TÉCNICO LISBOA

Planning, Learning and Decision Making

MSc in Computer Science and Engineering

Final examination – February 4, 2020

Instructions

- You can submit either just one of the tests or the whole exam. You have 90 minutes to complete a test, or 180 to complete the exam. If, after 90 minutes, you do not submit a test, you will be graded for the whole exam.
- Make sure that your test has a total of 14 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 exam has a total of 9 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.

= **BEGINNING OF TEST 1** =

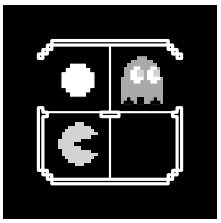


Figure 1: Simplified version of the Pacman game. Pacman (on the bottom) must avoid the moving ghosts while trying to capture the large pellet on the top.

Question 1. (**1.5 pts.**)
Consider the following simplified version of the Pacman game, depicted in Fig. 1. An agent controls the Pacman character (on the bottom-left) that must reach the large pellet (the large circle on the top-left) while avoiding the ghost (on the top-right). At each step, the ghost moves randomly to one of its adjacent cells. The Pacman character, however, cannot observe the position of the ghost, unless if the two are in the same horizontal corridor.

At any moment, Pacman can select any of four possible actions (up, down, left, right), which moves the character to the adjacent cell unless if a wall is in the way, in which case Pacman remains in the same cell. The world is "circular" in that a character (Pacman or ghost) exiting on the top-left cell reappears on the top-right cell and vice-versa.

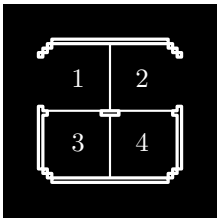
The agent loses the game if the ghost and Pacman stand in the same cell, and wins the game if it reaches the cell with the pellet. Upon termination (either winning or losing), the game resets to a random non-terminal game state.

Describe the decision problem faced by the agent controlling Pacman using a POMDP. In particular, you should indicate:

- The state space;
- The action space;
- The observation space;
- The transition probabilities for action "move right";
- The observation probabilities for action "move right";
- An immediate cost function that translates the goals of the game as described above. The cost function should be as simple as possible.

Solution 1.

Numbering the cells in the diagram as



the problem can be modeled as a POMDP $(\mathcal{X}, \mathcal{A}, \mathcal{Z}, \{\mathbf{P}_a\}, \{\mathbf{O}_a\}, c, \gamma)$, with

- $\mathcal{X} = \{(p, g), p = 1, \ldots, 4, g = 1, \ldots, 4\}$. p corresponds to the position of Pacman and g to the position of the ghost.
- $\mathcal{A} = \{u, d, l, r\}$, corresponding to the actions of the agent in the 4 directions.
- $\mathcal{Z} = \{(1,1), (1,2), (2,1), (2,2), (3,3), (3,4), (4,3), (4,4), (1,\emptyset), (2,\emptyset), (3,\emptyset), (4,\emptyset)\}$, where $\emptyset$ means that the ghost is not observable.
- As for the transition probabilities for the "move right" action, let $\mathrm{step}(p)$ denote the function

$$\mathrm{step}(p) = \begin{cases} 1 & \text{if } p = 2 \\ 2 & \text{if } p = 1 \\ 4 & \text{otherwise.} \end{cases}$$

  Additionally, let $\mathcal{X}_T$ denote the set of terminal states, i.e.,

$$\mathcal{X}_T = \{(1,1), (1,2), (1,3), (1,4), (2,2), (3,3), (4,4)\}.$$

  Then, the transition probabilities can be written as

$$\mathbf{P}((p', g') \mid (p, g), r) = \begin{cases} \frac{1}{9} & \text{if } (p, g) \in \mathcal{X}_T \text{ and } (p', g') \notin \mathcal{X}_T \\ 0.5 & \text{if } (p, g) \notin \mathcal{X}_T \text{ and } g \in \{1, 4\} \text{ and } g' \in \{2, 3\} \text{ and } p' = \mathrm{step}(p) \\ 0.5 & \text{if } (p, g) \notin \mathcal{X}_T \text{ and } g \in \{2, 3\} \text{ and } g' \in \{1, 4\} \text{ and } p' = \mathrm{step}(p) \\ 0.0 & \text{otherwise.} \end{cases}$$

- Similarly, for the observation probabilities, we get

$$\mathbf{O}((p', g') \mid (p, g), r) = \begin{cases} 1.0 & \text{if } p \in \{1, 2\} \text{ and } g \in \{1, 2\} \text{ and } (p', g') = (p, g) \\ 1.0 & \text{if } p \in \{3, 4\} \text{ and } g \in \{3, 4\} \text{ and } (p', g') = (p, g) \\ 1.0 & \text{if } p \in \{1, 2\} \text{ and } g \in \{3, 4\} \text{ and } (p', g') = (p, \emptyset) \\ 1.0 & \text{if } p \in \{3, 4\} \text{ and } g \in \{1, 2\} \text{ and } (p', g') = (p, \emptyset) \\ 0.0 & \text{otherwise.} \end{cases}$$

- Finally, a possible cost function penalizes whenever Pacman and the ghost are in the same cell, and is minimal when Pacman gets the pellet, i.e.,

$$c((p, g), a) = \begin{cases} 1 & \text{if } p = g \\ 0 & \text{if } p \neq g \text{ and } p = 1 \\ 0.1 & \text{otherwise.} \end{cases}$$

Question 2. (**1.5 pts.**)

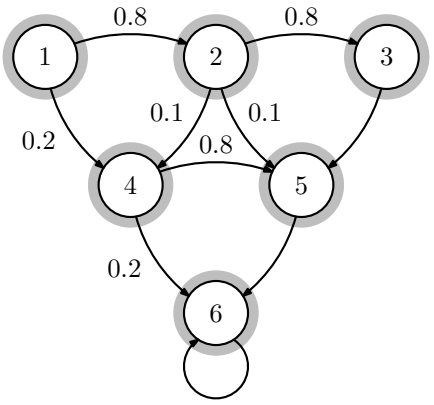


Figure 2: 6-state Markov chain.

Consider the Markov chain depicted in the transition diagram of Fig. 2.

(a) **(0.5 pts.)** Indicate the state-space and transition probabilities for the chain.

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

(c) **(0.5 pts.)** Select the correct alternative from those below, explaining your selection.

☐ The stationary distribution (with a precision of 10^{-3}) is

$$\mu^* = \begin{bmatrix} 0.408 & 0.408 & 0.408 & 0.408 & 0.408 & 0.408 \end{bmatrix}.$$

☐ The stationary distribution is

$$\mu^* = \begin{bmatrix} \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \end{bmatrix}.$$

☒ The stationary distribution is

$$\mu^* = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

☐ The chain does not possess a stationary distribution, since it is not irreducible.

Solution 2.

(a) The states for the chain correspond to the nodes in the graph, yielding $\mathcal{X} = \{1, 2, 3, 4, 5, 6\}$. As for the transition probabilities, we can get them directly from the edge labels, to get

$$\mathbf{P} = \begin{bmatrix} 0.0 & 0.8 & 0.0 & 0.2 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.8 & 0.1 & 0.1 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 1.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.8 & 0.2 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 1.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 1.0 \end{bmatrix}.$$

(b) The chain is not irreducible since there are multiple communicating classes.

(c) We have that

$$\begin{bmatrix}0 & 0 & 0 & 0 & 0 & 1\end{bmatrix}\begin{bmatrix}0.0 & 0.8 & 0.0 & 0.2 & 0.0 & 0.0\\ 0.0 & 0.0 & 0.8 & 0.1 & 0.1 & 0.0\\ 0.0 & 0.0 & 0.0 & 0.0 & 1.0 & 0.0\\ 0.0 & 0.0 & 0.0 & 0.0 & 0.8 & 0.2\\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 1.0\\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 1.0\end{bmatrix} = \begin{bmatrix}0 & 0 & 0 & 0 & 0 & 1\end{bmatrix},$$

meaning that the third alternative is a stationary distribution for the chain (the first two are not left eigenvectors of $\mathbf{P}$ and the last one is clearly false).

Question 3. (**1.5 pts.**)
Consider the HMM $(\mathcal{X}, \mathcal{Z}, \mathbf{P}, \mathbf{O})$ such that

- $\mathcal{X} = \{1, 2, 3, 4\}$;
- $\mathcal{Z} = \{u, v\}$;
- The transition and observation probabilities are given by

$$\mathbf{P} = \begin{bmatrix}0.2 & 0.6 & 0.15 & 0.05\\ 0.6 & 0.2 & 0.05 & 0.15\\ 0.05 & 0.15 & 0.6 & 0.2\\ 0.15 & 0.05 & 0.2 & 0.6\end{bmatrix}, \qquad \mathbf{O} = \begin{bmatrix}0.8 & 0.2\\ 0.0 & 1.0\\ 1.0 & 0.0\\ 0.2 & 0.8\end{bmatrix},$$

where the states and observations are ordered as in $\mathcal{X}$ and $\mathcal{Z}$ above, respectively.

Assuming that the HMM departs from the uniform initial distribution and the agent makes the sequence of observations $\boldsymbol{z}_{1:2} = \{u, v\}$, compute the most likely sequence of states between $t = 0$ and $t = 2$.

Solution 3.

Since there is no initial observation, we have:

$$\boldsymbol{m}_0 = \begin{bmatrix}0.25 & 0.25 & 0.25 & 0.25\end{bmatrix}^\top.$$

At time step $t = 1$, we get

$$\begin{aligned}\boldsymbol{m}_1 &= \operatorname{diag}(\mathbf{O}_{:,u}) \max\{\mathbf{P}^\top \operatorname{diag}(m_0)\}\\ &= \begin{bmatrix}0.8 & 0.0 & 0.0 & 0.0\\ 0.0 & 0.0 & 0.0 & 0.0\\ 0.0 & 0.0 & 1.0 & 0.0\\ 0.0 & 0.0 & 0.0 & 0.2\end{bmatrix} \cdot \max \begin{bmatrix}0.05 & 0.15 & 0.0125 & 0.0375\\ 0.15 & 0.05 & 0.0375 & 0.0125\\ 0.0375 & 0.0125 & 0.15 & 0.05\\ 0.0125 & 0.0375 & 0.05 & 0.15\end{bmatrix}\\ &= \begin{bmatrix}0.12 & 0.0 & 0.15 & 0.03\end{bmatrix}^\top,\end{aligned}$$

and

$$\boldsymbol{i}_1 = \text{argmax}\{\mathbf{P}^\top \text{diag}(m_0)\} = \begin{bmatrix} 2 & 1 & 3 & 4 \end{bmatrix}^\top .$$

At time step $t = 2$, we repeat the same computations to get

$$\boldsymbol{m}_2 = \text{diag}(\mathbf{O}_{:,u}) \max\left\{\mathbf{P}^\top \text{diag}(m_1)\right\} = \begin{bmatrix} 0.0048 & 0.072 & 0.0 & 0.024 \end{bmatrix}^\top ,$$

and

$$\boldsymbol{i}_2 = \text{argmax}\left\{\mathbf{P}^\top \text{diag}(m_1)\right\} = \begin{bmatrix} 1 & 1 & 3 & 3 \end{bmatrix}^\top .$$

From $\boldsymbol{m}_2$, it follows that $x_2^* = 2$, implying that $x_1^* = 1$ and $x_0^* = 2$. The most likely sequence is, thus, $\boldsymbol{x}_{0:2}^* = \{2, 1, 2\}$.

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

Consider an MDP $(\mathcal{X}, \mathcal{A}, \{\mathbf{P}_a\}, c, \gamma)$, where $\mathcal{X} = \{1, 2, 3\}$, $\mathcal{A} = \{A, B\}$, $\gamma = 0.9$ and

- The transition probabilities are given by

$$\mathbf{P}_A = \begin{bmatrix} 0.1 & 0.8 & 0.1 \\ 0.1 & 0.1 & 0.8 \\ 0.6 & 0.2 & 0.2 \end{bmatrix}, \qquad \mathbf{P}_B = \begin{bmatrix} 1.0 & 0.0 & 0.0 \\ 0.6 & 0.4 & 0.0 \\ 0.0 & 0.6 & 0.4 \end{bmatrix},$$

  where the states are ordered as in $\mathcal{X}$ above;

- The immediate cost function is given by

$$\boldsymbol{c} = \begin{bmatrix} 1.0 & 1.0 \\ 0.0 & 0.0 \\ 1.0 & 1.0 \end{bmatrix},$$

  where the actions are ordered as in $\mathcal{A}$ above.

(a) **(1.5 pt.)** Suppose that, for some policy π,

$$\mathbf{P}_\pi = \begin{bmatrix} 1.0 & 0.0 & 0.0 \\ 0.6 & 0.4 & 0.0 \\ 0.3 & 0.4 & 0.3 \end{bmatrix}, \qquad \boldsymbol{c}_\pi = \begin{bmatrix} 1.0 \\ 0.0 \\ 1.0 \end{bmatrix}.$$

Compute the cost-to-go associated with policy π.

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

$$\mathbf{A} = \begin{bmatrix} a & 0 & 0 \\ b & c & 0 \\ d & e & f \end{bmatrix}, \qquad \mathbf{A}^{-1} = \begin{bmatrix} \frac{1}{a} & 0 & 0 \\ -\frac{b}{ac} & \frac{1}{c} & 0 \\ \frac{be-cd}{acf} & -\frac{e}{df} & \frac{1}{f} \end{bmatrix}.$$

(b) **(1.0 pts.)** Compute the greedy policy with respect to the cost-to-go function from (a).

Note: If you did not answer (a) **and only in that case**, use

$$\boldsymbol{J}^\pi = \begin{bmatrix} 10.0 \\ 8.0 \\ 9.0 \end{bmatrix}.$$

(c) **(0.5 pts.)** Do you believe that the policy π is optimal? Explain your reasoning.

Solution 4.

(a) The cost-to-go, J^π, is given by

$$\begin{aligned}\boldsymbol{J}^\pi &= (\boldsymbol{I} - \gamma \mathbf{P}_\pi)^{-1} \boldsymbol{c}_\pi \\ &= \begin{bmatrix} 10.0 & 0.0 & 0.0 \\ 8.44 & 1.56 & 0.0 \\ 7.86 & 0.77 & 1.37 \end{bmatrix} \begin{bmatrix} 1.0 \\ 0.0 \\ 1.0 \end{bmatrix} \\ &= \begin{bmatrix} 10.0 \\ 8.44 \\ 9.23 \end{bmatrix}.\end{aligned}$$

(b) We compute the Q-values for each action $a \in \mathcal{A}$. We have:

$$\begin{aligned}\boldsymbol{Q}^\pi_{:,A} &= \boldsymbol{c}_{:,A} + \gamma \mathbf{P}_A \boldsymbol{J}^\pi \\ &= \begin{bmatrix} 1.0 \\ 0.0 \\ 1.0 \end{bmatrix} + 0.9 \begin{bmatrix} 0.1 & 0.8 & 0.1 \\ 0.1 & 0.1 & 0.8 \\ 0.6 & 0.2 & 0.2 \end{bmatrix} \begin{bmatrix} 10.0 \\ 8.44 \\ 9.23 \end{bmatrix} \\ &= \begin{bmatrix} 8.81 \\ 8.30 \\ 9.58 \end{bmatrix}\end{aligned}$$

Similarly, for action B,

$$\begin{aligned}\boldsymbol{Q}^\pi_{:,B} &= \boldsymbol{c}_{:,B} + \gamma \mathbf{P}_B \boldsymbol{J}^\pi \\ &= \begin{bmatrix} 1.0 \\ 0.0 \\ 1.0 \end{bmatrix} + 0.9 \begin{bmatrix} 1.0 & 0.0 & 0.0 \\ 0.6 & 0.4 & 0.0 \\ 0.0 & 0.6 & 0.4 \end{bmatrix} \begin{bmatrix} 10.0 \\ 8.44 \\ 9.23 \end{bmatrix} \\ &= \begin{bmatrix} 10.0 \\ 8.44 \\ 8.88 \end{bmatrix}.\end{aligned}$$

The resulting greedy policy is, therefore,

$$\boldsymbol{\pi}_g^{J^\pi} = \begin{bmatrix} 1.0 & 0.0 \\ 1.0 & 0.0 \\ 0.0 & 1.0 \end{bmatrix}.$$

(c) The policy π is, most likely, not optimal, since the greedy policy obtained from J^π is different from π—which can be observed by noting that the transition probabilities resulting from $\pi_g^{J^\pi}$ are different from $\mathbf{P}_\pi$. Therefore, since the greedy policy is what we would get after one step of policy iteration, π is, most likely, not optimal.

Question 5. (**1.5 pts.**)

Consider a one-step decision problem, where the decision-maker is faced with choosing one of two actions, $\mathcal{A} = \{a, b\}$. Depending of the actions of the agent, 3 possible outcomes may occur, namely outcome o_1, outcome o_2 and outcome o_3. The probability of the different outcomes given the action of the agent are

$$\begin{aligned} &\mathbb{P}\left[\mathrm{x} = o_1 \mid \mathrm{a} = a\right] = 0.35, && \mathbb{P}\left[\mathrm{x} = o_1 \mid \mathrm{a} = b\right] = 0.6, \\ &\mathbb{P}\left[\mathrm{x} = o_2 \mid \mathrm{a} = a\right] = 0.35, && \mathbb{P}\left[\mathrm{x} = o_2 \mid \mathrm{a} = b\right] = 0.4, \\ &\mathbb{P}\left[\mathrm{x} = o_3 \mid \mathrm{a} = a\right] = 0.30, && \mathbb{P}\left[\mathrm{x} = o_3 \mid \mathrm{a} = b\right] = 0.0.0 \end{aligned}$$

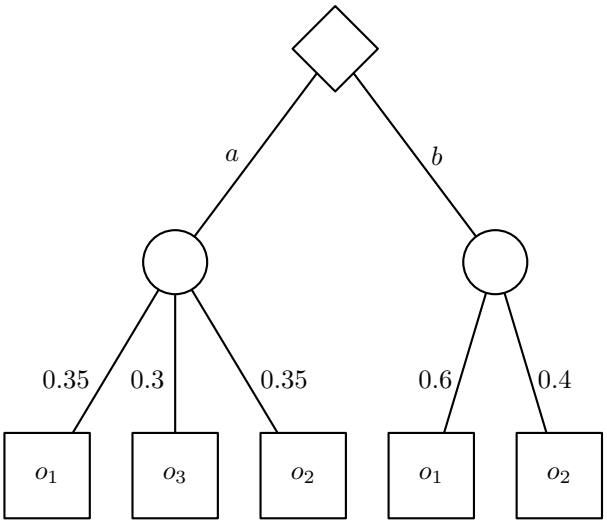
The *utility* of the three outcomes is, respectively, $u(o_1) = 0.3$, $u(o_2) = 1.0$ and $u(o_3) = 0.0$.

(a) **(0.75 pts.)** Draw the decision tree corresponding to the decision problem above.

(b) **(0.75 pts.)** Compute the optimal decision.

Solution 5.

(a) The decision tree for the problem is



(b) We have that

$$Q(a) = \sum_{x \in \mathcal{X}} \mathbb{P}\left[\mathrm{x} = x \mid \mathrm{a} = a\right] u(x) = 0.35 \times 0.3 + 0.35 \times 1 + 0.3 \times 0 = 0.455$$

$$Q(b) = \sum_{x \in \mathcal{X}} \mathbb{P}\left[\mathrm{x} = x \mid \mathrm{a} = b\right] u(x) = 0.6 \times 0.3 + 0.4 \times 1 + 0.0 \times 0 = 0.58.$$

The optimal decision is, therefore,

$$a^* = \operatorname*{argmax}_{a \in \mathcal{A}} Q(a) = b.$$

Question 6. (**1 pts.**)
Given a POMDP $(\mathcal{X}, \mathcal{A}, \mathcal{Z}, \{\mathbf{P}_a\}, \{\mathbf{O}_a\}, c, \gamma)$, let $\boldsymbol{b}_t$ denote the belief at time step t, where

$$b_t(x) = \mathbb{P}\left[\mathrm{x}_t = x \mid \mathbf{z}_{0:t} = \boldsymbol{z}_{0:t}, \mathbf{a}_{0:t-1} = \boldsymbol{a}_{0:t-1}\right].$$

Suppose that the actions are selected according to a policy π strictly dependent on the history, i.e., a_t is completely determined by $\boldsymbol{z}_{0:t}$ and $\boldsymbol{a}_{0:t-1}$. Show that

$$\mathbb{P}\left[\mathrm{z}_{t+1} = z \mid \mathbf{z}_{0:t} = \boldsymbol{z}_{0:t}, \mathbf{a}_{0:t} = \boldsymbol{a}_{0:t}\right] = \sum_{x,x'} \mathbf{O}(z \mid x', a_t)\mathbf{P}(x' \mid x, a_t)b_t(x).$$

Solution 6.

Using the total probability law,

$$\begin{aligned}
&\mathbb{P}\left[\mathrm{z}_{t+1} = z \mid \mathbf{z}_{0:t} = \boldsymbol{z}_{0:t}, \mathbf{a}_{0:t} = \boldsymbol{a}_{0:t}\right] \\
&= \sum_{x' \in \mathcal{X}} \mathbb{P}\left[\mathrm{z}_{t+1} = z \mid \mathbf{z}_{0:t} = \boldsymbol{z}_{0:t}, \mathbf{a}_{0:t} = \boldsymbol{a}_{0:t}, \mathrm{x}_{t+1} = x'\right] \mathbb{P}\left[\mathrm{x}_{t+1} = x' \mid \mathbf{z}_{0:t} = \boldsymbol{z}_{0:t}, \mathbf{a}_{0:t} = \boldsymbol{a}_{0:t}\right].
\end{aligned}$$

Since z_{t+1} is fully determined by x_{t+1} and a_t, we have

$$\mathbb{P}\left[\mathrm{z}_{t+1} = z \mid \mathbf{z}_{0:t} = \boldsymbol{z}_{0:t}, \mathbf{a}_{0:t} = \boldsymbol{a}_{0:t}\right] = \sum_{x' \in \mathcal{X}} \mathbf{O}(z \mid x', a_t)\mathbb{P}\left[\mathrm{x}_{t+1} = x' \mid \mathbf{z}_{0:t} = \boldsymbol{z}_{0:t}, \mathbf{a}_{0:t} = \boldsymbol{a}_{0:t}\right].$$

Again using the total probability law,

$$\begin{aligned}
&\mathbb{P}\left[\mathrm{z}_{t+1} = z \mid \mathbf{z}_{0:t} = \boldsymbol{z}_{0:t}, \mathbf{a}_{0:t} = \boldsymbol{a}_{0:t}\right] \\
&= \sum_{x,x' \in \mathcal{X}} \mathbf{O}(z \mid x', a_t)\mathbb{P}\left[\mathrm{x}_{t+1} = x' \mid \mathbf{z}_{0:t} = \boldsymbol{z}_{0:t}, \mathbf{a}_{0:t} = \boldsymbol{a}_{0:t}, \mathrm{x}_t = x\right] \mathbb{P}\left[\mathrm{x}_t = x \mid \mathbf{z}_{0:t} = \boldsymbol{z}_{0:t}, \mathbf{a}_{0:t} = \boldsymbol{a}_{0:t}\right]
\end{aligned}$$

which, since x_{t+1} is fully determined by x_t and a_t, becomes

$$\begin{aligned}
&\mathbb{P}\left[\mathrm{z}_{t+1} = z \mid \mathbf{z}_{0:t} = \boldsymbol{z}_{0:t}, \mathbf{a}_{0:t} = \boldsymbol{a}_{0:t}\right] \\
&= \sum_{x,x' \in \mathcal{X}} \mathbf{O}(z \mid x', a_t)\mathbf{P}(x' \mid x, a_t)\mathbb{P}\left[\mathrm{x}_t = x \mid \mathbf{z}_{0:t} = \boldsymbol{z}_{0:t}, \mathbf{a}_{0:t} = \boldsymbol{a}_{0:t}\right] \\
&= \sum_{x,x' \in \mathcal{X}} \mathbf{O}(z \mid x', a_t)\mathbf{P}(x' \mid x, a_t)\mathbb{P}\left[\mathrm{x}_t = x \mid \mathbf{z}_{0:t} = \boldsymbol{z}_{0:t}, \mathrm{a}_t = a_t, \mathbf{a}_{0:t-1} = \boldsymbol{a}_{0:t-1}\right].
\end{aligned}$$

Using Bayes rule, we can rewrite the expression above as

$$\begin{aligned}
&\mathbb{P}\left[\mathrm{z}_{t+1} = z \mid \mathbf{z}_{0:t} = \boldsymbol{z}_{0:t}, \mathbf{a}_{0:t} = \boldsymbol{a}_{0:t}\right] \\
&= \sum_{x,x' \in \mathcal{X}} \mathbf{O}(z \mid x', a_t)\mathbf{P}(x' \mid x, a_t)\frac{\mathbb{P}\left[\mathrm{a}_t = a_t \mid \mathrm{x}_t = x, \mathbf{z}_{0:t} = \boldsymbol{z}_{0:t}, \mathbf{a}_{0:t-1} = \boldsymbol{a}_{0:t-1}\right]}{\mathbb{P}\left[\mathrm{a}_t = a_t \mid \mathbf{z}_{0:t} = \boldsymbol{z}_{0:t}, \mathbf{a}_{0:t-1} = \boldsymbol{a}_{0:t-1}\right]}\mathbb{P}\left[\mathrm{x}_t = x \mid \mathbf{z}_{0:t} = \boldsymbol{z}_{0:t}, \mathbf{a}_{0:t-1} = \boldsymbol{a}_{0:t-1}\right].
\end{aligned}$$

Since the action a_t is fully determined by $\mathbf{z}_{0:t}$ and $\mathbf{a}_{0:t-1}$, we finally get

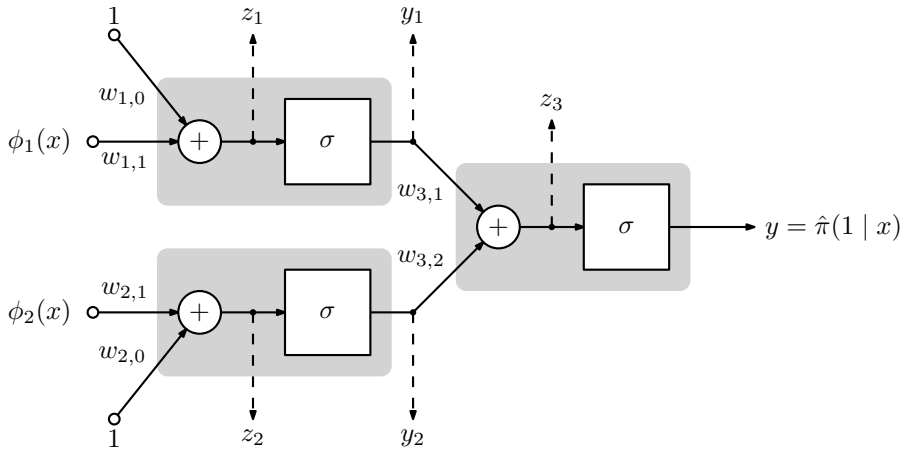
$$\begin{aligned}
&\mathbb{P}\left[\mathrm{z}_{t+1} = z \mid \mathbf{z}_{0:t} = \boldsymbol{z}_{0:t}, \mathbf{a}_{0:t} = \boldsymbol{a}_{0:t}\right] \\
&= \sum_{x,x' \in \mathcal{X}} \mathbf{O}(z \mid x', a_t)\mathbf{P}(x' \mid x, a_t)\mathbb{P}\left[\mathrm{x}_t = x \mid \mathbf{z}_{0:t} = \boldsymbol{z}_{0:t}, \mathbf{a}_{0:t-1} = \boldsymbol{a}_{0:t-1}\right] \\
&= \sum_{x,x' \in \mathcal{X}} \mathbf{O}(z \mid x', a_t)\mathbf{P}(x' \mid x, a_t)b_t(x).
\end{aligned}$$

= END OF TEST 1 =

= **BEGINNING OF TEST 2** =

Question 7. (**4.5 pts.**)

Consider a binary classification problem, where $\mathcal{A} = \{0, 1\}$ and the points x are described by two real-valued features, ϕ_1 and ϕ_2. Consider also the 2-layer neural network



where z_1, z_2, z_3, y_1 and y_2 are not real outputs, but merely auxiliary variables included for ease of reference. Suppose that the weights take the following values

$$w_{1,0} = 2; \qquad w_{2,0} = -1; \qquad w_{3,1} = 1;$$
$$w_{1,1} = 2; \qquad w_{2,1} = 3; \qquad w_{3,2} = -1.$$

(a) **(1 pts.)** Compute the output of the network when $\phi_1(x) = 0$ and $\phi_2(x) = 2$.

(b) **(2 pts.)** Compute the decision boundary for the neural network above with the provided weights. Is it a linear decision boundary?

(c) **(1.5 pts.)** Suppose that, for the point in (a), the correct output is $a = 1$. Assuming that the loss function used to train the network is the *negative log-likelihood*, Perform a *single* gradient-descent update to the weights of the network using the provided point. You should use back-propagation to compute the gradient. Use a step-size $\alpha = 1$.

Recall: The function σ appearing in the neural network diagram corresponds to the logistic function

$$\sigma(z) = \frac{1}{1 + e^{-z}}.$$

For the network above, using the negative log-likelihood as the loss, the gradient descent updates and back-propagation take the standard form:

$$\begin{aligned} &w_{i,0} \to w_{i,0} - \alpha\delta_i, \\ &w_{i,1} \to w_{i,1} - \alpha\delta_i\phi_i(x), \end{aligned} \qquad i = 1, 2 \qquad \delta_i = \delta_3 w_{3,i}\sigma'(z_i)$$

$$w_{3,i} \to w_{3,i} - \alpha\delta_3 y_i, \qquad i = 1, 2, \quad \text{with } \delta_3 = y - a,$$

where $a \in \{0, 1\}$ is the desired output for x and σ' is the derivative of σ.

Solution 7.

(a) Forward-propagating the inputs through the network, we get

$$z_1 = w_{1,0} + w_{1,1}\phi_1(x) = 2 + 2 \times 0 = 2 \qquad y_1 = \sigma(2) = 0.88$$
$$z_2 = w_{2,0} + w_{2,1}\phi_2(x) = -1 + 3 \times 2 = 5 \qquad y_2 = \sigma(5) = 0.99$$
$$z_3 = w_{3,1}y_1 + w_{3,2}y_2 = 0.88 - 0.99 = -0.11 \qquad y = \sigma(-0.11) = 0.47.$$

The output of the network is $y = 0.47$ and the point is thus classified as belonging to class 0.

(b) The decision boundary corresponds to the set

$$\text{DB} = \left\{ x \in \mathcal{X} \mid \hat{pi}(1 \mid x) = y = 0.5 \right\}.$$

Working through the network, we get:

$$y = \sigma(z_3) = 0.5 \Leftrightarrow z_3 = 0.0 \Leftrightarrow y_1 = y_2.$$

Since $y_1 = \sigma(z_1)$ and $y_2 = \sigma(z_2)$,

$$y_1 = y_2 \Leftrightarrow z_1 = z_2,$$

which can be written as

$$w_{1,0} + w_{1,1}\phi_1(x) = w_{2,0} + w_{2,1}\phi_2(x),$$

or

$$\phi_2(x) = \frac{w_{1,0} - w_{2,0}}{w_{2,1}} + \frac{w_{1,1}}{w_{2,1}}\phi_1(x) = 1 + \frac{2}{3}\phi_1(x),$$

which corresponds to a linear decision boundary.

(c) Noting that $\sigma'(z) = z(1 - z)$, we can now compute the gradient using back-propagation. We have

$$\delta_3 = y - a = 0.47 - 1 = -0.53$$
$$\delta_2 = \delta_3 w_{3,2} y_2(1 - y_2) = -0.53 \times (-1) \times 0.99 \times 0.01 = 0.005$$
$$\delta_1 = \delta_3 w_{3,1} y_1(1 - y_1) = -0.53 \times 1 \times 0.88 \times 0.12 = -0.056,$$

which leads to the updates

$$w_{1,0} = w_{1,0} - \alpha\delta_1 = 2 - 1 \times (-0.056) = 2.056$$
$$w_{2,0} = w_{2,0} - \alpha\delta_2 = -1 - 1 \times 0.005 = -1.005$$

$$w_{1,1} = w_{1,1} - \alpha\delta_1\phi_1(x) = 2 - 1 \times (-0.056) \times 0 = 2$$
$$w_{2,1} = w_{2,1} - \alpha\delta_2\phi_2(x) = 3 - 1 \times (0.052) \times 2 = 2.896$$

$$w_{3,1} = w_{3,1} - \alpha\delta_3 y_1 = 1 - 1 \times (-0.53) \times 0.88 = 1.466$$
$$w_{3,2} = w_{3,2} - \alpha\delta_3 y_2 = -1 - 1 \times (-0.53) \times 0.99 = -0.475.$$

Question 8. (**4 pts.**)

Consider an agent interacting with the MDP $(\mathcal{X}, \mathcal{A}, \{\mathbf{P}_a\}, c, \gamma)$ from Question 4, where $\mathcal{X} = \{1, 2, 3\}$ and $\mathcal{A} = \{A, B\}$. Suppose, however, that the transition probabilities $\{\mathbf{P}_a, a \in \mathcal{A}\}$ and the cost c are unknown. Consider that $\gamma = 0.9$.

Suppose that the agent wishes to evaluate the policy

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

It thus interacts with the environment, following policy π, and uses the TD(0) algorithm to compute an estimate of J^π. After t time-steps, the agent's estimate regarding J^π is given by

$$\boldsymbol{J}^{(t)} = \begin{bmatrix} 8.9 \\ 7.6 \\ 8.0 \end{bmatrix}.$$

(a) **(1 pt.)** Suppose that, at time step t, the agent is in state $x_t = 2$ and performs action $a = B$, upon which it observes a cost of 0.0 and transitions to state $x_{t+1} = 1$. Compute $\boldsymbol{J}^{(t+1)}$. Use a step-size $\alpha = 0.3$.

(b) **(1.5 pts.)** Suppose that, instead, the agent wants to evaluate the policy,

$$\boldsymbol{\pi}' = \begin{bmatrix} 0.7 & 0.3 \\ 0.1 & 0.9 \\ 0.1 & 0.9 \end{bmatrix}.$$

After running TD(0) for t steps using π', it obtains the estimate

$$\boldsymbol{J}^{(t)} = \begin{bmatrix} 5.5 \\ 4.7 \\ 5.5 \end{bmatrix}.$$

It is possible to use the sample from (a), obtained using a different policy (namely, policy π) to update the estimate of J^π by using an approach known as *importance sampling*. When using importance sampling, a transition (x_t, a_t, c_t, x_{t+1}) obtained with a policy π can be used to update the estimate of $J^{\pi'}$ by using the modified TD update

$$J^{(t+1)}(x_t) = J^{(t)}(x_t) + \alpha_t \frac{\pi'(a_t \mid x_t)}{\pi(a_t \mid x_t)} (c_t + \gamma J^{(t)}(x_{t+1}) - J^{(t)}(x_t)).$$

Use importance sampling to update the estimate for $J^{\pi'}$ using the sample from (a). Use a step-size $\alpha = 0.3$.

(c) **(1.5 pts.)** With enough samples, TD(0) converges to the solution J of the equation

$$\mathbb{E}_\pi \left[c(x_t, \mathrm{a}_t) + \gamma J(\mathrm{x}_{t+1}) - J(x_t) \right] = 0,$$

where the expectation is taken with respect to a_t—distributed according to $\pi(\cdot \mid x_t)$—and x_{t+1}—distributed according to $\mathbf{P}(\cdot \mid x_t, \mathrm{a}_t)$. The solution is none other than J^π.

Suppose that TD(0) with importance sampling—as defined in (b)—converges to the solution J of the equation

$$\mathbb{E}_\pi \left[\frac{\pi'(\mathrm{a}_t \mid x_t)}{\pi(\mathrm{a}_t \mid x_t)} \big(c(x_t, \mathrm{a}_t) + \gamma J(\mathrm{x}_{t+1}) - J(x_t) \big) \right] = 0,$$

where the expectation is, once again, taken with respect to a_t and x_{t+1}. Show that the solution for the equation above is $J^{\pi'}$.

Solution 8.

(a) The TD(0) update comes

$$\begin{aligned} J^{(t+1)}(x_t) &= J^{(t)}(x_t) + \alpha(c_t + \gamma J^{(t)}(x_{t+1}) - J^{(t)}(x_t)) \\ &= 7.6 + 0.3(0.0 + 0.9 \times 8.9 - 7.6) = 7.72, \end{aligned}$$

yielding the updated estimate

$$\boldsymbol{J}^{(t)} = \begin{bmatrix} 8.9 \\ 7.72 \\ 8.0 \end{bmatrix}.$$

(b) Using the provided update rule, we get

$$\begin{aligned} J^{(t+1)}(x_t) &= J^{(t)}(x_t) + \alpha \frac{\pi'(a_t \mid x_t)}{\pi(a_t \mid x_t)}(c_t + \gamma J^{(t)}(x_{t+1}) - J^{(t)}(x_t)) \\ &= 4.7 + 0.3\frac{0.9}{1.0}(0.0 + 0.9 \times 5.5 - 4.7) = 4.77, \end{aligned}$$

yielding the updated estimate

$$\boldsymbol{J}^{(t)} = \begin{bmatrix} 5.5 \\ 4.77 \\ 5.5 \end{bmatrix}.$$

(c) Expanding the expectation, we have

$$\begin{aligned} &\mathbb{E}_{\pi}\left[\frac{\pi'(\mathrm{a}_t \mid x_t)}{\pi(\mathrm{a}_t \mid x_t)}\big(c(x_t, \mathrm{a}_t) + \gamma J(\mathrm{x}_{t+1}) - J(x_t)\big)\right] \\ &= \sum_{a \in \mathcal{A}} \pi(a \mid x_t)\frac{\pi'(a \mid x_t)}{\pi(a \mid x_t)}\big(c(x_t, a) + \gamma \sum_{y \in \mathcal{X}} \mathbf{P}(y \mid x_t, a)J(y) - J(x_t)\big) \\ &= \sum_{a \in \mathcal{A}} \pi'(a \mid x_t)\big(c(x_t, a) + \gamma \sum_{y \in \mathcal{X}} \mathbf{P}(y \mid x_t, a)J(y) - J(x_t)\big) \\ &= \mathbb{E}_{\pi'}\left[c(x_t, \mathrm{a}_t) + \gamma J(\mathrm{x}_{t+1}) - J(x_t)\right], \end{aligned}$$

and the solution to

$$\mathbb{E}_{\pi'}\left[c(x_t, \mathrm{a}_t) + \gamma J(\mathrm{x}_{t+1}) - J(x_t)\right] = 0$$

is $J^{\pi'}$.

Question 9. (**1.5 pts.**)

Consider an agent interacting in a stochastic bandit setting, with actions $\mathcal{A} = \{a, b, c\}$. After interacting 10 times with the environment, the agent's estimates for the values of the different actions are

$$\hat{\boldsymbol{c}} = \begin{bmatrix} 0.3 & 0.8 & 0.1 \end{bmatrix},$$

where the total number of times each action was selected is

$$\boldsymbol{N} = \begin{bmatrix} 6 & 3 & 1 \end{bmatrix}.$$

(a) **(0.75 pts)** Determine the next action to be selected by the algorithm.

(b) **(0.75 pts.)** Suppose that the costs of the three actions, at time $t = 10$, are

$$\boldsymbol{c}_{10} = \begin{bmatrix} 0.1 & 0.2 & 0.8 \end{bmatrix}.$$

Compute the updated estimates for $\hat{\boldsymbol{c}}$.

Solution 9.

(a) Computing the UCB heuristic values,

$$\hat{Q}(a) = \hat{c}(a) - \sqrt{\frac{2\log t}{N_t(a)}} = 0.3 - \sqrt{\frac{2\log 10}{6}} = -0.57$$

$$\hat{Q}(b) = \hat{c}(b) - \sqrt{\frac{2\log t}{N_t(b)}} = 0.8 - \sqrt{\frac{2\log 10}{3}} = -0.43$$

$$\hat{Q}(c) = \hat{c}(c) - \sqrt{\frac{2\log t}{N_t(c)}} = 0.1 - \sqrt{\frac{2\log 10}{1}} = -2.05.$$

The agent will select action c.

(b) The agent will only update the estimate $\hat{c}(c)$ as

$$\hat{c}(c) = \hat{c}(c) + \frac{1}{N_{t+1}(c)}(c_{10}(c) - \hat{c}(c)) = 0.1 + 0.5 \times 0.7 = 0.45.$$

The resulting estimates are, therefore,

$$\hat{\boldsymbol{c}} = \begin{bmatrix} 0.3 & 0.8 & 0.45 \end{bmatrix}.$$