sum_{x' \in \mathcal{X}} \mathbf{P}(x' \mid x,a) J^{\pi}(x'),$$

which yields:

$$\boldsymbol{Q}^{\pi} = \begin{bmatrix} 0.82 & 0.82 & 0.82 \\ 0.84 & 1.0 & 1.0 \\ 1.74 & 1.0 & 1.0 \end{bmatrix},$$

and an updated estimate

$$\boldsymbol{J}_{\text{new}} = \begin{bmatrix} 0.82 \\ 0.84 \\ 1.0 \end{bmatrix} \neq \boldsymbol{J}^{\pi}.$$

We can thus conclude that the policy π is not optimal.

Question 4. (**5.5 pts.**)

Consider once again the MDP $(\mathcal{X}, \mathcal{A}, \{\mathbf{P}_a\}, c, \gamma)$, obtained from $\mathcal{M}$ by ignoring partial observability. Suppose that the agent interacts with the MDP using a policy π that selects actions uniformly at random, and observes the following trajectory:

$$\tau = \{2, A, 0.1, 1, B, 0.0, 2, B, 0.1, 3, B, 0.1, 2\},$$

comprising $x_0, a_0, c_0, x_1, a_1, \ldots, x_4$.

(a) **(2.0 pts.)** Suppose that the agent is using 2-step SARSA to estimate Q^{π}. Assuming that the initial estimate $Q_0 \equiv 0$, use the trajectory provided to update the estimate of Q^{π} in all states where an update is possible with the data available. Use a step-size $\alpha = 1.0$.

Note: Recall that the update for 2-step SARSA is given by:

$$Q_{t+1}(x_t, a_t) = Q_t(x_t, a_t) + \alpha[c_t + \gamma c_{t+1} + \gamma^2 Q_t(x_{t+2}, a_{t+2}) - Q_t(x_t, a_t)].$$

(b) **(1.0 pt.)** Is 2-step SARSA an on-policy or off-policy algorithm? Explain your answer.

(c) **(2.0 pt.)** Suppose that we wanted to use 2-step SARSA with function approximation. Indicate how you would modify the update for 2-step SARSA to account for function approximation.

(d) **(0.5 pt.)** "*SARSA is more stable than Q-learning.*" Do you agree with this statement? Why?

Solution 4.

(a) Given the available data, we can perform two updates:

$$Q(2, A) \leftarrow Q(2, A) + \alpha(0.1 + \gamma 0.0 + \gamma^2 Q(2, B)) = 0 + 1 \times (0.1 + 0.9 \times 0.0 + 0.81 \times 0.0) = 0.1,$$

and

$$Q(1, B) \leftarrow Q(1, B) + \alpha(0.0 + \gamma 0.1 + \gamma^2 Q(3, B)) = 0 + 1 \times (0.0 + 0.9 \times 0.1 + 0.81 \times 0.0) = 0.09.$$

No further updates are possible.

(b) 2-step SARSA is an on-policy algorithm, just like regular SARSA, since the update depends on the policy being used to collect the data.

(c) The update for 2-step SARSA with function approximation would be similar to that of SARSA with function approximation, modifying the corresponding target to $c_t + \gamma c_{t+1} + \gamma^2 Q_t(x_{t+2}, a_{t+2})$ instead of $c_t + \gamma Q_t(x_{t+1}, a_{t+1})$. Assuming a parameterized Q-function with parameters θ, Q_θ, we get:

$$\theta_{t+1} = \theta_t + \alpha \nabla_\theta Q_{\theta_t}(x_t, a_t)[c_t + \gamma c_{t+1} + \gamma^2 Q_{\theta_t}(x_{t+2}, a_{t+2}) - Q_{\theta_t}(x_t, a_t)].$$

(d) The statement is true in the sense that, when using linear function approximation, the convergence of SARSA may actually be established under mild conditions, which is not true for Q-learning.