TÉCNICO LISBOA

Planning, Learning and Decision Making

MSc in Computer Science and Engineering

Final examination – February 5, 2019

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 13 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 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.
- *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.**)

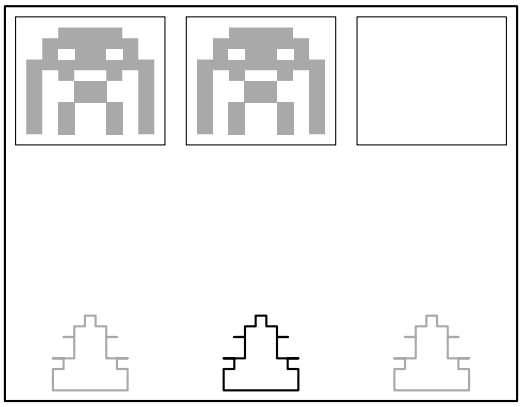


Figure 1: Simplified version of the Space Invaders game, from Atari 2600. A cannon (on the bottom) must shoot the moving aliens (on top) while avoid being killed.

Consider the following simplified version of the Space Invaders game, depicted in Fig. 1. An agent controls a "laser cannon" (on the bottom) that is used to shoot down the invading aliens. The cannon can be in any of the three indicated positions. At each moment, the agent can decide whether to move right, move left, or shoot.

Initially, as depicted in the Fig. 1, there are two aliens, side by side. At each moment, the aliens deterministically move left (if on the right side of the screen) or right (if on the left side of the screen. When there is a single alien left, it will still move from left to right all the way across the screen—now taking two steps to go from one end to the other.

Whenever the cannon moves under an alien, it is killed with probability 0.2 and the agent loses the game. The game then resets to the position in Fig. 1. Whenever the cannon shoots under an alien, it kills that alien with probability 0.5. When both aliens are killed, the agent wins the game, which then resets to the position in Fig. 1.

Describe the decision problem faced by the agent controlling the cannon using the adequate type of model. In particular, you should indicate:

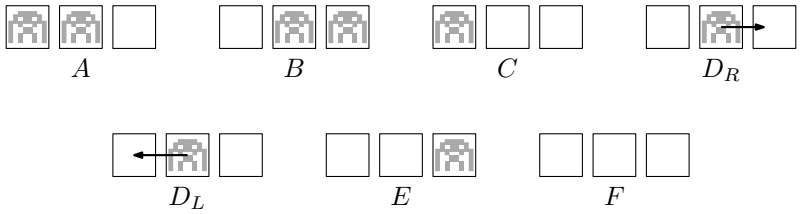
- The type of model needed to describe the decision problem of the taxi;
- The state space;
- The action space;
- The observation space (if relevant);
- The transition probabilities for actions "move right" and "shoot" from the state indicated in Fig. 1;
- An immediate cost function that translates the goals of the game as described above. The cost function should be as simple as possible.

Solution 1.

The state at time step t, x_t, should include the necessary information to predict the state at time step $t+1$. Therefore, it should include

- The configuration of the aliens—necessary to predict victory or defeat of the player, as well as the configuration at the next time step;
- The velocity of the aliens—i.e., whether they are moving right or left. For most configurations, this is trivial, but when there is a single alien in the middle, velocity information is necessary;
- The position of the player and whether the player died or not.

Taking the above into consideration, we get the following alien configurations:



which leads to the state-space:

$$\begin{aligned}\mathcal{X} = \{&(L,A),(L,B),(L,C),(L,D_R),(L,D_L),(L,E),(L,F),\\ &(M,A),(M,B),(M,C),(M,D_R),(M,D_L),(M,E),(M,F),\\ &(R,A),(R,B),(R,C),(R,D_R),(R,D_L),(R,E),(R,F),D\},\end{aligned}$$

where a state of the form (x_c, x_a) corresponds to a configuration where the cannon is in position x_c—either "left" (L), "middle" (M), or "right" (R)—and the aliens are in configuration x_a.

The problem is clearly fully observable, so an MDP is sufficient to describe the decision problem for the agent. The actions are just "move left" (L), "move right" (R), and "shoot" (S), leading to $\mathcal{A} = \{L, R, S\}$.

The cost function should translate the fact that dying is bad and killing all aliens is good. Therefore, we have

$$c(x,a) = \begin{cases} 1 & \text{if } x = D \\ 0 & \text{if } x = (*, F) \\ 0.5 & \text{otherwise.} \end{cases}$$

Finally, the state depicted in Fig. 1 corresponds to (M, A). The transition probabilities for the action "move right" come

$$\mathbf{P}(y \mid (M,A), R) = \begin{cases} 0.2 & \text{if } y = D \\ 0.8 & \text{if } y = (R, B) \\ 0 & \text{otherwise,} \end{cases}$$

and for the action "shoot",

$$\mathbf{P}(y \mid (M,A), S) = \begin{cases} 0.5 & \text{if } y = (M, D_R) \\ 0.5 & \text{if } y = (M, B) \\ 0 & \text{otherwise.} \end{cases}$$

In the remainder of the exam, consider the MDP $\mathcal{M} = (\mathcal{X}, \mathcal{A}, \{\mathbf{P}_a\}, c, \gamma)$ where

- $\mathcal{X} = \{1, 2, 3\}$;
- $\mathcal{A} = \{a, b, c\}$;
- The transition probabilities are

$$\mathbf{P}_a = \begin{bmatrix} 1.0 & 0.0 & 0.0 \\ 0.0 & 0.2 & 0.8 \\ 0.1 & 0.1 & 0.8 \end{bmatrix}; \qquad \mathbf{P}_b = \begin{bmatrix} 0.2 & 0.8 & 0.0 \\ 0.1 & 0.8 & 0.1 \\ 0.0 & 0.8 & 0.2 \end{bmatrix}; \qquad \mathbf{P}_c = \begin{bmatrix} 0.8 & 0.1 & 0.1 \\ 0.8 & 0.2 & 0.0 \\ 0.0 & 0.0 & 1.0 \end{bmatrix}.$$

- The cost function c is given by $c(x) = 1 - \mathbb{I}(x = 3)$.
- Finally, the discount is given by $\gamma = 0.9$.

You may also find useful the facts that, given a 3×3 matrix

$$\mathbf{A} = \begin{bmatrix} a & b & c \\ 0 & d & e \\ 0 & 0 & f \end{bmatrix},$$

it holds that

$$\mathbf{A}^{-1} = \begin{bmatrix} \frac{1}{a} & -\frac{b}{ad} & \frac{be-cd}{adf} \\ 0 & \frac{1}{d} & -\frac{e}{df} \\ 0 & 0 & \frac{1}{f} \end{bmatrix}.$$

Question 2. (**3 pts.**)
Let $\mathcal{M} = (\mathcal{X}, \mathcal{Z}, \mathbf{P}, \mathbf{O})$ denote a finite HMM and $\{z_1, \ldots, z_T\}$ an arbitrary sequence of observations obtained from $\mathcal{M}$. From the definition of forward and backward mappings, show that

$$\boldsymbol{\alpha}_0^\top \boldsymbol{\beta}_0 = \boldsymbol{\alpha}_T^\top \boldsymbol{\beta}_T.$$

Solution 2.

From the definition (and given that our observation sequence starts at $t = 1$),

$$\alpha_t(x) = \mathbb{P}\left[\mathrm{x}_t = x, \mathbf{z}_{1:t} = \boldsymbol{z}_{1:t}\right]$$
$$\beta_t(x) = \mathbb{P}\left[\mathbf{z}_{t+1:T} = \boldsymbol{z}_{t+1:T} \mid \mathrm{x}_t = x\right].$$

Therefore,

$$\begin{aligned} \boldsymbol{\alpha}_0^\top \boldsymbol{\beta}_0 &= \sum_{x \in \mathcal{X}} \alpha_0(x)\beta_0(x) \\ &= \sum_{x \in \mathcal{X}} \mathbb{P}\left[\mathrm{x}_0 = x\right] \mathbb{P}\left[\mathbf{z}_{1:T} = \boldsymbol{z}_{1:T} \mid \mathrm{x}_0 = x\right] \\ &= \mathbb{P}\left[\mathbf{z}_{1:T} = \boldsymbol{z}_{1:T}\right]. \end{aligned}$$

Conversely,

$$\begin{aligned}\boldsymbol{\alpha}_T^\top \boldsymbol{\beta}_T &= \sum_{x \in \mathcal{X}} \alpha_T(x)\beta_T(x) \\ &= \sum_{x \in \mathcal{X}} \mathbb{P}\left[\mathrm{x}_T = x, \mathbf{z}_{1:T} = \boldsymbol{z}_{1:T}\right] \\ &= \mathbb{P}\left[\mathbf{z}_{1:T} = \boldsymbol{z}_{1:T}\right],\end{aligned}$$

as desired.

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

Consider the MDP $\mathcal{M}$ and the Q-function

$$\boldsymbol{Q} = \begin{bmatrix} 2.95 & 2.25 & 2.65 \\ 1.20 & 2.05 & 2.75 \\ 0.30 & 0.85 & 0.0 \end{bmatrix}.$$

(a) **(1 pt.)** Compute the greedy policy with respect to the Q-function above and draw the transition diagram for the Markov chain induced by such policy.

(b) **(2.5 pts.)** Compute the cost-to-go J^π associated with the policy π from Question (a).

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

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

(c) **(2.5 pts.)** Using the policy π from Question (a), perform one step of policy iteration. Is the policy π optimal? Explain.

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

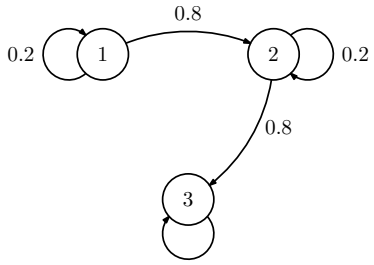
$$\pi = \begin{bmatrix} 0.5 & 0.0 & 0.5 \\ 1.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 1.0 \end{bmatrix}.$$

Solution 3.

(a) The greedy policy selects actions with minimal Q-value,

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

leading to the transition diagram



(b) The cost-to-go can be computed as

$$\begin{aligned}\boldsymbol{J}^{\pi} &= (\mathbf{I} - \gamma \mathbf{P}_{\pi})^{-1} \boldsymbol{c}_{\pi} \\ &= \left(\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} - 0.9 \begin{bmatrix} 0.2 & 0.8 & 0.0 \\ 0.0 & 0.2 & 0.8 \\ 0.0 & 0.0 & 1.0 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} \\ &= \begin{bmatrix} 1.22 & 1.07 & 7.71 \\ 0.0 & 1.22 & 8.78 \\ 0.0 & 0.0 & 10.0 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 2.29 \\ 1.22 \\ 0.0 \end{bmatrix}.\end{aligned}$$

(c) To perform one step of policy iteration, we depart from J^{π} from Question (b) and compute:

$$\boldsymbol{Q}_a = \boldsymbol{c}_a + \gamma \mathbf{P}_a \boldsymbol{J}^{\pi} = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} + 0.9 \begin{bmatrix} 1.0 & 0.0 & 0.0 \\ 0.0 & 0.2 & 0.8 \\ 0.1 & 0.1 & 0.8 \end{bmatrix} \begin{bmatrix} 2.29 \\ 1.22 \\ 0.0 \end{bmatrix} = \begin{bmatrix} 3.06 \\ 1.22 \\ 0.316 \end{bmatrix};$$

$$\boldsymbol{Q}_b = \boldsymbol{c}_b + \gamma \mathbf{P}_b \boldsymbol{J}^{\pi} = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} + 0.9 \begin{bmatrix} 0.2 & 0.8 & 0.0 \\ 0.1 & 0.8 & 0.1 \\ 0.0 & 0.8 & 0.2 \end{bmatrix} \begin{bmatrix} 2.29 \\ 1.22 \\ 0.0 \end{bmatrix} = \begin{bmatrix} 2.29 \\ 2.08 \\ 0.878 \end{bmatrix};$$

$$\boldsymbol{Q}_c = \boldsymbol{c}_c + \gamma \mathbf{P}_c \boldsymbol{J}^{\pi} = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} + 0.9 \begin{bmatrix} 0.8 & 0.1 & 0.1 \\ 0.8 & 0.2 & 0.0 \\ 0.0 & 0.0 & 1.0 \end{bmatrix} \begin{bmatrix} 2.29 \\ 1.22 \\ 0.0 \end{bmatrix} = \begin{bmatrix} 2.76 \\ 2.87 \\ 0.0 \end{bmatrix}.$$

We thus get the updated policy

$$\boldsymbol{\pi}_{\text{new}} = \begin{bmatrix} 0.0 & 1.0 & 0.0 \\ 1.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 1.0 \end{bmatrix},$$

which matches π. Therefore, we can conclude that π is optimal.

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

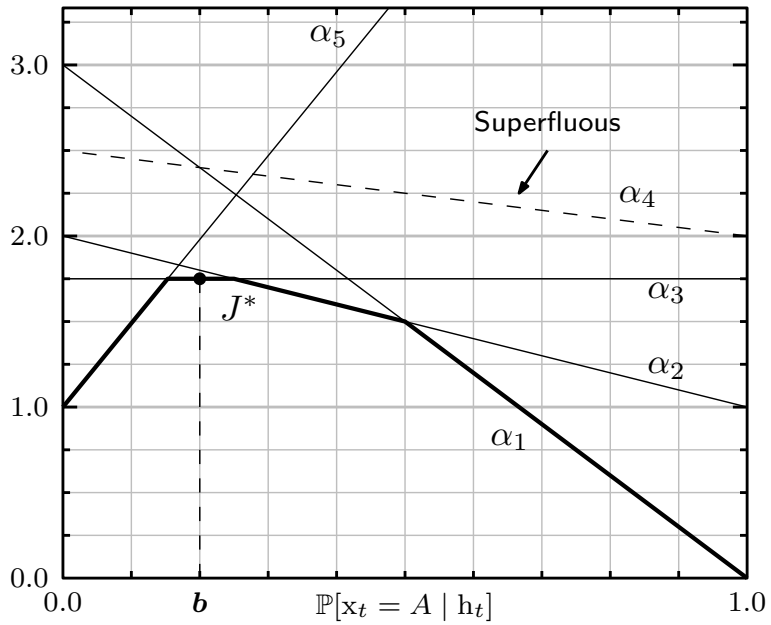
After running exact VI on a POMDP with $\mathcal{X} = \{A, B\}$ and $\mathcal{A} = \{a, b, c\}$, we got the following set of α-vectors, representing the optimal cost-to-go function:

$$\Gamma = \left\{ \begin{bmatrix} 0.0 \\ 3.0 \end{bmatrix}, \begin{bmatrix} 1.0 \\ 2.0 \end{bmatrix}, \begin{bmatrix} 1.75 \\ 1.75 \end{bmatrix}, \begin{bmatrix} 2.0 \\ 2.5 \end{bmatrix}, \begin{bmatrix} 6.0 \\ 1.0 \end{bmatrix} \right\},$$

associated with the actions $\{a, c, b, b, a\}$, respectively.

(a) **(1 pt.)** Does Γ provide a parsimonious representation for the optimal cost-to-go J^* for the POMDP? Explain your answer.

(b) **(1 pt.)** Plot the optimal cost-to-go J^* according to Γ using the grid below.



(c) **(1 pt.)** Compute the optimal action for belief $\boldsymbol{b} = [0.2, 0.8]$.

Solution 4.

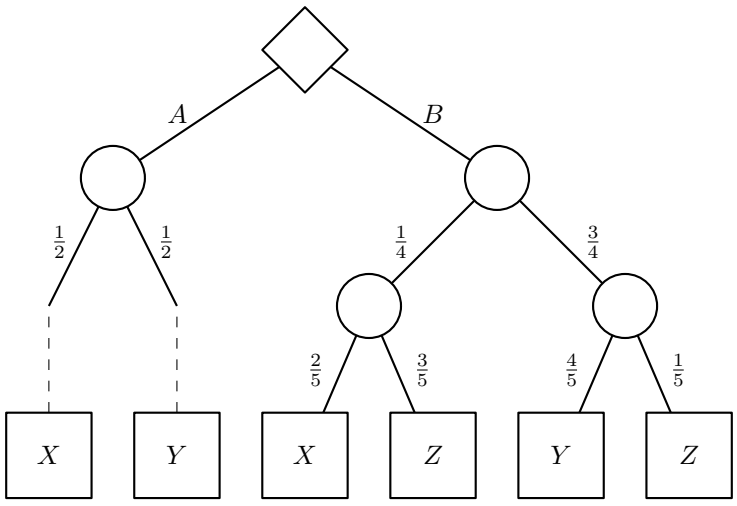
(a) Numbering the α-vectors $\alpha_1, \ldots, \alpha_5$ from left to right, we can observe that α_4 strictly dominates α_2. Therefore, α_4 is unnecessary to represent J^* and the representation is not parsimonious.

(b) The optimal cost-to-go is the bold black line depicted in the grid. Note how, as seen in Question (a), α-vector α_4 is superfluous to represent J^*.

(c) The belief $\boldsymbol{b} = [0.2, 0.8]$ is depicted in the grid, and we can see that it belongs to the witness region of α_3. For this reason, the optimal action at $\boldsymbol{b}$ is action b.

Question 5. (**2 pts.**)

Consider the decision problem described by the following decision tree.



Compute a utility function u such that

(a) **(0.8 pts.)** $A \succ B$.

(b) **(0.6 pts.)** $A \prec B$.

(c) **(0.6 pts.)** $A \sim B$.

Solution 5.

We have that

$$Q(A) = 0.5u(X) + 0.5u(Y) \qquad Q(B) = 0.1u(X) + 0.6u(Y) + 0.3u(Z).$$

Rather arbitrarily, we set throughout $u(Y) = 0$ and $u(Z) = 1$, leading to

$$Q(A) = 0.5u(X) \qquad Q(B) = 0.1u(X) + 0.3.$$

Then,

(a) To have $A \succ B$, we need to have $Q(A) > Q(B)$, which can be ensured by setting

$$0.5u(X) > 0.1u(X) + 0.3,$$

or, equivalently, $u(X) > 0.75$. We thus set, for example, $u(X) = 0.8$.

(b) Following the previous question, to have $A \prec B$, we need to have $Q(A) < Q(B)$, which can be ensured by setting, for example, $u(X) = 0.7$.

(c) Finally, to have $A \sim B$, we need to have $Q(A) = Q(B)$, which can be ensured by setting $u(X) = 0.75$.

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

Consider a Markov chain $(\mathcal{X}, \mathbf{P})$ defined by the transition probability matrix

$$\mathbf{P} = \begin{bmatrix} 0.0 & 0.3 & 0.7 & 0.0 & 0.0 \\ 1.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 1.0 & 0.0 \\ 0.0 & 0.0 & 0.5 & 0.0 & 0.5 \\ 0.5 & 0.0 & 0.0 & 0.5 & 0.0 \end{bmatrix}.$$

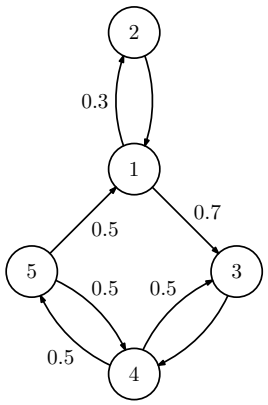
(a) **(1 pt.)** Is the chain irreducible? Explain your answer.

(b) **(1 pt.)** Is the chain aperiodic? Explain your answer.

(c) **(1 pt.)** Is the chain ergodic? Explain your answer.

Solution 6.

(a) We start by drawing the transition diagram for the chain:



We can immediately conclude that there is a single communicating class, since all states are communicating. The chain is, therefore, irreducible.

(b) Since the chain is irreducible, all states have the same period. However, the chain is not aperiodic since, for example, $\mathbf{P}^t(1 \mid 1) > 0$ only if t is even. Therefore, $d_1 = 2$ and all states have a period of 2.

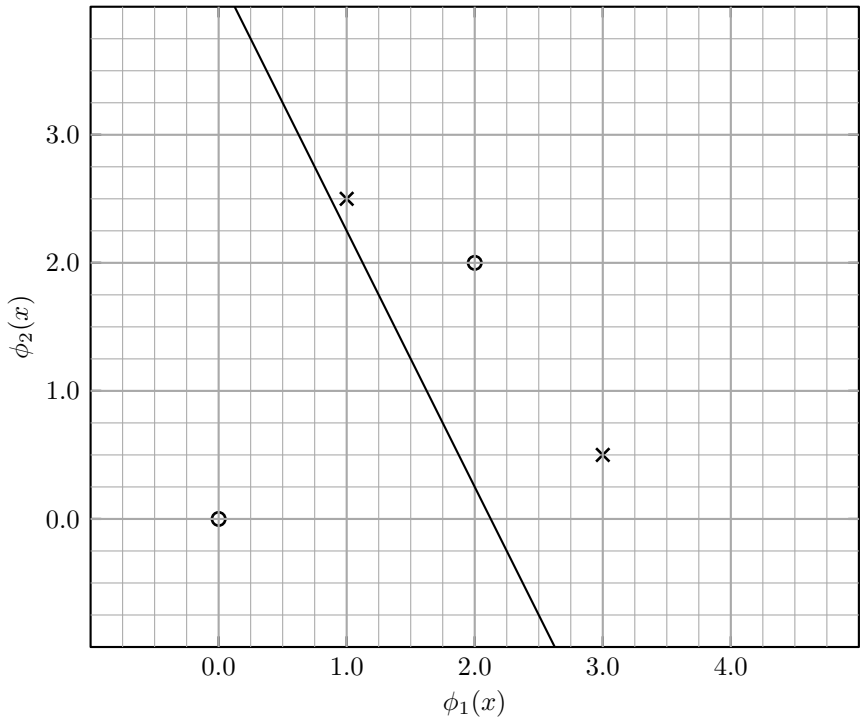
(c) It follows from Question (b) that the chain is not ergodic, since its distribution does not converge to a stationary distribution (although one such distribution does exist, in this case). For example, letting μ_t denote the state distribution at time step t, we have that $\mu_t(1) > 0$ for t even and $\mu_t(1) = 0$ for t odd.

= END OF TEST 1 =

= **BEGINNING OF TEST 2** =

Question 7. (**9 pts.**)

Consider the dataset depicted in the grid below.



(a) **(3 pts.)** Compute the parameters for a Naive Bayes classifier trained from the data above.

(b) **(3 pts.)** Using Naive Bayes, compute the class for the point x such that $\phi_1(x) = 3$ and $\phi_2(x) = 3$.

Note: If you did not solve Question (a), use $\mathbb{P}\left[\text{a} = \text{“}\circ\text{”}\right] = \mathbb{P}\left[\text{a} = \text{“}\times\text{”}\right] = 0.5$, $\mu_{1,\circ} = \mu_{2,\circ} = 0$, $\mu_{1,\times} = 3$, $\mu_{2,\times} = 0$. Assume, also, that $\sigma_{1,\circ} = \sigma_{2,\circ} = \sigma_{1,\times} = \sigma_{2,\times} = 1$.

Recall:

For a random variable $\text{x} \sim \mathsf{Normal}(\mu, \sigma)$, we have that

$$\text{pdf}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

(c) **(3 pts.)** Compute the decision boundary for the Naive Bayes classifier, and plot it in the grid of page 3.

Note: If you did not solve Question (a), use $\mathbb{P}\left[\text{a} = \text{“}\circ\text{”}\right] = \mathbb{P}\left[\text{a} = \text{“}\times\text{”}\right] = 0.5$, $\mu_{1,\circ} = \mu_{2,\circ} = 0$, $\mu_{1,\times} = 3$, $\mu_{2,\times} = 0$. Assume, also, that $\sigma_{1,\circ} = \sigma_{2,\circ} = \sigma_{1,\times} = \sigma_{2,\times} = 1$.

Solution 7.

(a) Starting with the priors,

$$\mathbb{P}[\circ] = \frac{2}{4} = 0.5 \qquad\qquad \mathbb{P}[\times] = \frac{2}{4} = 0.5.$$

The data is described by continuous attributes, so the class-conditional distributions will correspond to Gaussian distributions. For class "$\circ$" we have:

$$\phi_1(x) \mid \circ \sim \mathsf{Normal}(\mu_{1,\circ}, \sigma^2_{1,\circ}) \qquad\qquad \phi_2(x) \mid \circ \sim \mathsf{Normal}(\mu_{2,\circ}, \sigma^2_{2,\circ}),$$

with

$$\mu_{1,\circ} = \frac{0+2}{2} = 1, \qquad\qquad \sigma^2_{1,\circ} = \frac{(0-1)^2 + (2-1)^2}{2} = 1,$$
$$\mu_{2,\circ} = \frac{0+2}{2} = 1, \qquad\qquad \sigma^2_{2,\circ} = \frac{(0-1)^2 + (2-1)^2}{2} = 1.$$

Similarly, for class "$\times$",

$$\phi_1(x) \mid \times \sim \mathsf{Normal}(\mu_{1,\times}, \sigma^2_{1,\times}) \qquad\qquad \phi_2(x) \mid \times \sim \mathsf{Normal}(\mu_{2,\times}, \sigma^2_{2,\times}),$$

with

$$\mu_{1,\times} = \frac{1+3}{2} = 2, \qquad\qquad \sigma^2_{1,\times} = \frac{(1-2)^2 + (3-2)^2}{2} = 1,$$
$$\mu_{2,\times} = \frac{0.5+2.5}{2} = 1.5, \qquad\qquad \sigma^2_{2,\times} = \frac{(0.5-1.5)^2 + (2.5-1.5)^2}{2} = 1.$$

(b) We compute the probability of the point $(3, 3)$ belonging to each of the two classes. For class "$\circ$", using the expression for the Normal pdf, we get

$$\mathbb{P}[a = \circ \mid \phi_1(x) = 3, \phi_2(x) = 3] \propto \underbrace{\frac{1}{2}}_{\text{prior}} \cdot \underbrace{\frac{1}{\sqrt{2\pi}} e^{-\frac{(3-1)^2}{2}}}_{\text{prob. feature } \phi_1 \mid \circ} \cdot \underbrace{\frac{1}{\sqrt{2\pi}} e^{-\frac{(3-1)^2}{2}}}_{\text{prob. feature } \phi_2 \mid \circ} = \frac{1}{4\pi} e^{-4} = 0.0015.$$

Similarly, for class "$\times$",

$$\mathbb{P}[a = \times \mid \phi_1(x) = 3, \phi_2(x) = 3] \propto \underbrace{\frac{1}{2}}_{\text{prior}} \cdot \underbrace{\frac{1}{\sqrt{2\pi}} e^{-\frac{(3-2)^2}{2}}}_{\text{prob. feature } \phi_1 \mid \times} \cdot \underbrace{\frac{1}{\sqrt{2\pi}} e^{-\frac{(3-1.5)^2}{2}}}_{\text{prob. feature } \phi_2 \mid \times} = \frac{1}{4\pi} e^{-1.625} = 0.0157.$$

The point is, therefore, classified as belonging to class "$\times$".

(c) To compute the decision boundary for the Naive Bayes (and ignoring the normalization factor, which is the same for both classes), we set

$$\mathbb{P}[\circ]\,\mathbb{P}[\phi_1(x) \mid \circ]\,\mathbb{P}[\phi_2(x) \mid \circ] = \mathbb{P}[\times]\,\mathbb{P}[\phi_1(x) \mid \times]\,\mathbb{P}[\phi_2(x) \mid \times].$$

Replacing, we get

$$\frac{1}{2} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{(\phi_1(x)-1)^2}{2}} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{(\phi_2(x)-1)^2}{2}} = \frac{1}{2} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{(\phi_1(x)-2)^2}{2}} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{(\phi_2(x)-1.5)^2}{2}}$$

which, after some manipulation, becomes

$$e^{-\frac{(\phi_1(x)-1)^2 + (\phi_2(x)-1)^2}{2}} = e^{-\frac{(\phi_1(x)-2)^2 + (\phi_2(x)-1.5)^2}{2}}$$

or, equivalently,

$$(\phi_1(x)-1)^2 + (\phi_2(x)-1)^2 = (\phi_1(x)-2)^2 + (\phi_2(x)-1.5)^2.$$

Expanding the squares, we get

$$-2\phi_1(x) + 1 - 2\phi_2(x) + 1 = -4\phi_1(x) + 4 - 3\phi_2(x) + 2.25$$

which, solving for $\phi_2(x)$, finally yields the decision boundary

$$\phi_2(x) = 4.25 - 2\phi_1(x),$$

which is plotted in the grid.

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

Consider a RL agent interacting with the MDP $\mathcal{M}$ in page 4. Suppose that, after interacting with $\mathcal{M}$ for t steps, the agent's current estimate for Q^* is

$$\boldsymbol{Q}^{(t)} = \begin{bmatrix} 2.95 & 2.25 & 2.65 \\ 1.20 & 2.05 & 2.75 \\ 0.30 & 0.85 & 0.0 \end{bmatrix}.$$

Finally, consider the following excerpt of trajectory:

$$\ldots, x_t = 1, a_t = a, c_t = 1, x_{t+1} = 1, a_{t+1} = c, c_{t+1} = 1, x_{t+2} = 2, a_{t+2} = b, \ldots$$

(a) **(2.5 pts.)** Using the trajectory excerpt above, perform one SARSA update. Use $\alpha = 0.1$.

(b) **(2.5 pts.)** Using the trajectory excerpt above, perform one Q-learning update. Use $\alpha = 0.1$.

(c) **(3 pts.)** The standard SARSA update relies on the temporal difference

$$\delta_t = c_t + \gamma Q^{(t)}(x_{t+1}, a_{t+1}) - Q^{(t)}(x_t, a_t),$$

and is sometimes referred as *SARSA with 1-step lookahead*. More generally, *SARSA with n-step look-ahead* uses the generalized temporal difference

$$\delta_{t,n} = c_t + \gamma c_{t+1} + \gamma^2 c_{t+2} + \ldots + \gamma^{n-1} c_{t+n-1} + \gamma^n Q^{(t)}(x_{t+n}, a_{t+n}) - Q^{(t)}(x_t, a_t).$$

Using the trajectory excerpt above, perform one update of SARSA with 2-step lookahead. Use $\alpha = 0.1$.

Solution 8.

(a) From the trajectory above we get the transition $(1, a, 1, 1, c)$, which leads to the SARSA update

$$\begin{aligned} Q^{(t+1)}(1, a) &= Q^{(t)}(1, a) + \alpha(c_t + \gamma Q^{(t)}(1, c) - Q^{(t)}(1, a)) \\ &= 2.95 + 0.1 \times (1 + 0.9 \times 2.65 - 2.95) = 2.99. \end{aligned}$$

(b) From the trajectory above we get the transition $(1, a, 1, 1)$, which leads to the Q-learning update

$$\begin{aligned} Q^{(t+1)}(1, a) &= Q^{(t)}(1, a) + \alpha(c_t + \gamma \min_u Q^{(t)}(1, u) - Q^{(t)}(1, a)) \\ &= 2.95 + 0.1 \times (1 + 0.9 \times 2.25 - 2.95) = 2.96. \end{aligned}$$

(c) Using a 2-step look-ahead, we use a transition $(1, a, (1, 1), 2, b)$, where the pair $(1, 1)$ correspond to the two costs incurred at times t and $t + 1$. We get the update

$$\begin{aligned} Q^{(t+1)}(1, a) &= Q^{(t)}(1, a) + \alpha(c_t + \gamma c_{t+1} + \gamma^2 Q^{(t)}(2, b) - Q^{(t)}(1, a) \\ &= 2.95 + 0.1 \times (1 + 0.9 + 0.81 \times 2.05 - 2.95) = 3.01. \end{aligned}$$

Question 9. (**3 pts.**)

Consider an EXP3 agent with actions $\mathcal{A} = \{a, b, c\}$. After running several steps of EXP3, the agent's weights are

$$\boldsymbol{w} = \begin{bmatrix} 0.018 & 0.001 & 0.368 \end{bmatrix}.$$

(a) **(1.0 pts)** Determine the probability of selecting each action at the next time step.

(b) **(2.0 pts.)** Suppose that the agent selects action c and observes/pays a cost of 0.8. Compute the updated weights. Use $\eta = 2$.

Solution 9.

(a) The probability of selecting an action a according to EXP3 is given by

$$p(a) = \frac{w(a)}{\sum_{a' \in \mathcal{A}} w(a')}.$$

In our case, we get:

$$p(a) = \frac{0.018}{0.018 + 0.001 + 0.368} = 0.047 \qquad p(b) = \frac{0.001}{0.387} = 0.003 \qquad p(c) = \frac{0.368}{0.387} = 0.95.$$

(b) The update equation for EXP3 is, for the action a_t executed,

$$w_{t+1}(a_t) \leftarrow w_t(a_t) e^{-\eta \frac{c_t}{p(a_t)}}.$$

In our case, we get

$$w_{t+1}(a) = w_t(a); \qquad w_{t+1}(b) = w_t(b); \qquad w_{t+1}(c) = 0.368 e^{-2 \cdot \frac{0.8}{0.95}} = 0.0683.$$