Learning and Decision Making 2016-2017

MSc in Computer Science and Engineering

Recovery examination – July 7, 2017

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 13 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 9 questions, with a maximum score of 40 points. The questions have different levels of difficulty. The point value of each question is provided next to the question number and corresponds to the exam grading (for the tests, the values should be doubled).
- *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. (**3 pts.**)

Consider the following stochastic matrix:

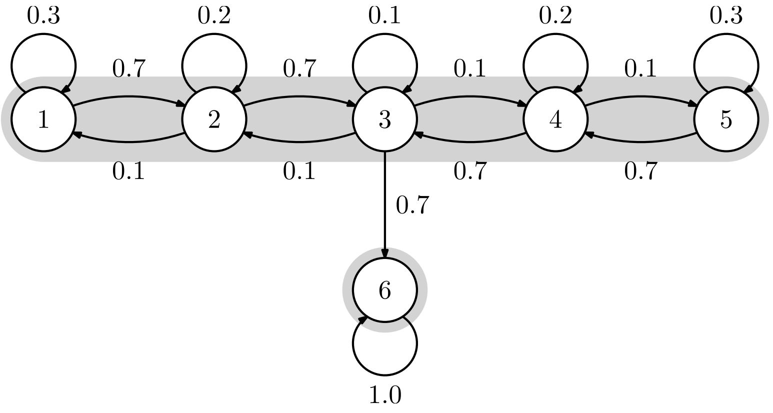
$$\mathbf{P} = \begin{bmatrix} 0.3 & 0.7 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.1 & 0.2 & 0.7 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.1 & 0.1 & 0.1 & 0.0 & 0.7 \\ 0.0 & 0.0 & 0.7 & 0.2 & 0.1 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.7 & 0.3 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 1.0 \end{bmatrix}.$$

a) **(1.5 pt.)** Represent the Markov chain with transition probability matrix $\mathbf{P}$ using a transition diagram.

b) **(1.5 pt.)** Identify the communicating classes for the chain in a), marking them in your transition diagram. Is the chain irreducible? Why?

Solution 1.

a) The chain can be represented using the transition diagram



b) The chain possesses two communicating classes, marked in the diagram above, namely:

$$\mathcal{C}_1 = \{1, 2, 3, 4, 5\} \qquad \text{and} \qquad \mathcal{C}_2 = \{6\}.$$

Therefore, it is not irreducible.

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

Given a Markov chain $(\mathcal{X}, \mathbf{P})$ and a state $x \in \mathcal{X}$ with period d, show that the period of any state $y \in \mathcal{X}$ such that $y \leftrightarrow x$ is also d.

Solution 2.

If $y \leftrightarrow x$, then there is $m, n > 0$ such that

$$\mathbf{P}^m(y \mid x) > 0, \qquad \text{and} \qquad \mathbf{P}^n(x \mid y) > 0.$$

Since x has period d, then for any $t > 0$ for which $\mathbf{P}^t(x \mid x) > 0$ there is k_t such that $t = k_t d$. But then, from the Champman-Kolmogorov equation,

$$\mathbf{P}^{m+n}(x \mid x) \geq \mathbf{P}^m(y \mid x)\mathbf{P}^n(x \mid y) > 0,$$

and $m + n = k_1 d$ for some $k_1 > 0$. Then, given any $t > 0$ such that $\mathbf{P}^t(y \mid y) > 0$,

$$\mathbf{P}^{m+t+n}(x \mid x) \geq \mathbf{P}^m(y \mid x)\mathbf{P}^t(y \mid y)\mathbf{P}^n(x \mid y) > 0,$$

and thus $m + t + n = k_2 d$ for some $k_2 > 0$. It follows that $t = (k_2 - k_1)d$ and y has also period d.

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

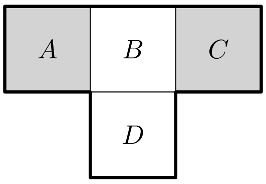
Given a strict preference relation $\succ$ over a set $\mathcal{X}$ of outcomes, show that if $x \succ y$ and $y \sim z$, then $x \succ z$, for any $x, y, z \in \mathcal{X}$.

Solution 3.

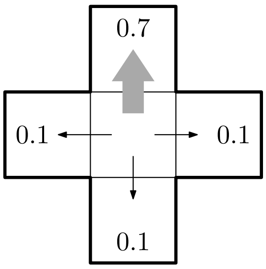
If $x \succ y$, by the negative transitivity of $\succ$, any $z \in \mathcal{X}$ verifies $x \succ z$ or $z \succ y$ (or both). But then, if $y \sim z$, this implies that $z \not\succ y$ and it must hold that $x \succ z$.

Question 4. (**12 pts.**)

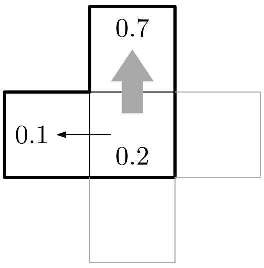
Consider a mobile robot moving in the following office environment, where each cell corresponds to a different location and the only walls correspond to the exterior contour.



The robot has available four actions, move up (u), move down (d), move left (l), and move right (r). The actions are noisy and sometimes lead to unexpected movements. Specifically, actions succeed with a probability 0.7, moving the robot to the contiguous cell in the right direction (if there is one). With a probability 0.3, the robot moves randomly to one of the other adjacent cells (when there are some). If there is no adjacent cell in a given direction, the corresponding transition keeps the robot in the same cell. The diagram below illustrates the transition probabilities for an unobstructed and an obstructed motion.



Transition probabilities
for the action u when
there are no obstructions.



Transition probabilities
for the action u when
there are obstructions
to the right and bottom.

The robot has available a set of laser sensors that the robot uses to localize. Due to hardware problems, the sensor sometimes fails and provides no reading (with a probability 0.2), but is otherwise is state-of-the-art, allowing for accurate localization. Therefore, upon receiving a reading from the sensor (which happens with probability 0.8), the robot is able to unambiguously locate itself except in the two shaded cells which look alike and both lead to the same "hallway" observation.

The goal of the agent is to reach cell D.

a) **(2 pts.)** Describe the navigation problem of the robot described above as a partially observable Markov decision problem.

b) **(2 pts.)** Suppose that the robot is completely lost, i.e., it believes that it is equally likely to be in any of the possible states. The robot takes action l and makes the "hallway" observation. What is the probability that the robot is at cell C? Indicate the relevant computations.

c) **(2 pt.)** Suppose now that, after the action and observation in b), the robot takes action l again and observes B. Compute the most likely sequence of states experienced by the robot, in light of its actions and observations. Indicate the relevant computations.

d) **(2 pts.)** Consider the policy

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

where the states are ordered from A to D and the actions as u, d, l, and r. Further suppose that associated Q-function is

$$Q^\pi = \begin{bmatrix} 6.04 & 6.04 & 6.04 & 5.42 \\ 5.17 & 4.32 & 5.80 & 5.80 \\ 6.04 & 6.04 & 5.42 & 6.04 \\ 3.68 & 2.83 & 2.83 & 2.83 \end{bmatrix}.$$

Show that π is optimal for the underlying MDP.

e) **(2 pts.)** Suppose again that the agent is completely lost. Compute the action prescribed by the Q-MDP heuristic (use the Q-function from the previous question).

f) **(2 pts.)** Suppose that the α-vectors used to represent the optimal cost-to-go for the POMDP and the associated actions are

$$\boldsymbol{\alpha}_1 = \begin{bmatrix}6.25 & 6.6 & 6.8 & 3.15\end{bmatrix}^\top, \quad \text{Action: } r;$$

$$\boldsymbol{\alpha}_2 = \begin{bmatrix}6.65 & 4.85 & 7.75 & 3.2\end{bmatrix}^\top, \quad \text{Action: } d;$$

$$\boldsymbol{\alpha}_3 = \begin{bmatrix}6.8 & 6.6 & 6.25 & 3.15\end{bmatrix}^\top, \quad \text{Action: } l;$$

$$\boldsymbol{\alpha}_4 = \begin{bmatrix}7.75 & 7.25 & 5.95 & 3.15\end{bmatrix}^\top, \quad \text{Action: } l;$$

$$\boldsymbol{\alpha}_5 = \begin{bmatrix}5.95 & 7.25 & 7.75 & 3.15\end{bmatrix}^\top, \quad \text{Action: } r;$$

$$\boldsymbol{\alpha}_6 = \begin{bmatrix}7.75 & 4.85 & 6.65 & 3.2\end{bmatrix}^\top, \quad \text{Action: } d.$$

Again suppose that the agent is completely lost. Determine the optimal action from the α-vectors above.

Solution 4.

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

- $\mathcal{X} = \{A, B, C, D\}$;
- $\mathcal{A} = \{u, d, l, r\}$
- $\mathcal{Z} = \{B, D, H, \emptyset\}$, where H is the "hallway observation, common in states A and C, and $\emptyset$ is the "null" observation;
- $\{\mathbf{P}_a, a \in \mathcal{A}\}$ are the transition probability matrices

$$\mathbf{P}_u = \begin{bmatrix}0.9 & 0.1 & 0.0 & 0.0\\ 0.1 & 0.7 & 0.1 & 0.1\\ 0.0 & 0.1 & 0.9 & 0.0\\ 0.0 & 0.7 & 0.0 & 0.3\end{bmatrix}; \qquad \mathbf{P}_d = \begin{bmatrix}0.9 & 0.1 & 0.0 & 0.0\\ 0.1 & 0.1 & 0.1 & 0.7\\ 0.0 & 0.1 & 0.9 & 0.0\\ 0.0 & 0.1 & 0.0 & 0.9\end{bmatrix};$$

$$\mathbf{P}_l = \begin{bmatrix}0.9 & 0.1 & 0.0 & 0.0\\ 0.7 & 0.1 & 0.1 & 0.1\\ 0.0 & 0.7 & 0.3 & 0.0\\ 0.0 & 0.1 & 0.0 & 0.9\end{bmatrix}; \qquad \mathbf{P}_r = \begin{bmatrix}0.3 & 0.7 & 0.0 & 0.0\\ 0.1 & 0.1 & 0.7 & 0.1\\ 0.0 & 0.1 & 0.9 & 0.0\\ 0.0 & 0.1 & 0.0 & 0.9\end{bmatrix};$$

- $\{\mathbf{O}_a, a \in \mathcal{A}\}$ are the observation probability matrices

$$\mathbf{O}_u = \mathbf{O}_d = \mathbf{O}_l = \mathbf{O}_r = \begin{bmatrix} 0.0 & 0.0 & 0.8 & 0.2 \\ 0.8 & 0.0 & 0.0 & 0.2 \\ 0.0 & 0.0 & 0.8 & 0.2 \\ 0.0 & 0.8 & 0.0 & 0.2 \end{bmatrix};$$

- c is the cost function, represented in matrix form as

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

- Finally, γ is the discount.

b) Since there is no initial observation, we have

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

A standard forward computation yields

$$\begin{aligned} \boldsymbol{\alpha}_1^\top &= \boldsymbol{\alpha}_0^\top \mathbf{P}_l \text{diag}(\mathbf{O}_{:,H}) \\ &= \begin{bmatrix} 0.25 & 0.25 & 0.25 & 0.25 \end{bmatrix} \begin{bmatrix} 0.9 & 0.1 & 0.0 & 0.0 \\ 0.7 & 0.1 & 0.1 & 0.1 \\ 0.0 & 0.7 & 0.3 & 0.0 \\ 0.0 & 0.1 & 0.0 & 0.9 \end{bmatrix} \begin{bmatrix} 0.8 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.8 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 \end{bmatrix} \\ &= \begin{bmatrix} 0.32 & 0.0 & 0.08 & 0.0 \end{bmatrix}. \end{aligned}$$

Finally, upon normalizing, we get

$$\boldsymbol{\mu}_{1|0:1} = \begin{bmatrix} 0.8 & 0.0 & 0.2 & 0.0 \end{bmatrix}.$$

c) Since we want to perform joint smoothing, we can run the Viterbi algorithm. Again, we do not have an initial observation, so we set

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

Following the steps of the algorithm yields

$$\begin{aligned} \boldsymbol{m}_1^\top &= \max\left\{\text{diag}(\boldsymbol{m}^\top)\mathbf{P}_l\right\}\text{diag}(\mathbf{O}_H) \\ &= \begin{bmatrix} 0.18 & 0.0 & 0.06 & 0 \end{bmatrix}. \end{aligned}$$

In the expression above, the maximum is taken over each column. The maximizing indices are

$$\boldsymbol{i}_0^\top = \begin{bmatrix} 1 & 3 & 3 & 4 \end{bmatrix}.$$

Repeating the process for the second step,

$$\boldsymbol{m}_2^\top = \begin{bmatrix} 0.0 & 0.0336 & 0.0 & 0.0 \end{bmatrix}$$
$$\boldsymbol{i}_1 = \begin{bmatrix} 1 & 3 & 3 & * \end{bmatrix}.$$

The resulting sequence is, thus, $C \to C \to B$.

d) To show that the policy π is optimal, it suffices to note that it is greedy with respect to its own Q-function. This property is distinctive of an optimal policy, so π must be optimal.

e) Using the Q matrix from the previous question, we have:

$$\begin{aligned} \boldsymbol{b}\,\mathbf{Q} &= \begin{bmatrix} 0.25 & 0.25 & 0.25 & 0.25 \end{bmatrix} \begin{bmatrix} 6.04 & 6.04 & 6.04 & 5.42 \\ 5.17 & 4.32 & 5.80 & 5.80 \\ 6.04 & 6.04 & 5.42 & 6.04 \\ 3.68 & 2.83 & 2.83 & 2.83 \end{bmatrix} \\ &= \begin{bmatrix} 5.24 & 4.81 & 5.02 & 5.02 \end{bmatrix}, \end{aligned}$$

and Q-MDP would prescribe action d.

f) Concatenating all α-vectors in a single matrix $\mathbf{\Lambda}$, we get

$$\begin{aligned} \boldsymbol{b}\,\mathbf{\Lambda} &= \begin{bmatrix} 0.25 & 0.25 & 0.25 & 0.25 \end{bmatrix} \begin{bmatrix} 6.25 & 6.65 & 6.80 & 7.75 & 5.95 & 7.75 \\ 6.60 & 4.85 & 6.60 & 7.25 & 7.25 & 4.85 \\ 6.80 & 7.75 & 6.25 & 5.95 & 7.75 & 6.65 \\ 3.15 & 3.20 & 3.15 & 3.15 & 3.15 & 3.20 \end{bmatrix} \\ &= \begin{bmatrix} 5.7 & 5.61 & 5.7 & 6.02 & 6.02 & 5.61 \end{bmatrix}, \end{aligned}$$

and the agent will select the action corresponding to either $\boldsymbol{\alpha}_2$ or $\boldsymbol{\alpha}_6$, which is d.

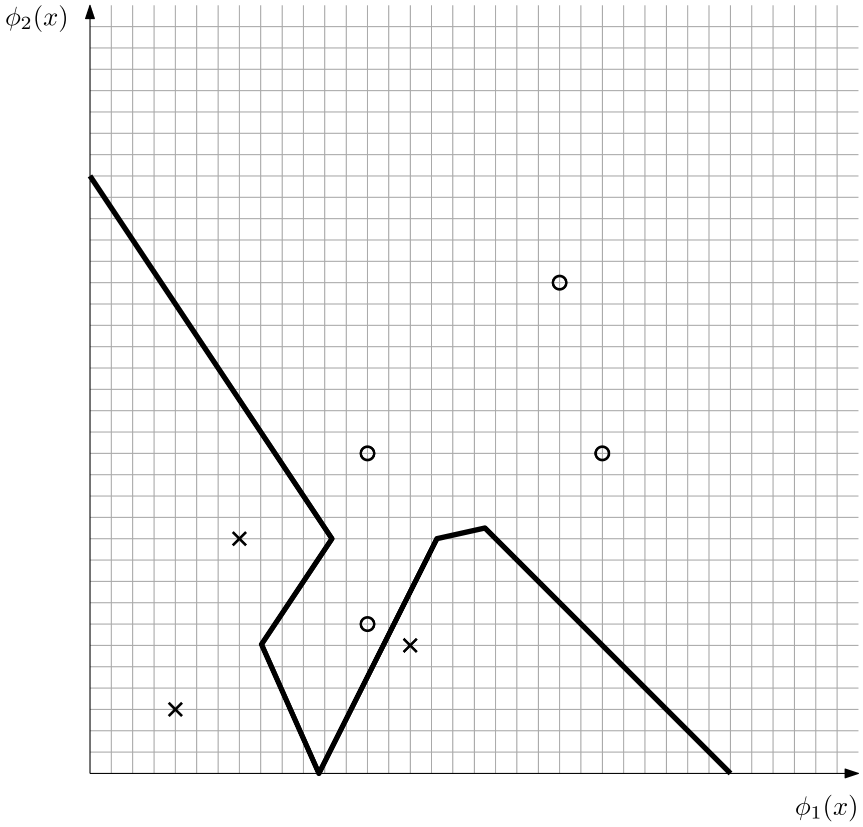
END OF TEST 1

BEGINNING OF TEST 2

Question 5. (2 pts.)
Explain what overfitting in the context of classification and indicate one possible way to minimize its negative impact on the performance of a classifier.

Solution 5.
Overfitting is a phenomenon that occurs in supervised learning in general (and classification in particular), in which the performance of a classifier in the training set is an overestimate of its actual performance. In other words, a classifier overfits when its empirical risk is much smaller than its actual risk. To minimize the impact of overfitting, there are several common strategies, such as restricting the hypothesis space or, more generally, bias the learning algorithm to prefer simpler hypothesis over more complex models (e.g., using regularization).

Question 6. (5 pts.)
Consider the training set below, where each point x_n is represented as by features, ϕ_1 and ϕ_2.



a) **(1 pts.)** Is the data linearly separable? Why?

b) **(2.5 pts.)** Indicate in the plot the decision boundary corresponding to the 1-NN classifier.

c) **(1.5 pts.)** Suppose that the classifier in b) is overfit. How can you modify the parameters of the algorithm to decrease overfitting?

Solution 6.

a) The data is not linearly separable, as there is no hyperplane that perfectly separates the two classes.

c) Overfitting can be improved by considering, for example, a k-NN classifier with $k > 1$.

Question 7. (2 pts.)

Given a training set $\mathcal{D} = \{(x_n, a_n), n = 1, \ldots, N\}$, show that a discriminative binary classifier π that minimizes the negative log-likelihood of the data in $\mathcal{D}$ also maximizes the functional

$$\ell(\mathcal{D}; \pi) = \sum_{n=1}^{N} \left[a_n \log(\pi(1 \mid x_n)) + (1 - a_n) \log(1 - \pi(1 \mid x_n))\right],$$

assuming that the samples in $\mathcal{D}$ are independent and identically distributed.

Solution 7.

The negative log likelihood of the data is given by

$$L(\mathcal{D} \mid \pi) = -\log \mathbb{P}\left[\{(x_n, a_n), n = 1, \ldots, N\} \mid \pi\right].$$

By virtue of the independence of the data, we have that

$$L(\mathcal{D} \mid \pi) = -\log \prod_{n=1}^{N} \mathbb{P}\left[(x_n, a_n) \mid \pi\right].$$

Moreover, since

$$\mathbb{P}\left[(x_n, a_n) \mid \pi\right] = \pi(a_n \mid x_n)\mathbb{P}\left[x_n\right],$$

we have

$$\begin{aligned}\min_{\pi} L(\mathcal{D} \mid \pi) &= \max_{\pi} \log \prod_{n=1}^{N} \pi(a_n \mid x_n) \prod_{n=1}^{N} \mathbb{P}\left[x_n\right] \\ &= \max_{\pi} \sum_{n=1}^{N} \log \pi(a_n \mid x_n),\end{aligned}$$

where the factor $\prod_{n=1}^{N} \mathbb{P}\left[x_n\right]$ can be removed as it does not depend on π. Finally, noting that

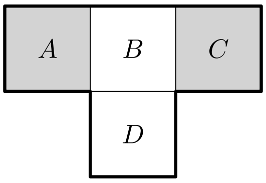
$$\pi(a_n \mid x_n) = \pi(1 \mid x_n)^{a_n} (1 - \pi(1 \mid x_n))^{1-a_n},$$

we finally get

$$\begin{aligned}\min_{\pi} L(\mathcal{D} \mid \pi) &= \max_{\pi} \sum_{n=1}^{N} [a_n \log \pi(1 \mid x_n) + (1 - a_n) \log(1 - \pi(1 \mid x_n))] \\ &= \max_{\pi} \ell(\mathcal{D} \mid \pi).\end{aligned}$$

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

Consider once again the scenario from Question 4. A mobile robot moves in the following office environment, where each cell corresponds to a different location and the only walls correspond to the exterior contour.



The robot has available four actions, move up (u), move down (d), move left (l), and move right (r). However, we now assume that the robot is a reinforcement learning agent and has no knowledge regarding the effect of its actions or the task to be learned.

Suppose that, after interacting with the environment for some time, the robot has the following estimated Q-function:

$$\hat{\mathbf{Q}} = \begin{bmatrix} 2.77 & 2.54 & 2.64 & 2.56 \\ 2.32 & 1.97 & 2.29 & 2.39 \\ 2.39 & 2.48 & 2.40 & 2.50 \\ 0.73 & 0.54 & 0.64 & 0.69 \end{bmatrix}. \tag{1}$$

Then, following an ε-greedy policy, the agent observes the following sequence of states, actions and costs:

$$\{(B, d, 1), (D, d, 0), (D, u, 0), (B, d, 1), (D, l, 0)\},$$

where a triplet (x_t, a_t, c_t) includes the state x_t, the action a_t and the cost c_t observed at time step t, and consecutive triplets correspond to consecutive time steps.

a) **(3 pts.)** Assuming that the agent is following the SARSA algorithm, compute the updates resulting from the sequence above and indicate the resulting Q-function. Use $\gamma = 0.95$ and $\alpha = 0.1$.

b) **(1 pt.)** Compute the greedy policy resulting from the Q-function you computed in a).

Note: If you have not completed a), formulate your answer with respect to the matrix in (1).

c) **(1 pt.)** According to the computations in a), for each of the actions in the sequence above indicate whether it is an exploration of exploitation action and why.

Note: If you have not completed a), formulate your answer with respect to the matrix in (1).

d) **(3 pts.)** Suppose, now, that the agent is learning using the REINFORCE policy gradient algorithm. The policy of the agent takes the form

$$\pi_{\boldsymbol{\theta}}(a \mid x) = \frac{e^{-\theta_{xa}}}{\sum_{a' \in \mathcal{A}} e^{-\theta_{xa'}}};$$

where the probability $\pi_{\boldsymbol{\theta}}(a \mid x)$ for each pair (x, a) is defined by the value θ_{xa}. If all parameters $\boldsymbol{\theta}$ are initialized to zero, compute the parameter vector resulting from two REINFORCE updates using the sequence above (in particular, use the first two samples in the sequence). In your updates, use $\alpha = 0.1$.

Note: Recall that the REINFORCE update associated with a triplet (x, a, c) is given by

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \alpha \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(a \mid x) c.$$

Solution 8.

a) In view of SARSAs update, we have data to perform only 4 updates, namely

$$\begin{aligned}
Q(B, d) &= 1.97 + 0.1 \times (1 + 0.95 \times 0.54 - 1.97) = 1.92;\\
Q(D, d) &= 0.54 + 0.1 \times (0 + 0.95 \times 0.73 - 0.54) = 0.56;\\
Q(D, u) &= 0.73 + 0.1 \times (0 + 0.95 \times 1.92 - 0.73) = 0.84;\\
Q(B, d) &= 1.92 + 0.1 \times (1 + 0.95 \times 0.64 - 1.92) = 1.89.
\end{aligned}$$

b) The Q-function resulting from the SARSA updates is

$$Q_{\text{updt}} = \begin{bmatrix} 2.77 & 2.54 & 2.64 & 2.56 \\ 2.32 & 1.89 & 2.29 & 2.39 \\ 2.39 & 2.48 & 2.40 & 2.50 \\ 0.84 & 0.56 & 0.64 & 0.69 \end{bmatrix},$$

yielding the greedy policy

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

c) The SARSA updates do not change the greedy policy, so we can directly compare which actions are in accordance with the greedy policy identified in the previous question. In particular, the actions u and l performed in state D (third and last samples) are not in accordance with the greedy policy, so they are exploration actions. The remaining actions (although one may not be sure) can all be interpreted as exploitation actions.

d) To perform the REINFORCE updates, we first compute the gradient of the log policy,

$$\frac{\partial}{\partial \theta_{x'a'}} \log \pi(a \mid x) = -\mathbb{I}(x = x') \left[\mathbb{I}(a = a') - \pi(a' \mid x')\right].$$

Then, given a sample (x, a, c),

$$\begin{aligned}
\theta_{xa} &\leftarrow \theta_{xa} - \alpha(1 - \pi(a \mid x)c\\
\theta_{xa'} &\leftarrow \theta_{xa'} + \alpha\pi(a' \mid x)c,
\end{aligned}$$

for $a' \neq a$. For the first sample, we get

$$\begin{aligned}
\theta_{Bu} &= 0 + 0.1 \times 0.25 \times 1 = 0.025;\\
\theta_{Bd} &= 0 - 0.1 \times (1 - 0.25) \times 1 = -0.075;\\
\theta_{Bl} &= 0 + 0.1 \times 0.25 \times 1 = 0.025;\\
\theta_{Br} &= 0 + 0.1 \times 0.25 \times 1 = 0.025.
\end{aligned}$$

Repeating this process for the second sample, we get

$$\begin{aligned}
\theta_{Du} &= 0 + 0.1 \times 0.25 \times 0 = 0;\\
\theta_{Dd} &= 0 - 0.1 \times (1 - 0.25) \times 0 = 0;\\
\theta_{Dl} &= 0 + 0.1 \times 0.25 \times 0 = 0;\\
\theta_{Dr} &= 0 + 0.1 \times 0.25 \times 0 = 0.
\end{aligned}$$

The resulting parameter vector is, therefore,

$$\boldsymbol{\theta} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0.025 & -0.075 & 0.025 & 0.025 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

Question 9. (3 pts.)

Consider an agent that must, at each time-step t, select one of three possible actions: a, b and c. After selecting the action, the agent pays a cost c_t that depends on the action selected at time-step t. Suppose that the agent selected the actions:

$$a \qquad\qquad b \qquad\qquad c \qquad\qquad b$$

and incurred the costs

$$0.5 \qquad\qquad 0.5 \qquad\qquad 1.0 \qquad\qquad 0.5.$$

Indicate the probability of selecting each action at the next time step if the agent follows the EXP3 algorithm with parameter $\eta = 1$. Include all relevant computations.

Solution 9.

Initially, $p(a) = p(b) = p(c) = \frac{1}{3}$. Following the EXP3 algorithm, we have

$$w_a = w_a \times e^{-\eta c/p(a)} = e^{-0.5 \times 3} = e^{-1.5} = 0.22,$$

yielding the probabilities $p(a) = 0.1$, $p(b) = p(c) = 0.45$. Then,

$$w_b = w_b \times e^{-\eta c/p(b)} = e^{-0.5 \times 2.22} = e^{-1.11} = 0.33,$$

and we get $p(a) = 0.14$, $p(b) = 0.21$ and $p(c) = 0.65$. Repeating this process, we get

$$w_c = w_c \times e^{-\eta c/p(c)} = e^{-1.0\times 1.55} = e^{-1.55} = 0.21,$$

with $p(a) = 0.29$, $p(b) = 0.43$ and $p(c) = 0.28$, and

$$w_b = w_b \times e^{-\eta c/p(b)} = 0.33e^{-0.5\times 2.32} = e^{-1.55} = 0.10,$$

finally yielding $p(a) = 0.41$, $p(b) = 0.19$ and $p(c) = 0.40$.