TÉCNICO LISBOA

Planning, Learning and Decision Making

MSc in Computer Science and Engineering

First examination — April 11, 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.**)

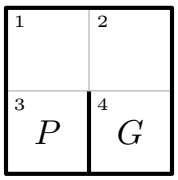

Figure 1: A taxi driver must pick up a passenger standing in the cell marked with "P" and drop it in the cell marked with "G".

Consider a taxi driver moving in the grid-world environment of Fig. 1, similar to the one you encountered in the lab assignments. 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 six actions available:

- *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.
- *Pick up.* When the taxi is in cell marked with "P" and has not yet picked the passenger, the action "Pick up" successfully picks up the passenger. Otherwise, the action "Pick up" has no effect. When the passenger is successfully picked up, it will remain in the taxi until it is dropped off in the cell marked with "G".
- *Drop off.* When the taxi is in the cell marked with "G" and the passenger is in the taxi, the action "Drop off" successfully drops the passenger. Otherwise, the action "Drop off" has no effect. When the passenger is successfully dropped off, the problem should transition to a final state indicating that the task is completed.

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" and "Drop off";
- 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.

In order to choose its actions, the taxi driver should keep track of its own position in the grid and the position of the passenger. Both elements are fully observable, so the model should be a Markov decision problem. The state-space corresponds to all possible combinations of the information just identified, leading to:

$$\mathcal{X} = \{(1, P), (1, T), (2, P), (2, T), (3, P), (3, T), (4, P), (4, T), \text{Final}\},$$

where P indicates that the passenger is in cell P, and T indicates that the passenger is in the taxi. It would also be possible to consider an additional state $(4, G)$, corresponding to the passenger being dropped in G, but this state is redundant with the final state and so it is left out.

The actions available to the agent are $\mathcal{A} = \{\mathrm{Up}, \mathrm{Down}, \mathrm{Left}, \mathrm{Right}, \mathrm{Pickup}, \mathrm{Dropoff}\}$.

Since the environment is fully observable, it is unnecessary to indicate the observations, as they correspond to the states themselves.

The transition probabilities for the actions "Up" and "Drop off" are given by:

$$\mathbf{P}_{\mathrm{Up}} = \begin{bmatrix} 1.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 1.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 1.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 1.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.8 & 0.0 & 0.0 & 0.0 & 0.2 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.8 & 0.0 & 0.0 & 0.0 & 0.2 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.8 & 0.0 & 0.0 & 0.0 & 0.2 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.8 & 0.0 & 0.0 & 0.0 & 0.2 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 1.0 \end{bmatrix},$$

$$\mathbf{P}_{\mathrm{Dropoff}} = \begin{bmatrix} 1.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 1.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 1.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 1.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 1.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 1.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 1.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 1.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 1.0 \end{bmatrix}.$$

Finally, the cost function can be given by:

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

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

- $\mathcal{X} = \{1, 2, 3, 4\}$;
- $\mathcal{A} = \{A, B\}$;
- $\mathcal{Z} = \{u, v, w\}$;
- The transition probabilities are

$$\mathbf{P}_A = \begin{bmatrix} 0.2 & 0.0 & 0.0 & 0.8 \\ 0.0 & 0.35 & 0.65 & 0.0 \\ 0.0 & 0.0 & 1.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 1.0 \end{bmatrix}; \qquad \mathbf{P}_B = \begin{bmatrix} 0.2 & 0.0 & 0.8 & 0.0 \\ 0.0 & 0.35 & 0.0 & 0.65 \\ 0.0 & 0.0 & 1.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 1.0 \end{bmatrix}.$$

- The observation probabilities are

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

- The cost function c is given by

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

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

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 that always selects action A. Then, ...

- ☐ ... the distribution $\mu_1 = \begin{bmatrix} 0 & 0 & 1 & 0 \end{bmatrix}$ is a stationary distribution for the induced Markov chain $(\mathcal{X}, \mathbf{P}_\pi)$.
- ☐ ... the distribution $\mu_2 = \begin{bmatrix} 0 & 0 & 0 & 1 \end{bmatrix}$ is a stationary distribution for the induced Markov chain $(\mathcal{X}, \mathbf{P}_\pi)$.
- ☐ ... neither μ_1 nor μ_2 above are stationary distributions for the induced Markov chain $(\mathcal{X}, \mathbf{P}_\pi)$.
- ☒ **... both μ_1 and μ_2 above are stationary distributions for the induced Markov chain $(\mathcal{X}, \mathbf{P}_\pi)$.**

(b) **(0.8 pts.)** A Markov decision process is said *ergodic* if every deterministic stationary policy induces an irreducible Markov chain. Consider once again the MDP obtained from $\mathcal{M}$ by ignoring partial observability, which corresponds to the tuple $(\mathcal{X}, \mathcal{A}, \{\mathbf{P}_a\}, c, \gamma)$. Then, ...

- ☐ The MDP is ergodic.
- ☐ The MDP is ergodic only for the uniform policy.
- ☒ **The MDP is not ergodic.**
- ☐ There is not enough information to determine whether the MDP is ergodic.

(c) **(0.8 pts.)** Consider the POMDP $\mathcal{M}$ and suppose that the state x_0 is drawn from the initial distribution $\mu_0 = \begin{bmatrix} 0.5 & 0.5 & 0 & 0 \end{bmatrix}$. Further suppose that the agent selects action A in time steps $t = 0$ and $t = 1$ and observes $z_{1:2} = \{u, u\}$. The most likely state at time step $t = 2$...

- ☐ ... is $x_2 = A$.
- ☒ **... is $x_2 = B$.**
- ☐ ... are both A and B, with equal probability.
- ☐ ... is neither A nor B.

(d) **(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 selects action A in time steps $t = 0$ and $t = 1$ and observes $z_{1:2} = \{u, u\}$. Applying the forward algorithm, we get that...

- ☐ ... $\alpha_1 = \begin{bmatrix} 0.02 & 0.06 & 0.0 & 0.0 \end{bmatrix}^\top$.
- ☐ ... $\mu_{1|1} = \begin{bmatrix} 0.1 & 0.175 & 0.0 & 0.0 \end{bmatrix}^\top$.
- ☐ ... $\beta_2 = \begin{bmatrix} 0.02 & 0.06 & 0.0 & 0.0 \end{bmatrix}^\top$.
- ☒ **...** $\mu_{2|1:2} = \begin{bmatrix} 0.25 & 0.75 & 0.0 & 0.0 \end{bmatrix}^\top$**.**

(e) **(0.8 pts.)** Suppose that $\succ$ is a strict preference relation on $\mathcal{X}$. Then, ...

- ☒ **... it is transitive and negative transitive.**
- ☐ ... it is an equivalence relation.
- ☐ ... it is a rational preference.
- ☐ None of the above.

(f) **(0.8 pts.)** Consider again the POMDP $\mathcal{M}$, now with initial distribution $\mu_0 = \begin{bmatrix}0.16 & 0.84 & 0 & 0\end{bmatrix}$, and suppose that the optimal policy for the underlying MDP is given by

$$\pi^*_{\text{MDP}} = \begin{bmatrix}1.0 & 0.0\\ 0.0 & 1.0\\ 0.5 & 0.5\\ 0.5 & 0.5\end{bmatrix}.$$

Suppose that the agent follows the AV heuristic. Then, the action selected by the agent at time step $t = 0$ will be...

- ☐ ... action A.
- ☒ **... action B.**
- ☐ Both actions are equally likely.
- ☐ The action will depend on the observation of the agent at time step $t = 1$.

(g) **(0.8 pts.)** Suppose that the α-vectors used to represent the optimal cost-to-go function for $\mathcal{M}$ are

$$\Gamma^* = \left\{\begin{bmatrix}0.12 & 6.0 & 10.0 & 0.0\end{bmatrix}^\top, \begin{bmatrix}7.32 & 0.15 & 10.0 & 0.0\end{bmatrix}^\top\right\},$$

where the first α-vector is associated with action A and the second to action B. Further consider the belief $\boldsymbol{b} = \begin{bmatrix}0.5 & 0.5 & 0 & 0\end{bmatrix}$. Then, ...

- ☒ **...** $J^*(\boldsymbol{b}) = 3.06$.
- ☐ ... $J^*(\boldsymbol{b}) = 5.99$.
- ☐ ... the optimal action at belief $\boldsymbol{b}$ is B.
- ☐ None of the above.

(h) **(0.8 pts.)** The FIB heuristic for POMDPs...

- ☐ ... assumes that there will be no partial observability from the next time step on.
- ☐ ... accounts less for partial observability than the action-voting heuristic.
- ☒ **... accounts more for partial observability than the Q-MDP heuristic.**
- ☐ None of the above.

(i) **(0.8 pts.)** The weighted majority algorithm...

- ☐ ... chooses each actions with a probability proportional to the corresponding weight.
- ☒ **... makes a number of mistakes that is logarithmic in the number of sources.**
- ☐ ... is most adequate for adversarial multi-armed bandits.
- ☐ None of the above.

(j) **(0.8 pts.)** The weight update in EXP3...

- ☐ ... assumes that at least one of the actions always has a cost of 0.
- ☐ ... uses information regarding the cost of all actions.
- ☐ ... assumes that the costs come from a fixed distribution, and thus corresponds to maintaining a running average of the costs observed thus far for each action.
- ☒ **None of the above.**

Question 3. (**2 pts.**)

Consider the MDP $(\mathcal{X}, \mathcal{A}, \{\mathbf{P}_a\}, c, \gamma)$, obtained from $\mathcal{M}$ by ignoring partial observability. Suppose that the cost-to-go associated with the policy in Question 2.(f) is given by

$$\boldsymbol{J}^* = \begin{bmatrix} 0.12 \\ 0.15 \\ 10.0 \\ 0.0 \end{bmatrix}.$$

Show that the policy is optimal.

Solution 3.

To show that the policy is optimal, we perform a step of value iteration and show that J^* remains unchanged. We have, for action A,

$$\begin{aligned} \boldsymbol{Q}^*_{:,A} &= \boldsymbol{C}_{:,A} + \gamma \mathbf{P}_A \boldsymbol{J}^* \\ &= \begin{bmatrix} 0.1 \\ 0.1 \\ 1.0 \\ 0.0 \end{bmatrix} + 0.9 \begin{bmatrix} 0.2 & 0.0 & 0.0 & 0.8 \\ 0.0 & 0.35 & 0.65 & 0.0 \\ 0.0 & 0.0 & 1.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 1.0 \end{bmatrix} \begin{bmatrix} 0.12 \\ 0.15 \\ 10.0 \\ 0.0 \end{bmatrix} = \begin{bmatrix} 0.12 \\ 5.99 \\ 10.0 \\ 0.0 \end{bmatrix}, \end{aligned}$$

and for action B,

$$\begin{aligned} \boldsymbol{Q}^*_{:,B} &= \boldsymbol{C}_{:,B} + \gamma \mathbf{P}_B \boldsymbol{J}^* \\ &= \begin{bmatrix} 0.1 \\ 0.1 \\ 1.0 \\ 0.0 \end{bmatrix} + 0.9 \begin{bmatrix} 0.2 & 0.0 & 0.8 & 0.0 \\ 0.0 & 0.35 & 0.0 & 0.65 \\ 0.0 & 0.0 & 1.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 1.0 \end{bmatrix} \begin{bmatrix} 0.12 \\ 0.15 \\ 10.0 \\ 0.0 \end{bmatrix} = \begin{bmatrix} 7.32 \\ 0.15 \\ 10.0 \\ 0.0 \end{bmatrix}. \end{aligned}$$

Taking the minimum across actions, we get

$$\boldsymbol{J}_{\text{new}} = \min \left(\begin{bmatrix} 0.12 \\ 5.99 \\ 10.0 \\ 0.0 \end{bmatrix}, \begin{bmatrix} 7.32 \\ 0.15 \\ 10.0 \\ 0.0 \end{bmatrix} \right) = \begin{bmatrix} 0.12 \\ 0.15 \\ 10.0 \\ 0.0 \end{bmatrix} = \boldsymbol{J}^*,$$

and the conclusion follows.

Question 4. (**7.5 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:

$$\tau = \{2, A, 0.1, 2, B, 0.1, 4, B, 0.0, 4, B, 0.0, 4\},$$

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

(a) **(2.0 pts.)** Suppose that the agent is using first-visit Monte-Carlo RL to estimate J^π. Assuming that the initial estimate $J_0 \equiv 0$, use the trajectory provided to update the estimate of the cost-to-go function. Use a step-size $\alpha = 1.0$.

(b) **(2.0 pt.)** Let $\tau = \{x_0, a_0, c_0, x_1, a_1, \ldots, x_T\}$ denote a trajectory obtained with an arbitrary policy π, and let $\mathbb{P}_\pi[\tau]$ denote the probability of τ under policy π. Then, given a second policy π', show that

$$\frac{\mathbb{P}_{\pi'}[\tau]}{\mathbb{P}_\pi[\tau]} = \prod_{t=0}^{T-1} \frac{\pi'(a_t \mid x_t)}{\pi(a_t \mid x_t)}.$$

Hint: Note that, given a trajectory $\tau = \{x_0, a_0, c_0, \ldots, x_T\}$,

$$\mathbb{P}_\pi[\tau] = \mathbb{P}[\mathrm{x}_0 = x_0] \prod_{t=0}^{T-1} \pi(a_t \mid x_t) \mathbf{P}(x_{t+1} \mid x_t, a_t),$$

assuming that the costs depend deterministically on the state and action.

(c) **(2.0 pt.)** *Importance sampling* is a technique that enables the computation of an expectation with respect to a distribution p using samples from a different distribution q, by taking advantage of the fact that

$$\mathbb{E}_{\mathrm{x}\sim p}[f(\mathrm{x})] = \mathbb{E}_{\mathrm{x}\sim q}\left[\frac{p(\mathrm{x})}{q(\mathrm{x})} f(\mathrm{x})\right].$$

Given a trajectory $\tau = \{x_0, a_0, c_0, x_1, a_1, \ldots, x_T\}$, let $L(\tau)$ denote the total discounted cost of trajectory τ, i.e.,

$$L(\tau) = \sum_{t=0}^{T-1} \gamma^t c_t.$$

Taking into consideration that

$$J^\pi(x_0) \approx \mathbb{E}_\pi[L(\tau)],$$

use the idea of importance sampling and the result in (b) to indicate how the Monte Carlo RL update can be adapted to compute the cost-to-go for policy π' using a trajectory obtained using policy π.

(d) **(1.5 pt.)** What are *exploring starts*? Why are they necessary in Monte Carlo RL?

Solution 4.

(a) First-visit MC means that we only update the states visited in the trajectory once, using the total discounted cost observed after the first visit to each one. In the observed trajectory, we visit only states 2 (at time steps $t = 0$ and $t = 1$) and 4 (at time steps $t = 2$, $t = 3$, and $t = 4$). For state 2, we get

$$L_0(\tau) = c_0 + \gamma c_1 + \gamma^2 c_2 + \gamma^3 c_3 = 0.1 + 0.9 \times 0.1 + 0.9^2 \times 0.0 + 0.9^3 \times 0.0 = 0.19.$$

For state 4, we get:

$$L_2(\tau) = c_2 + \gamma c_3 = 0.0 + 0.9 \times 0.0 = 0.0$$

We thus get the updates:

$$J(2) \leftarrow J(2) + \alpha(L_0(\tau) - J(2)) = 0.0 + 1.0 \times (0.19 - 0.0) = 0.19;$$
$$J(4) \leftarrow J(4) + \alpha(L_2(\tau) - J(4)) = 0.0 + 1.0 \times (0.0 - 0.0) = 0.0.$$

(b) Using the expression provided in the hint, we have that:

$$\frac{\mathbb{P}_{\pi'}[\tau]}{\mathbb{P}_{\pi}[\tau]} = \frac{\mathbb{P}[\mathrm{x}_0 = x_0]}{\mathbb{P}[\mathrm{x}_0 = x_0]} \prod_{t=0}^{T-1} \frac{\pi'(a_t \mid x_t)}{\pi(a_t \mid x_t)} \cdot \frac{\mathbf{P}(x_{t+1} \mid x_t, a_t)}{\mathbf{P}(x_{t+1} \mid x_t, a_t)} = \prod_{t=0}^{T-1} \frac{\pi'(a_t \mid x_t)}{\pi(a_t \mid x_t)},$$

as desired.

(c) We have that

$$J^{\pi'}(x_0) \approx \mathbb{E}_{\pi'}[L_0(\tau)],$$

where τ is a trajectory sampled using π'. Using importance sampling, however, we have that

$$J^{\pi'}(x_0) \approx \mathbb{E}_{\pi}\left[\frac{\mathbb{P}_{\pi'}[\tau]}{\mathbb{P}_{\pi}[\tau]} L_0(\tau)\right],$$

where now τ is a trajectory sampled using π. The update for MCRL using importance sampling thus comes

$$J(x_0) \leftarrow J(x_0) + \alpha\left(\frac{\mathbb{P}_{\pi'}[\tau]}{\mathbb{P}_{\pi}[\tau]} L_0(\tau) - J(x_0)\right)$$

which, using the result from (b), yields

$$J(x_0) \leftarrow J(x_0) + \alpha\left(\prod_{t=0}^{T-1} \frac{\pi'(a_t \mid x_t)}{\pi(a_t \mid x_t)} L_0(\tau) - J(x_0)\right).$$

(d) *Exploring starts* means that we use the starting state/state-action pair of a trajectory to ensure sufficient exploration. This is a fundamental requirement of Monte Carlo methods, since these methods rely on trajectories obtained using fixed policies and, as such, sufficient visitation to every state/state-action pair cannot be ensured unless if exploring starts are available. The use of exploring starts is a heavy requirement in the sense that it essentially implies that the system can be reset to any state/state-action pair at will, which may not be possible in many real-world problems.