Learning and Decision Making 2016-2017x

MSc in Computer Science and Engineering

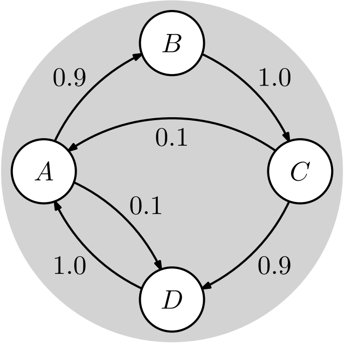
First test – April 4, 2017

Instructions

- You have 90 minutes to complete the test.
- Make sure that your test has a total of 11 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 5 questions, with a maximum score of 19 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 Markov chain:



(a) **(1 pt.)** Write the transition probability matrix for the chain.

(b) **(0.5 pt.)** Indicate in the transition diagram the communicating classes for the chain. Is the chain irreducible? Why?

(c) **(0.5 pt.)** Compute the period of state C. Is the chain aperiodic? Why?

(d) **(1 pt.)** Which of the following possibilities best describes the stationary distribution for the chain? Briefly explain your reasoning.

i) $\boldsymbol{\mu} = [0.5385 \quad 0.4847 \quad 0.4847 \quad 0.4901]$;

ii) $\boldsymbol{\mu} = [0.2696 \quad 0.2426 \quad 0.2426 \quad 0.2453]$;

iii) $\boldsymbol{\mu} = [0.25 \quad 0.25 \quad 0.25 \quad 0.25]$;

iv) The chain does not have a stationary distribution.

Solution 1.

(a) The transition probability matrix can be extracted directly from the probabilities in the transition diagram, yielding

$$\mathbf{P} = \begin{bmatrix} 0.0 & 0.9 & 0.0 & 0.1 \\ 0.0 & 0.0 & 1.0 & 0.0 \\ 0.1 & 0.0 & 0.0 & 0.9 \\ 1.0 & 0.0 & 0.0 & 0.0 \end{bmatrix}.$$

(b) The single communication class is indicated as the shaded circle. Since there is a single communicating class, the chain is irreducible.

(c) Using the Chapman-Kolmogorov inequality, we have that

$$\mathbb{P}\left[\mathrm{x}_3 = C \mid \mathrm{x}_0 = C\right] \geq 0.1 \times 0.9 = 0.09 > 0$$

and

$$\mathbb{P}\left[\mathrm{x}_4 = C \mid \mathrm{x}_0 = C\right] \geq 0.9 \times 0.9 = 0.19 > 0.$$

Since $\gcd\{3, 4\} = 1$, we can conclude that the period of C is 1. Moreover, since the chain is irreducible, all states have the same period and thus the chain is aperiodic.

(d) The hypothesis (i) is not a distribution, so it cannot be the stationary distribution. Hypothesis (iv) is also false, since the chain is irreducible and aperiodic. As for the hypothesis (iii), it yields

$$\boldsymbol{\mu}\mathbf{P} = \begin{bmatrix} 0.275 & 0.225 & 0.5 & 0.25 \end{bmatrix},$$

thus it is not stationary. Finally, hypothesis (ii) yields

$$\boldsymbol{\mu}\mathbf{P} = \begin{bmatrix} 0.2696 & 0.2426 & 0.2426 & 0.2453 \end{bmatrix}.$$

Therefore, the stationary distribution is (ii).[1]

Question 2. (4 pts.)

Daniel Atlas hides a ball under one of three cups and explains the audience that he will shuffle the cups by successively selecting two of them and switching their positions. From his explanation, you realize that half of the times he switches the two cups without the ball, and the other half he switches the cup with the ball with one of the other two.

Once the explanation is over, Atlas places the ball under the central cup and starts shuffling the cups around furiously. You are more or less able to follow where the ball is, but after each movement there is a probability of 0.4 that you perceive the ball in the wrong cup—i.e., think that the ball is in one cup when it is in a different one.

(a) **(2 pts.)** From the audience's perspective, the motion of the ball can be modeled as an HMM. Describe the HMM model, indicating all corresponding parameters.

(b) **(2 pt.)** Suppose that, after each of the two initial motions, you note, respectively,

1. The ball on the middle;
2. The ball on the left.

What is the probability that, after the motions, the ball is under the middle cup?

Note: The observations above should be considered as resulting from performed movements and, therefore, take place at steps $t = 1$ and $t = 2$. On the other hand, since the audience observed where the ball was placed at time $t = 0$, you need not to consider an initial observation.

**** If you did not do part (a) ****, consider instead the HMM $(\mathcal{X}, \mathcal{Z}, \mathbf{P}, \mathbf{O})$ where:

- $\mathcal{X} = \{1, 2, 3\}$
- $\mathcal{Z} = \{a, b, c\}$

[1] The actual stationary distribution is given by

$$\boldsymbol{\mu}\mathbf{P} = \begin{bmatrix} 0.269541778975741 & 0.242587601078167 & 0.242587601078167 & 0.245283018867925 \end{bmatrix}$$

which corresponds to the provided answer but for small roundoff errors.

- The transition and observation probability matrices are

$$\mathbf{P} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix} \qquad \mathbf{O} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} \end{bmatrix}.$$

Compute the most likely state at time $t = 3$ given that $\mathrm{x}_0 = 1$ and $\mathrm{z}_1 = \mathrm{z}_2 = c$. **If you did part (a) and instead solve this HMM, you will lose points**.

Solution 2.

(a) The information of interest is the position of the ball (in which of the cups it is). The motion of the cups can be seen as a motion of the ball. Therefore, the HMM can be described as a tuple $(\mathcal{X}, \mathcal{Z}, \mathbf{P}, \mathbf{O})$ where

- $\mathcal{X} = L, M, R$, where L stands for the left cup, M for the middle cup and R for the right cup.
- Since the audience can follow the position of the ball, $\mathcal{Z} = \mathcal{X} = \{L, M, R\}$.
- The transition probabilities essentially describe how the ball moves. The transition and observation probability matrices thus come:

$$\mathbf{P} = \begin{bmatrix} 0.5 & 0.25 & 0.25 \\ 0.25 & 0.5 & 0.25 \\ 0.25 & 0.25 & 0.5 \end{bmatrix}, \qquad \mathbf{O} = \begin{bmatrix} 0.6 & 0.2 & 0.2 \\ 0.2 & 0.6 & 0.2 \\ 0.2 & 0.2 & 0.6 \end{bmatrix}.$$

(b) We use the forward algorithm to determine the desired probability. Since there is no initial observation and the ball is placed in the middle cup,

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

Continuing with the algorithm,

$$\boldsymbol{\alpha}_1^\top = \boldsymbol{\alpha}_0^\top \mathbf{P} \mathrm{diag}(\mathbf{O}_{:,M}) = \begin{bmatrix} 0.25 & 0.5 & 0.25 \end{bmatrix} \begin{bmatrix} 0.2 & 0 & 0 \\ 0 & 0.6 & 0 \\ 0 & 0 & 0.2 \end{bmatrix} = \begin{bmatrix} 0.05 & 0.3 & 0.05 \end{bmatrix}.$$

and

$$\boldsymbol{\alpha}_2^\top = \boldsymbol{\alpha}_1^\top \mathbf{P} \mathrm{diag}(\mathbf{O}_{:,L}) = \begin{bmatrix} 0.1125 & 0.175 & 0.1125 \end{bmatrix} \begin{bmatrix} 0.6 & 0 & 0 \\ 0 & 0.2 & 0 \\ 0 & 0 & 0.2 \end{bmatrix} = \begin{bmatrix} 0.0675 & 0.035 & 0.0225 \end{bmatrix}.$$

After normalizing, we finally have:

$$\boldsymbol{\mu}_{2|0:2} = \begin{bmatrix} 0.54 & 0.28 & 0.18 \end{bmatrix}$$

and the desired probability is, thus, 0.28.

◇

As for the alternative problem, the solution is obtained in a similar manner. We have:

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

Continuing with the algorithm,

$$\boldsymbol{\alpha}_1^\top = \boldsymbol{\alpha}_0^\top \mathbf{P} \mathrm{diag}(\mathbf{O}_{:,c}) = \begin{bmatrix} 0.5 & 0.5 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0.5 & 0 \\ 0 & 0 & 0.5 \end{bmatrix} = \begin{bmatrix} 0 & 0.25 & 0 \end{bmatrix}.$$

and

$$\boldsymbol{\alpha}_2^\top = \boldsymbol{\alpha}_1^\top \mathbf{P} \mathrm{diag}(\mathbf{O}_{:,c}) = \begin{bmatrix} 0 & 0.25 & 0.125 \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0.5 & 0 \\ 0 & 0 & 0.5 \end{bmatrix} = \begin{bmatrix} 0 & 0.0625 & 0.0625 \end{bmatrix}.$$

After normalizing, we finally have:

$$\boldsymbol{\mu}_{2|0:2} = \begin{bmatrix} 0 & 0.5 & 0.5 \end{bmatrix}$$

and we conclude that states 2 and 3 are both equally likely.

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

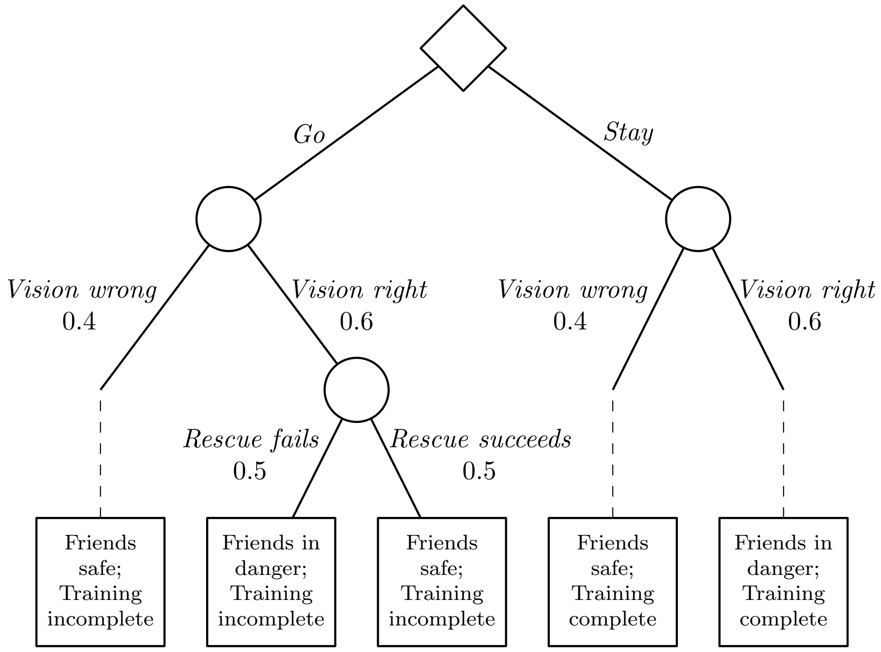
Luke Skywalker is facing a dilemma. In the middle of his Jedi training he had a vision in which he watched the suffering of his friend Han Solo at the hands of Darth Vader. Therefore, he must decide between going to Cloud City and try to save his friend or stay at Dagobah and complete his training.

He knows that his visions are wrong with a probability of 0.4. Moreover, even if he goes to Cloud City, there is a 0.5 probability that he will fail to save his friend. Finally, if he does decide to go to Cloud City, he knows that his Jedi training will be incomplete. We know that Luke's preferred outcome is to *have his friends safe and his training complete*. If both are not possible, he *prefers to have his friends safe* than to *have his training complete*. What he wants to avoid at all costs is to *have his friends suffering and his training incomplete*.

Knowing that Luke decided to go to Cloud City to try to save his friend, draw the decision tree for Luke's decision process and indicate a utility function consistent with the preferences above that explains Luke's decision. Of the outcomes described above, assume that the utility for the first is maximal (1) and the utility for the last is minimal (0).

Solution 3.

The decision tree for the problem is provided below.


Go
Stay
Vision wrong
0.4
Vision right
0.6
Vision wrong
0.4
Vision right
0.6
Rescue fails
0.5
Rescue succeeds
0.5
Friends safe; Training incomplete
Friends in danger; Training incomplete
Friends safe; Training incomplete
Friends safe; Training complete
Friends in danger; Training complete

Let SC denote the outcome where Luke's friends are safe and his training complete; SI denote the outcome where his friends are safe but the training incomplete; DC the outcome where his friends are in danger but his training complete; and DI the option where his friends are in danger and his training incomplete.

We know that $SC \succ SI \succ DC \succ DI$, and $u(SC) = 1$ and $u(DI) = 0$. The expected utility of alternative "Go to Cloud City" is, therefore,

$$Q(\textit{Go}) = 0.4 \times u(SI) + 0.6 \times 0.5 \times u(DI) + 0.6 \times 0.5 \times u(SI) = 0.7u(SI)$$

and the expected utility of alternative "Stay in Dagobah" is, in turn,

$$Q(\textit{Stay}) = 0.4 \times u(SC) + 0.6 \times u(DC) = 0.4 + 0.6u(DC).$$

Since Luke went to Cloud City, we have that $Q(\textit{Go}) > Q(\textit{Stay})$ or, equivalently, $0.7u(SI) > 0.4 + 0.6u(DC)$. Letting, for example, $u(DC) = 0.1$, this means that $u(SI) > 0.657$. Finally, the utility function

$$\boldsymbol{u}^\top = \begin{bmatrix} 1 & 0.7 & 0.1 & 0 \end{bmatrix}$$

explains Luke's behavior.

Question 4. (8 pts.)

Consider the MDP with parameters

- $\mathcal{X} = \{A, B, C, D\}$;
- $\mathcal{A} = \{a, b\}$;

- The transition probability matrices are:

$$\mathbf{P}_a = \begin{bmatrix} 0.5 & 0.25 & 0.25 & 0 \\ 0.25 & 0.5 & 0.25 & 0 \\ 0.25 & 0.25 & 0.5 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \qquad \mathbf{P}_b = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix};$$

- The cost function is:

$$\mathbf{C} = \begin{bmatrix} 0.1 & 1 \\ 0.1 & 0 \\ 0.1 & 1 \\ 0 & 0 \end{bmatrix};$$

- $\gamma = 0.0$.

(a) **(1 pt.)** Draw the transition diagram for the MDP.

(b) **(1.5 pts.)** Consider the policy

$$\pi = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0.5 & 0.5 \end{bmatrix}.$$

Compute the cost-to-go function associated with π. Indicate all relevant computations.

(c) **(1.5 pt.)** Using the uniform policy as the initial policy, compute the optimal policy for the MDP using policy iteration. Indicate all relevant computations.

Note: Recall that the uniform policy selects randomly between all actions with equal probability.

(d) **(2 pt.)** Suppose now that $\gamma = 0.99$ and that the cost-to-go function associated with the policy π from part (b) is

$$\boldsymbol{J}^{\pi} = \begin{bmatrix} 0.3883 \\ 0 \\ 0.3883 \\ 0 \end{bmatrix}.$$

Show that the policy π optimal, indicating all relevant computations.

(e) **(2 pts.)** Suppose now that you construct a POMDP $(\mathcal{X}, \mathcal{A}, \mathcal{Z}, \mathbf{P}, \mathbf{O}, c, \gamma)$ from the MDP above, where $\gamma = 0.99$ and

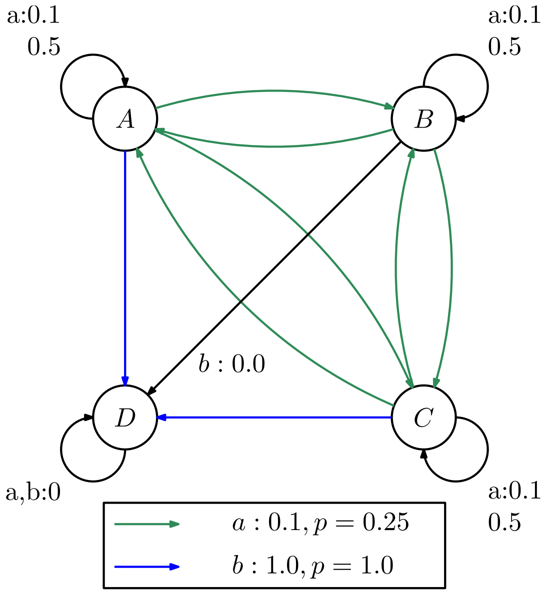
- $\mathcal{Z} = \{A, B, C, D\}$;
- The observation probability matrices are:

$$\mathbf{O}_a = \mathbf{O}_b = \begin{bmatrix} 0.5 & 0.25 & 0.25 & 0 \\ 0.25 & 0.5 & 0.25 & 0 \\ 0.25 & 0.25 & 0.5 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Indicate the action prescribed by the Q-MDP heuristic when $\mathbf{b}_t = [0.3 \quad 0.5 \quad 0.2 \quad 0]$. Indicate all relevant computations.

Solution 4.

(a) The transition diagram for the MDP is depicted below where, to avoid cluttering the diagram, we use a color code to distinguish the properties of each transition.



(b) The cost to go associated with a policy π can be computed as

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

In our case, since $\gamma = 0$, we have that

$$\boldsymbol{J}^\pi = \boldsymbol{c}_\pi = \begin{bmatrix} 0.1 \\ 0 \\ 0.1 \\ 0 \end{bmatrix}.$$

(c) We have:

$$\pi^{(0)} = \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \\ 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix}$$

and

$$\boldsymbol{J}^{(0)} = \boldsymbol{c}_{\pi^{(0)}} = \begin{bmatrix} 0.55 \\ 0.05 \\ 0.55 \\ 0 \end{bmatrix}.$$

Computing the associated Q-function yields

$$Q^{(0)} = \mathbf{C} + \gamma \mathbf{P} \boldsymbol{J} = \mathbf{C} = \begin{bmatrix} 0.1 & 1 \\ 0.1 & 0 \\ 0.1 & 1 \\ 0 & 0 \end{bmatrix}$$

Yielding the policy

$$\pi^{(1)} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0.5 & 0.5 \end{bmatrix}.$$

The corresponding cost-to-go was computed before and yields:

$$\boldsymbol{J}^{\pi} = \boldsymbol{c}_{\pi} = \begin{bmatrix} 0.1 \\ 0 \\ 0.1 \\ 0 \end{bmatrix}.$$

Then,

$$Q^{(1)} = \begin{bmatrix} 0.1 & 1 \\ 0.1 & 0 \\ 0.1 & 1 \\ 0 & 0 \end{bmatrix},$$

yielding the policy

$$\pi^{(2)} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0.5 & 0.5 \end{bmatrix}.$$

The algorithm terminates.

(d) To show that the policy is optimal, we can, for example, perform one step of value iteration to get:

$$\begin{aligned} \boldsymbol{J}_{\text{new}} &= \min(\mathbf{C} + \gamma\,[\mathbf{P}_a\boldsymbol{J}_{\text{old}} \quad \mathbf{P}_b\boldsymbol{J}_{\text{old}}]) \\ &= \min\left(\begin{bmatrix} 0.1 & 1 \\ 0.1 & 0 \\ 0.1 & 1 \\ 0 & 0 \end{bmatrix} + 0.99 \begin{bmatrix} 0.2912 & 0 \\ 0.1941 & 0 \\ 0.2912 & 0 \\ 0 & 0 \end{bmatrix}\right) \\ &= \min\left(\begin{bmatrix} 0.3883 & 1 \\ 0.2922 & 0 \\ 0.3883 & 1 \\ 0 & 0 \end{bmatrix}.\right) = \begin{bmatrix} 0.3883 \\ 0 \\ 0.3883 \\ 0 \end{bmatrix}. \end{aligned}$$

Thus $J_{\text{old}} = J^*$ and the policy π is optimal.

(e) In part (d), we have computed the optimal Q-function for the underlying MDP (henceforth denoted Q_{MDP}). Then,

$$\begin{aligned} \pi_{\text{Q-MDP}}(\boldsymbol{b}_t) &= \operatorname*{argmin}_{a\in\mathcal{A}} \boldsymbol{b}_t \cdot \mathbf{Q}_{\text{MDP}} \\ &= \operatorname*{argmin}_{a\in\mathcal{A}} \begin{bmatrix} 0.3 & 0.5 & 0.2 & 0 \end{bmatrix} \begin{bmatrix} 0.3883 & 1 \\ 0.2922 & 0 \\ 0.3883 & 1 \\ 0 & 0 \end{bmatrix} \\ &= \operatorname*{argmin}_{a\in\mathcal{A}} \begin{bmatrix} 0.34 & 0.5 \end{bmatrix}. \end{aligned}$$

The Q-MDP heuristic thus prescribes action a.

Question 5. (2 pts.)

Explain what the belief-MDP is and its relevance for the solution of POMDPs.

Solution 5.

For a given POMDP, the belief MDP is an equivalent MDP model whose state space is the set of possible distributions over the state space of the original POMDP (also known as the belief space). Belief MDPs are thus a “bridge” between the theory of MDPs and that of POMDPs, allowing us to use MDP solution methods such as value and policy iteration to solve POMDPs.