**Learning and Decision Making 2015-2016**

MSc in Computer Science and Engineering

Recovery examination – July 1, 2016

# 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 exam has a total of 10 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 exam has a total of 5 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

**Question 1.** (**2 pts.**)

Consider the following problem. A first year student finished her final project for a course, and she is on her way to I.S.T. to submit the project report. Suddenly, she realizes that her printer failed to print half of the report. As such, she has got two possibilities:

**(R)** Return home and print the remaining pages.

**(I)** Print the remaining pages once she gets to I.S.T.

If she decides to return home, there is a 0.6 probability that she will submit the project late, due to traffic on her way back to I.S.T. Late projects are penalized with 2 values in the final grade.

On the other hand, if she decides to print at I.S.T., there is a 0.3 probability that she may not find a printer there (she is a first year student, after all) and, in that case, she will submit her project in time but with missing pages. The missing pages will cost her 3 values in the final grade. Even if she finds a printer, there is a 0.5 probability that the printer is busy. In that case, she will again submit the project late (but complete), incurring the corresponding penalty.

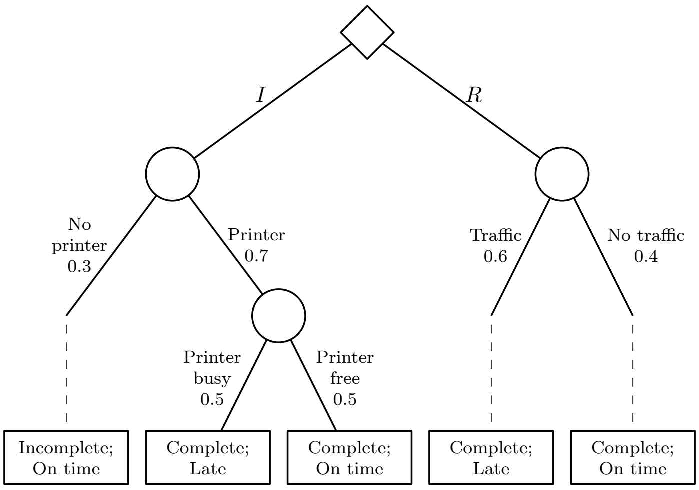
Compute the best action ($R$ or $I$) according to the expected utility theory, representing the utility of each outcome by the corresponding point loss (in terms of grade).

**Solution 1.**

There are 3 possible outcomes for the event, with utilities

- Submit a complete work on time (with a utility of 0).
- Submit a complete work but late (with a utility of $-2$).
- Submit an incomplete work, on time (with a utility of $-3$).

The decision tree associated with the decision process is, therefore,



The expected utility associated with each action can now be computed as follows:

$$Q(R) = 0.6 \times (-2) = -1.2$$
$$Q(I) = 0.3 \times (-3) + 0.35 \times (-2) = -0.9 - 0.7 = -1.6.$$

She should, therefore, return home and print the work there.

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

Consider the Tiger problem discussed in class. This problem can be modeled as a POMDP $(\mathcal{X}, \mathcal{A}, \mathcal{Z}, \{\boldsymbol{P}_a\}, \{\boldsymbol{O}_a\}, c, \gamma)$, where $\mathcal{X} = \{x_\ell, x_r\}$ ($x_\ell$ indicates the tiger on the left and $x_r$ the tiger on the right), $\mathcal{A} = \{a_0, a_\ell, a_r\}$ ($a_0$ is the action listen, $a_\ell$ and $a_r$ are the actions for opening the left and right door, respectively), $\mathcal{Z} = \{z_\ell, z_r\}$, and the transition and observation probabilities are given by:

$$\boldsymbol{P}_{a_0} = \begin{bmatrix} 1.0 & 0.0 \\ 0.0 & 1.0 \end{bmatrix} \qquad \boldsymbol{P}_{a_\ell} = \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix} \qquad \boldsymbol{P}_{a_r} = \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix}$$

$$\boldsymbol{O}_{a_0} = \begin{bmatrix} 0.85 & 0.15 \\ 0.15 & 0.85 \end{bmatrix} \qquad \boldsymbol{O}_{a_\ell} = \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix} \qquad \boldsymbol{O}_{a_r} = \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix}$$

where states, actions and observations are ordered as in the sets indicated above. Knowing that the tiger is initially placed randomly with a probability 0.5 behind one of the two doors, and the agent selected actions $a_0 = a_\ell$ and $a_1 = a_0$, use the forward-backward algorithm to determine the probability that the state at time step $t = 1$ is $\mathrm{x}_t = x_\ell$ given the observation sequence $\boldsymbol{z}_{1:2}\,\{z_\ell, z_\ell\}$.

**Solution 2.**

Note that, since we know the actions taken by the agent, the POMDP reduces to a (non-homogeneous) HMM. To compute $\mathbb{P}\left[\mathrm{x}_1 = x_\ell \mid \mathbf{z}_{1:2} = \{z_\ell, z_\ell\}\right]$ using the forward-backward algorithm, we should use, in each step, the transition/observation matrices associated with the action selected at that time step. We have, thus

$$\gamma_1(x_\ell) = \frac{\alpha_1(x_\ell)\beta_1(x_\ell)}{\boldsymbol{\alpha}_1^\top \boldsymbol{\beta}_1}.$$

Starting with the backward computation, we have that $\boldsymbol{\beta}_2 = \mathbf{1}$ and

$$\begin{aligned} \boldsymbol{\beta}_1 &= \boldsymbol{P}_{a_0}\,\mathrm{diag}(\boldsymbol{O}_{a_0, z_\ell})\boldsymbol{\beta}_2 \\ &= \begin{bmatrix} 1.0 & 0.0 \\ 0.0 & 1.0 \end{bmatrix} \begin{bmatrix} 0.85 & 0.00 \\ 0.00 & 0.15 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \\ &= \begin{bmatrix} 0.85 \\ 0.15 \end{bmatrix}. \end{aligned}$$

As for the forward computation, since there is no initial observation, we have:

$$\begin{aligned} \boldsymbol{\alpha}_0 &= \begin{bmatrix} 0.5 & 0.5 \end{bmatrix} \\ \boldsymbol{\alpha}_1 &= \boldsymbol{\alpha}_0 \boldsymbol{P}_{a_\ell}\,\mathrm{diag}(\boldsymbol{O}_{a_\ell, z_\ell}) \\ &= \begin{bmatrix} 0.5 & 0.5 \end{bmatrix} \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix} \begin{bmatrix} 0.5 & 0.0 \\ 0.0 & 0.5 \end{bmatrix} \\ &= \begin{bmatrix} 0.5 & 0.5 \end{bmatrix}. \end{aligned}$$

and we finally get

$$\gamma_1(x_\ell) = \frac{0.425}{0.425 + 0.075} = 0.85.$$

**Question 3.** (**6 pts.**)

A retired boxer is considering participating in a comeback match. Everyday, he faces the choice between resuming his training ($T$) or relaxing and postponing the training to the next day ($R$). By training he will regain his physical shape, but it is painful and stressful. Relaxing, on the other hand, allows the boxer to spend some time with his family and engage in relaxing activities.

There is some probability that a comeback match may appear at any moment. If a fight appears when the boxer is unfit, he will get injured, corresponding to a cost of 1. If the boxer is fit, he will surely win the fight, corresponding to a cost of 0. The effort of training corresponds to a cost of 0.6, while for relaxing is 0.4 except in the fight state, where both actions yield a cost of 0.

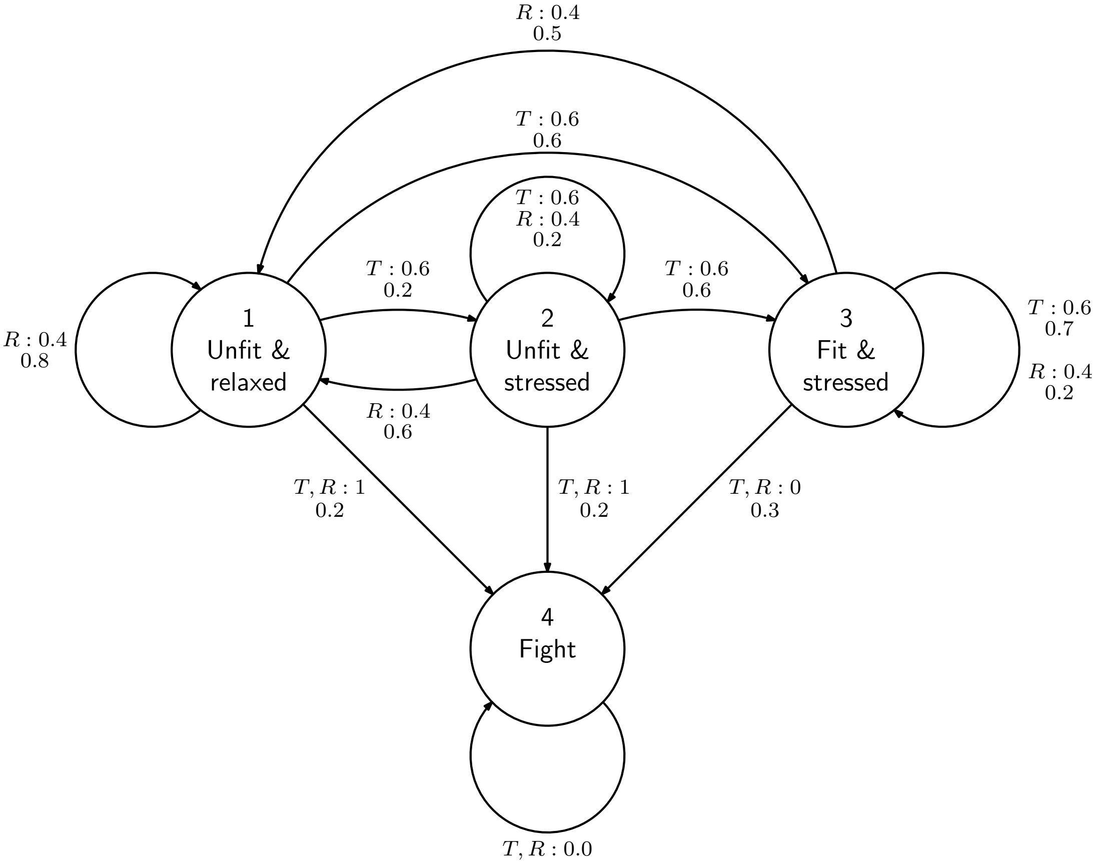


Figure 1: States, actions and transitions for the decision problem faced by the boxer.

The decision problem of the boxer can be modeled as an MDP $(\mathcal{X}, \mathcal{A}, \{\boldsymbol{P}_a\}, c, \gamma)$, where states, actions and transitions are depicted in the diagram of Fig. 1, where edge labels take the form $\frac{\langle\text{action}\rangle:\langle\text{cost}\rangle}{\langle\text{trans. prob.}\rangle}$.

(a) **(1.5 pts.)** For the MDP above, indicate the state-space, action-space, transition probabilities and cost. Your cost should be a function of the state and action.

(b) **(1.5 pts.)** Consider the cost-to-go function

$$J = \begin{bmatrix} 1.72 & 1.72 & 1.25 & 0 \end{bmatrix}^\top,$$

where the states are ordered as in Fig. 1. Determine whether the value function above is optimal for the MDP. In your computations, use $\gamma = 0.95$.

(c) **(1 pt.)** Assuming that the value function from the previous question is indeed optimal, compute the optimal $Q$-function and the optimal policy for the MDP.

(d) **(2 pts.)** Consider the policy that selects the action $T$ in all states, and the Markov chain induced by such policy. Represent such chain as a transition diagram. Is the chain irreducible? Explain your conclusion.

**Solution 3.**

(a) We have

- We refer to the states by the corresponding indices, where $1$ is the state "Unfit and relaxed"; $2$ is the state "Unfit and stressed"; $3$ is the state "Fit and stressed"; and $4$ is the state "Fight". We have, then, $\mathcal{X} = \{1, 2, 3, 4\}$.
- $\mathcal{A} = \{T, R\}$, where $T$ stands for the action "Train" and $R$ for the action "Relax".
- The transition probabilities follow directly from the diagram:

$$\boldsymbol{P}_T = \begin{bmatrix} 0.0 & 0.2 & 0.6 & 0.2 \\ 0.0 & 0.2 & 0.6 & 0.2 \\ 0.0 & 0.0 & 0.7 & 0.3 \\ 0.0 & 0.0 & 0.0 & 1.0 \end{bmatrix}, \qquad \boldsymbol{P}_R = \begin{bmatrix} 0.8 & 0.0 & 0.0 & 0.2 \\ 0.6 & 0.2 & 0.0 & 0.2 \\ 0.5 & 0.0 & 0.2 & 0.3 \\ 0.0 & 0.0 & 0.0 & 1.0 \end{bmatrix}.$$

- Finally, we note that the cost depends on the action of the agent ($T$ implies a cost of $0.6$; $R$ implies a cost of $.4$) but also on the *next state*—for if a transition to state "Fight" occurs from an unfit state, the agent will incur a cost of $1$, while from a fit state it will incur a cost of $0$. Since we must represent the reward as depending on state and action, the component that depends on the next state must be *averaged out* according to the corresponding transition probabilities. This yields

$$c = \begin{bmatrix} 0.68 & 0.52 \\ 0.68 & 0.52 \\ 0.42 & 0.28 \\ 0.0 & 0.0 \end{bmatrix}$$

Note how, interestingly, training is always more costly than relaxing.

(b) To verify the optimality of the provided value function, we run one step of value iteration. We have, for action $T$,

$$\begin{aligned} \boldsymbol{Q}_T &= \boldsymbol{C}_{:,T} + \gamma \boldsymbol{P}_T \boldsymbol{J} \\ &= \begin{bmatrix} 0.68 \\ 0.68 \\ 0.42 \\ 0 \end{bmatrix} + 0.95 \times \begin{bmatrix} 0.0 & 0.2 & 0.6 & 0.2 \\ 0.0 & 0.2 & 0.6 & 0.2 \\ 0.0 & 0.0 & 0.7 & 0.3 \\ 0.0 & 0.0 & 0.0 & 1.0 \end{bmatrix} \begin{bmatrix} 1.72 \\ 1.72 \\ 1.25 \\ 0.0 \end{bmatrix} = \begin{bmatrix} 1.72 \\ 1.72 \\ 1.25 \\ 0.0 \end{bmatrix}. \end{aligned}$$

As for action $R$, we get

$$\begin{aligned} \boldsymbol{Q}_R &= \boldsymbol{C}_{:,R} + \gamma \boldsymbol{P}_R \boldsymbol{J} \\ &= \begin{bmatrix} 0.52 \\ 0.52 \\ 0.28 \\ 0 \end{bmatrix} + 0.95 \times \begin{bmatrix} 0.8 & 0.0 & 0.0 & 0.2 \\ 0.6 & 0.2 & 0.0 & 0.2 \\ 0.5 & 0.0 & 0.2 & 0.3 \\ 0.0 & 0.0 & 0.0 & 1.0 \end{bmatrix} \begin{bmatrix} 1.72 \\ 1.72 \\ 1.25 \\ 0.0 \end{bmatrix} = \begin{bmatrix} 1.83 \\ 1.83 \\ 1.33 \\ 0.0 \end{bmatrix}. \end{aligned}$$

Finally, we have that

$$\boldsymbol{J}_{\text{new}} = \min\{\boldsymbol{Q}_T, \boldsymbol{Q}_R\} = \begin{bmatrix} 1.72 \\ 1.72 \\ 1.25 \\ 0.0 \end{bmatrix} = \boldsymbol{J},$$

where the $\min$ is taken component-wise. Since $\boldsymbol{J}$ and $\boldsymbol{J}_{\text{new}}$ are equal, we can conclude that $J$ is indeed optimal.

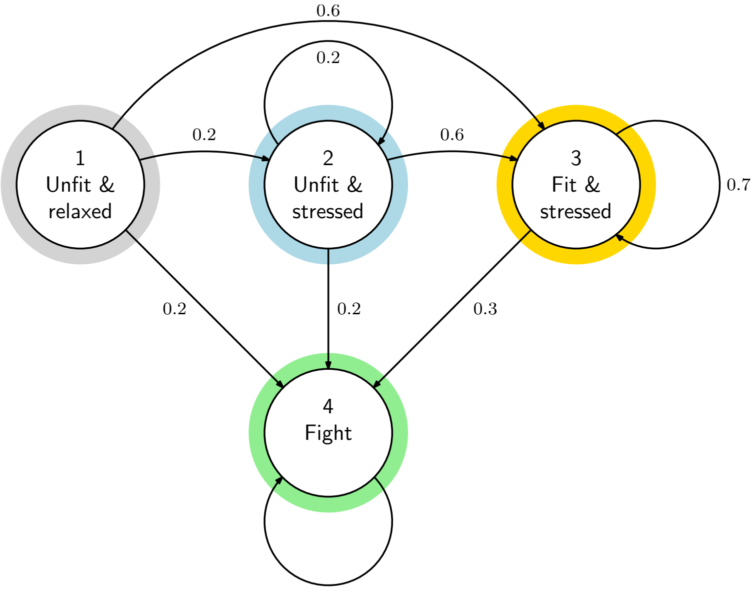
(c) Following the previous question, we have immediately that

$$\boldsymbol{Q}^* = \begin{bmatrix} 1.72 & 1.83 \\ 1.72 & 1.83 \\ 1.25 & 1.33 \\ 0.0 & 0.0 \end{bmatrix}.$$

One optimal policy is, for example,

$$\pi^* = \begin{bmatrix} T & T & T & T \end{bmatrix}^\top.$$

(d) The transition diagram for the chain is



where we shaded different communicating classes using different colors. Since there is more than a single communicating class, we can conclude that the chain is not irreducible.

## END OF TEST 1

# BEGINNING OF TEST 2

**Question 4.** (**3 pts.**)
Consider the following data, corresponding to the amount of study (in hours) of several ADI students and whether they passed the exam or not.

| Study time (hours) | 0.50 | 1.75 | 2.00 | 3.25 | 5.00 |
|---|---|---|---|---|---|
| Passed | No | Yes | No | Yes | Yes |

Suppose that you wish to train, using Newton's method, a logistic regression classifier to determine the probability of a student passing as a function of the number of hours that she studied. Considering an initial parameter vector $\boldsymbol{w} = [0, 0]^\top$, compute one iteration of Netwon's method using the data above and a step-size $\beta = 1$.

**Note:** Recall that, in Newton's method, $\boldsymbol{w}$ is updated as in the expression

$$\boldsymbol{w} = \boldsymbol{w} - \beta \boldsymbol{H}^{-1}\boldsymbol{g},$$

with

$$\boldsymbol{g} = \frac{1}{N}\sum_{n=1}^{N} a_n \boldsymbol{\phi}(x_n)(\pi_{\text{LR}}(a_n \mid x_n) - 1), \quad \boldsymbol{H} = \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{\phi}(x_n)\boldsymbol{\phi}^\top(x_n)\pi_{\text{LR}}(a_n \mid x_n)(1 - \pi_{\text{LR}}(a_n \mid x_n))$$

where, in this case, $\boldsymbol{\phi}(x) = [\phi_0(x), \phi_1(x)]^\top$, with $\phi_0(x) \equiv 1$ and $\phi_1(x_n)$ the study time of student $x$. Recall also that the inverse of a $2 \times 2$ matrix can be computed as

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad - bc}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}.$$

**Solution 4.**

We begin by computing the gradient and Hessian for the current problem. Using the provided expressions, we get

$$\begin{aligned}
\boldsymbol{g} &= \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{\phi}(x_n)\boldsymbol{\phi}^\top(x_n)\pi_{\text{LR}}(a_n \mid x_n)(1 - \pi_{\text{LR}}(a_n \mid x_n)) \\
&= 0.1\left(\begin{bmatrix} 1 \\ 0.5 \end{bmatrix} - \begin{bmatrix} 1 \\ 1.75 \end{bmatrix} + \begin{bmatrix} 1 \\ 2.0 \end{bmatrix} - \begin{bmatrix} 1 \\ 3.25 \end{bmatrix} - \begin{bmatrix} 1 \\ 5 \end{bmatrix}\right) \\
&= -\begin{bmatrix} 0.1 \\ 0.75 \end{bmatrix}.
\end{aligned}$$

As for the Hessian, we get

$$\begin{aligned}
\boldsymbol{H} &= \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{x}_n \pi_{\text{LR}}(1 \mid \boldsymbol{x}_n)(1 - \pi_{\text{LR}}(1 \mid \boldsymbol{x}_n))\boldsymbol{x}_n^\top \\
&= 0.05\left(\begin{bmatrix} 1 & 0.5 \\ 0.5 & 0.25 \end{bmatrix} + \begin{bmatrix} 1 & 1.75 \\ 1.75 & 3.06 \end{bmatrix} + \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix} + \begin{bmatrix} 1 & 3.25 \\ 3.25 & 10.56 \end{bmatrix} + \begin{bmatrix} 1 & 5 \\ 5 & 25 \end{bmatrix}\right) \\
&= \begin{bmatrix} 0.25 & 0.625 \\ 0.625 & 2.1438 \end{bmatrix},
\end{aligned}$$

with the inverse

$$\boldsymbol{H}^{-1} = \frac{1}{0.145}\begin{bmatrix} 2.1438 & -0.625 \\ -0.625 & 0.25 \end{bmatrix} = \begin{bmatrix} 14.75 & -4.30 \\ -4.30 & 1.72 \end{bmatrix}.$$

We can now perform an update according to Newton's method, to get

$$\begin{aligned} \boldsymbol{w} &= \boldsymbol{w} - \boldsymbol{H}^{-1}\boldsymbol{g} \\ &= \begin{bmatrix} 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 14.75 & -4.30 \\ -4.30 & 1.72 \end{bmatrix}\begin{bmatrix} 0.1 \\ 0.75 \end{bmatrix} \\ &= \begin{bmatrix} -1.75 \\ 0.86 \end{bmatrix} \end{aligned}$$

**Question 5.** (**7 pts.**)

Consider the 2-state POMDP depicted in Fig. 2, where transitions associated with all actions are deterministic, and there is a single observation $z_0$, made in both states with probability 1. The rewards are indicated in the diagram of Fig. 2.

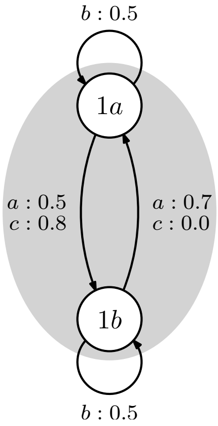


Figure 2: 2-state POMDP. The shaded region indicates that the agent makes the same observation in both states. Edge labels take the form ⟨action⟩ : ⟨cost⟩.

(a) **(2 pts.)** Knowing that, for a discount of $\gamma = 0.95$, the optimal $Q$-function for the underlying MDP is

$$Q^*_{\text{MDP}} = \begin{bmatrix} 5.13 & 5.37 & 5.43 \\ 5.57 & 5.13 & 4.87 \end{bmatrix},$$

compute the action prescribed by each of the three heuristics—most likely state, action voting, and $Q$-MDP—for the belief $\boldsymbol{b} = [0.6, 0.4]$.

(b) **(0.5 pts.)** Suppose that the (unknown) initial state is $\mathrm{x}_0 = 1a$, and the agent executes the following sequence of actions:

$$\mathbf{a}_{0:7} = \{a, b, c, c, a, b, c, b\}.$$

Indicate the corresponding sequence of costs.

(c) **(2.0 pts.)** Using the action and cost data from the previous question, determine the action selected by the agent at time step $t = 8$, if the agent follows the UCB algorithm. **Note:** If you did not complete the previous question use the following sequence of costs:

$$\mathbf{c}_{0:7} = \{0.8, 0.6, 0.7, 0.3, 0.4, 0.4, 0.9, 0.5\}\,.$$

(d) **(2.5 pts.)** Considering now the MDP underlying the model in Fig. 2, suppose that the agent is using model-based reinforcement learning to compute the optimal $Q$-function for this MDP. Further consider that the initial estimates for $\boldsymbol{P}$, $c$ and $Q^*$ are

$$\hat{\boldsymbol{P}}_a = \hat{\boldsymbol{P}}_b = \hat{\boldsymbol{P}}_c = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad\qquad \hat{\boldsymbol{C}} = \hat{\boldsymbol{Q}} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

Assuming that the initial state is $\mathrm{x}_0 = 1a$ and that the first actions of the agent are $\mathbf{a}_{0:1} = \{a, b\}$, compute two step of model-based reinforcement learning, updating the estimates $\hat{\boldsymbol{P}}$, $\hat{\boldsymbol{c}}$ and $\hat{\boldsymbol{Q}}$.

**Solution 5.**

(a) The optimal policy for the underlying MDP, computed from the given $Q$-function, is

$$\pi_{\text{MDP}} = \begin{bmatrix} a & c \end{bmatrix}^\top.$$

Let us then consider each heuristic separately:

- Given the belief $\boldsymbol{b} = [0.6, 0.4]$, the most likely state is $1a$. The action prescribed by the MLS heuristic is, thus, action $a$.
- Again, given the belief $\boldsymbol{b} = [0.6, 0.4]$, action $a$ will have an associated weight of $0.6$, while action $c$ will have an associated weight of $0.4$. Therefore, the AV heuristic prescribes action $a$.
- Finally, computing the $Q$-MDP value associated with each action we get

$$Q_{\text{MDP}}(\boldsymbol{b}, \cdot) = \begin{bmatrix} 5.31 & 5.27 & 5.21 \end{bmatrix}^\top,$$

and the $Q$-MDP heuristic will prescribe action $c$.

(b) The given sequence of actions will lead to the sequence of states

$$\mathbf{x}_{0:8} = \{1a, 1b, 1b, 1a, 1b, 1a, 1a, 1b, 1b\}$$

and the associated sequence of costs is

$$\mathbf{c}_{0:7} = \{0.5, 0.5, 0.0, 0.8, 0.7, 0.5, 0.8, 0.5\}\,.$$

(c) To compute the action selection for UCB, we first determine the average reward associated with each action. This yields

$$\hat{Q}(a) = 0.6 \qquad\qquad \hat{Q}(b) = 0.5 \qquad\qquad \hat{Q}(c) = 0.533.$$

The coefficients of each action, according to UCB (and noting that time starts in $t = 0$ instead of $t = 1$), are then given by

$$\hat{Q}(a) - \sqrt{\frac{2\log(9)}{2}} = 0.6 - 1.48 = -0.882$$
$$\hat{Q}(b) - \sqrt{\frac{2\log(9)}{3}} = 0.5 - 1.21 = -0.710$$
$$\hat{Q}(c) - \sqrt{\frac{2\log(9)}{3}} = 0.533 - 1.21 = -0.677$$

and UCB will select action $\mathrm{a}_8 = a$.

(d) Given the initial state $\mathrm{x}_0 = 1a$ and the two actions $\mathrm{a}_0 = a$ and $\mathrm{a}_1 = b$, we get the following two transitions for the RL updates: $(1a, a, 0.5, 1b)$ and $(1b, b, 0.5, 1b)$. For the first transition we start by updating $\hat{\boldsymbol{P}}$. Only $\hat{\boldsymbol{P}}(\cdot \mid 1a, a)$ will be updated as

$$\hat{\boldsymbol{P}}(y \mid x_t, a_t) \leftarrow \left(1 - \frac{1}{N(x_t, a_t) + 1}\right)\hat{\boldsymbol{P}}(y \mid x_t, a_t) + \frac{1}{N(x_t, a_t) + 1}\mathbb{I}(x_{t+1} = y).$$

Since $N \equiv 0$ for all state-action pairs, we get

$$\hat{\boldsymbol{P}}(1a \mid 1a, a) \leftarrow (1 - 1)\hat{\boldsymbol{P}}(1a \mid 1a, a) = 0$$
$$\hat{\boldsymbol{P}}(1b \mid 1a, a) \leftarrow (1 - 1)\hat{\boldsymbol{P}}(1b \mid 1a, a) + 1 = 1.$$

As for $\hat{c}$, we have

$$\hat{c}(x_t, a_t) \leftarrow \hat{c}(x_t, a_t) + \frac{1}{N_t(x_t, a_t) + 1}\big(c_t - \hat{c}(x_t, a_t)\big),$$

yielding

$$\hat{c}(1a, a) \leftarrow 0 + (0.5 - 0) = 0.5.$$

Finally, for $Q$ we get

$$\hat{Q}(x_t, a_t) \leftarrow \hat{c}(x_t, a_t) + \gamma \sum_y \hat{\boldsymbol{P}}(y \mid x_t, a_t) \min_a \hat{Q}(y, a),$$

yielding

$$\hat{Q}(1a, a) \leftarrow 0.5 + 0.95 \begin{bmatrix} 0 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \end{bmatrix} = 0.5.$$

Repeating the process for the second transition, we get

$$\hat{\boldsymbol{P}}(1a \mid 1b, b) \leftarrow (1 - 1)\hat{\boldsymbol{P}}(1a \mid 1b, b) = 0$$
$$\hat{\boldsymbol{P}}(1b \mid 1b, b) \leftarrow (1 - 1)\hat{\boldsymbol{P}}(1b \mid 1b, b) + 1 = 1,$$

and

$$\hat{c}(1b, b) \leftarrow 0 + (0.5 - 0) = 0.5.$$

as well as

$$\hat{Q}(1b, b) \leftarrow 0.5 + 0.95 \begin{bmatrix} 0 & 1 \end{bmatrix} \begin{bmatrix} 0.5 \\ 0.0 \end{bmatrix} = 0.5,$$

yielding

$$\hat{\boldsymbol{P}}_a = \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix} \qquad \hat{\boldsymbol{P}}_b = \hat{\boldsymbol{P}}_c = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad \hat{c} = \hat{Q} = \begin{bmatrix} 0.5 & 0 & 0 \\ 0 & 0.5 & 0 \end{bmatrix}.$$