Exercises (without solutions)

Francisco S. Melo

INESC-ID and Instituto Superior Técnico

In this document, we list a number of exercises to practice. There are no solutions available. For solution approaches, check examples in the lecture notes/recommended readings.

2 Markov chains

Exercise 1.

A surveillance robot moves in an office environment as described in the diagram of Fig. 1. The robot must monitor the environment, essentially moving back and forth along the main corridor, as indicated by the arrows. However, due to the noise in its actuation, it sometimes enters a private area (marked as a shaded cell in the diagram). In that case, the robot moves back to the corridor as quickly as possible, also as indicated by the black arrows.

Probabilistically, the motion of the robot can be described as follows: when moving in a given direction, the motion succeeds with a probability 0.8 but, with a probability 0.2, the motion "fails" and the robot ends up either in the same cell or in one of the contiguous cells, as depicted in Fig. 2. For example, when moving up from cell 22, the robot ends up in cell 19 with probability 0.8 and in each of the cells 21, 22, 23 and 24 with a probability of 0.05. Similarly, when moving right from cell 7 the robot ends up in cell 8 with probability 0.8; it ends up in cell 6 with probability 0.05; and it remains in cell 7 with probability 0.15.

a) Write down the Markov chain model for the robot in this example.

b) If the robot departs, at time $t = 0$, from cell 13, compute the probability that, at time step $t = 5$, the robot will be in cell 5.

c) Determine whether, at time step 2, it is more likely that the robot is in cell 6 or 7.

d) Identify the communication classes for the chain. Is the chain aperiodic?

e) Is the chain ergodic? If so, compute the stationary distribution and interpret it, in terms of the motion of the robot.

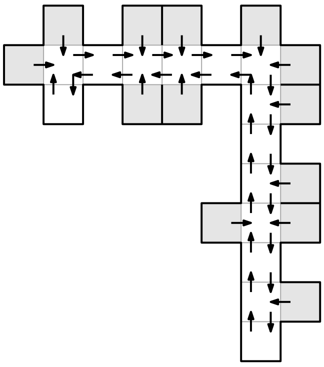


Figure 1: Motion of the robot in the office environment of Exercise 1. Whenever the robot finds itself in a private area (shaded cells), it moves out to the main corridor (black arrows). Otherwise, the robot moves along the corridor in one direction, reverting the direction in the two offices at the ends of the corridor.

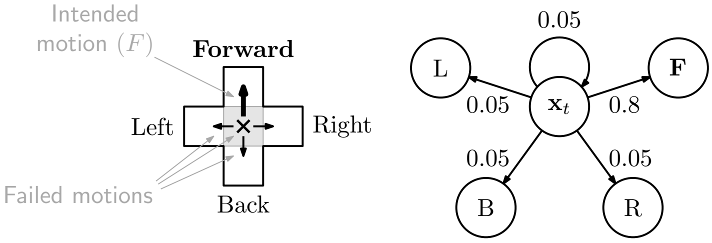


Figure 2: Probabilistic description of a single step of the robot (corresponding to each arrow in Fig. 1). With a probability of 0.8 the motion succeeds and the robot moves forward. However, with a probability 0.2, the motion fails and the robot ends up either in the same cell or one of the other contiguous cells. If no such cell exists, the associated probability adds up with that of remaining in the same cell.

Exercise 2.

Consider a Markov chain $(\mathcal{X}, \boldsymbol{P})$ where $\mathcal{X} = \{1, 2, 3, 4\}$ and

$$\boldsymbol{P} = \begin{bmatrix} \frac{1}{4} & \frac{3}{4} & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ \frac{1}{3} & 0 & \frac{1}{3} & \frac{1}{3} \end{bmatrix}.$$

a) Draw the transition diagram for the chain.

b) Identify the communicating classes for the chain. Is the chain irreducible?

c) Identify the period d for the chain. Is the chain aperiodic?

d) Determine all invariant probability distributions for the chain.

Exercise 3.
Consider a Markov chain $(\mathcal{X}, \boldsymbol{P})$ where $\mathcal{X} = \{1, 2, 3, 4, 5, 6\}$ and

$$\boldsymbol{P} = \begin{bmatrix} 0 & \frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{2} & 0 & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{2} & 0 & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 & 0 \end{bmatrix}.$$

a) Draw the transition diagram for the chain.

b) Identify the communicating classes for the chain. Is the chain irreducible?

c) Identify the period d for the chain. Is the chain aperiodic?

d) Determine all invariant probability distributions for the chain.

Exercise 4.
In a Markov chain with countable state-space $\mathcal{X}$, show that if a state $x \in \mathcal{X}$ is recurrent and $x \to y$, then y is also recurrent.

Exercise 5.
Show that, if ν is an invariant for the chain $(\mathcal{X}, \boldsymbol{P})$ it is also an invariant for any ε-resolvent $(\mathcal{X}, \mathsf{K}_\varepsilon)$.

Exercise 6.
Show that, if $(\mathcal{X}, \boldsymbol{P})$ is an ergodic Markov chain with invariant distribution μ^*, and f is a bounded real-valued function defined on $\mathcal{X}$, then $f(\mathrm{x}_t) \to \mu^* f$ as $t \to \infty$, where

$$\mu^* f = \sum_{x \in \mathcal{X}} \mu^*(x) f(x).$$

Exercise 7.

a) Show that the PageRank vector in the book is, in fact, the invariant distribution of a Markov chain built from the "web" of documents that accommodates the "teleportation" mechanism.

b) Discuss the effects of the two "adjustments" described in the PageRank example in the book, in light of the properties of the Markov chain and their impact on the existence of stationary distributions.

Exercise 8. Markov-chain Monte Carlo

In this exercise you will look the MCMC method in detail.

a) Suppose that z is a random variable that models some quantity of interest. Further suppose that z takes values in $\mathbb{R}$ and follows a Gaussian distribution, with density

$$p_{\mathrm{z}}(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}.$$

Further assume that you have at your disposal a random number generator that provides uniformly distributed samples in the interval $[0, 1]$. Assume that successive samples from the generator can be considered independent.

Explain how to use the uniform random number generator to obtain samples distributed according to p_{z}.

b) *Monte Carlo methods* are a class of algorithms that rely on random sampling to address complex numerical problems. Examples of numerical problems that can be addressed using Monte Carlo sampling include *integration* and *optimization.* Suppose, for example, that you want to compute

$$F = \mathbb{E}\left[f(\mathrm{z})\right] \triangleq \int_{\mathcal{Z}} p_{\mathrm{z}}(z) f(z) dz, \tag{1}$$

where f is some arbitrary nonlinear function. Since the integral in (1) is difficult to compute, we can resort to *sampling* to compute an estimate for F.

Suppose then that you have at your disposal a set $\mathcal{D} = \{z_1, \ldots, z_n\}$, where $z_1, \ldots, z_n$ are independent samples of z obtained, for example, using the approach in part a). Indicate how to use the sample set $\mathcal{D}$ to compute an estimate $\hat{F}_n$ for F, and show that your estimate is *consistent*, i.e., that $\hat{F}_n \to F$ as the number of samples $n \to \infty$.

In part a) you used a uniform random number generator to obtain a set of normally distributed independent samples. The *central limit theorem* (CLT) ensures that such samples approximately follow the desired distribution. In part b) you then used a Monte Carlo approach to perform numerical integration.

The Monte Carlo method relies on independent samples of the quantity of interest (in this case, the r.v. z) and is consistent, in the sense that the obtained estimate does converge to the desired quantity as the number of samples increases.

Suppose, now, that the random variable z is not normally distributed but, instead, follows some complicated distribution with density

$$p_{\mathrm{z}}(z) = \frac{e^{-h(z)}}{\int_{\mathcal{Z}} e^{-h(\zeta)} d\zeta},$$

where h is an arbitrary non-negative function. We can no longer rely on the CLT to generate the desired samples.

To address this difficulty, one common possibility is to construct a Markov chain whose invariant distribution matches the distribution we wish to sample from. If the chain is properly constructed, ergodicity can be guaranteed and we know that the samples of the chain will (eventually) follow the desired distribution.

Such methods, which combine a Monte Carlo approach with Markov-chain sampling, are collectively known as *Markov chain Monte Carlo* (MCMC). We now briefly go over a simple installment of MCMC, known as the *Metropolis-Hastings algorithm.*

Let then $q_0(\cdot \mid z)$ represent a density function associated with a distribution over $\mathcal{Z}$ that can easily be sampled (known as the *proposal distribution*), and let $\{\mathrm{z}_t, t \in \mathbb{N}\}$ represent a Markov chain constructed as follows:

- At each time step t, if $\mathrm{z}_t = z$, sample a new value z^* according to the distribution $q_0(\cdot \mid z)$;

- Compute the acceptance probability

$$\gamma(z, z^*) = \min\left\{1, \frac{p_{\mathrm{z}}(z^*)}{p_{\mathrm{z}}(z)} \cdot \frac{q_0(z \mid z^*)}{q_0(z^* \mid z)}\right\};$$

- With probability γ_t, set $\mathrm{z}_{t+1} = z^*$; otherwise, let $\mathrm{z}_{t+1} = z$;

- Set $t = t + 1$ and repeat;

c) Show that the transition probabilities for the chain $\{\mathrm{z}_t, t \in \mathbb{N}\}$ can be described in terms of the density

$$\boldsymbol{P}(z' \mid z) = \gamma(z, z')q_0(z' \mid z) + \mathbb{I}(z = z') \int_{\mathcal{Z}} (1 - \gamma(z, \zeta))q_0(\zeta \mid z)d\zeta.$$

d) Assume that q_0 is selected so that the chain $\{\mathrm{z}_t, t \in \mathbb{N}\}$ is irreducible. Using Proposition 2.8 from the book, show that the chain is positive recurrent and that p_{z} represents the density associated with the invariant distribution.

To conclude, we note that, upon careful design of the proposal distribution, it is possible to ensure that the chain $\{\mathrm{z}_t, t \in \mathbb{N}\}$ is geometrically ergodic, which ensures that the use of samples from z_t to build the sample set $\mathcal{D}$ still provides a consistent estimator for F.

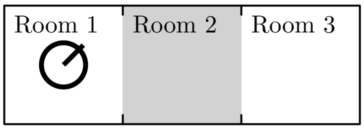


Figure 3: 3-room office environment.

3 Hidden Markov models

Exercise 9.

Consider a mobile robot navigating the 3-room office environment of Fig. 3. The rooms have colored walls: Rooms 1 and 3 have white walls, while Room 2 has gray walls. The robot has a vision sensor that enables the detection of colored walls, although lighting differences may sometimes cause the sensor to detect the wrong color. You want to track the position of the robot at every time step. To this purpose, you program the robot to move in a predefined pattern and periodically send sensorial information to you.

In particular,

- The robot is programmed to move to the adjacent room to the right at every time step. Each movement succeeds with probability 0.75 and fails with probability 0.25, leaving the position of the robot unchanged. Upon reaching Room 3, the robot remains there;
- When in a white room, the vision sensor detects a white wall with probability 0.9 and a gray wall with probability 0.1. Conversely, when in Room 2, the vision sensor detects a white wall with probability 0.3 and a gray wall with probability 0.7.

a) Write down an HMM model for this problem. In particular, define the state and observation spaces, the transition probability matrix $\boldsymbol{P}$, the observation probability matrix $\boldsymbol{O}$, and the initial distribution μ_0, supposing that the initial location of the robot is unknown (i.e., it can be any of the three rooms with equal probability).

b) Determine the probability of observing the sequence $\boldsymbol{z}_{1:3} = (w, g, w)$ using the backward computation.

c) Determine the probability of observing the sequence $\boldsymbol{z}_{1:3} = (w, g, w)$ using the forward computation. Do the results match those of part b)?

d) Determine the most likely state-sequence given the sequence of observations $\boldsymbol{z}_{1:3} = (w, g, w)$.

4 Utility theory

Exercise 10.
Consider the following situation. An agent is offered the possibility of selecting between two envelopes, A and B. The agent is told that one of the envelopes contains *twice* as much money as the other, but the agent does not know how much there is in any of the envelopes, or which one has the largest amount.

a) Suppose that the agent selects envelope A, and observes that it contains 10 EUR. The person hosting the game then gives the agent the possibility of keeping the 10 EUR or switching envelopes.

   Compute the expected value of each alternative, and indicate which should be the choice according to the expected value theory.

b) Show that the above reasoning remains valid for whichever value is in envelope A. What are the implications of such fact?

Exercise 11.
Show that if $x \succ y$ and $y \sim z$ or $x \sim y$ and $y \succ z$, then $x \succ z$.

Exercise 12.
Consider a game where the player is placed before two alternatives:

> A: The player receives $100,000$ EUR;
>
> B: A fair coin is flipped; if tails comes out, the player receives $300,000$ EUR; if heads comes out, the player receives nothing.

A study has shown that most people would prefer alternative A to alternative B. Is this result in accordance with the expected value theory?

Exercise 13.
Consider a simple game where a player flips a coin and, if tail comes out, the player wins. Compute the expected value of such a game in each of the following situations:

a) The player pays nothing for playing, and receives 90 EUR for winning. The coin is rigged and tails only come out with a probability of 0.1.

b) The player pays nothing for playing, and receives 200 EUR for winning. The coin is fair.

c) The player pays 100 EUR for playing, and receives 100 EUR for winning. The coin is rigged and tails come out with a probability 0.85.

d) The player pays nothing for playing. She receives 10 EUR for losing and 50 EUR for winning. The coin is fair.

e) The player pays nothing for playing. She receives 30 EUR for losing and 90 EUR for winning. The coin is rigged and tails come out with a probability 0.7.

f) The player pays 20 EUR for playing. She receives 50 EUR for winning. The coin is rigged and tails come out with a probability 0.25.

Exercise 14.
Consider the following game: a player selects any 5 integers from 1 to 50. A raffle then takes place, and the player receives a prize of $10,000,000$ EUR if she gets the 5 numbers right. According to the expected value theory, how much should a player pay to join the game for the game to be fair?

Exercise 15.
Suppose that Adam is going to purchase an automobile, and is having difficulties in choosing which of three possible choices, A, B and C, he should acquire. In particular, Adam prefers automobile A to automobile B since it prefers large vehicles to smaller ones and A is clearly larger than B. On the other hand, between B and C, Adam prefers B, since it is faster than C and Adam likes fast cars. Finally, vehicle C is much more environmentally friendly than car A, so between the two Adam surely prefers C.

Explain why this is not a proper preference relation, indicating which of the two axioms is violated by means of an example.

Exercise 16.
The use of utility functions allow us to extend the notion of preferences from outcomes to actions. Show that the relation $\succeq$ over a set of actions $\mathcal{A}$ is:

a) *Complete*, i.e., for any $a, b \in \mathcal{A}$, either $a \succeq b$ or $b \succeq a$ or both.

b) *Transitive*, i.e., for any $a, b, c \in \mathcal{A}$, if $a \succeq b$ and $b \succeq c$, then $a \succeq c$.

c) *Independent*, i.e, given three actions $a, b, c \in \mathcal{A}$ and $t \in (0, 1)$, if $a \succ b$, then $ta+(1-t)c \succ tb+(1-t)c$, where $ta+(1-t)c$ corresponds to selecting action a with probability t and action c with probability $1 - t$.

d) *Continuous*, i.e., given any three actions $a, b, c \in \mathcal{A}$ such that $a \succeq b$, $b \succeq c$ and $a \succ c$, then there is $t \in [0; 1]$ such that $b \sim ta + (1 - t)c$.

Exercise 17. Ellsberg paradox
In this exercise, we look into the well-known *Ellsberg paradox*, one of several empirical studies whose results disagree with expected utility theory.

A player is presented with an urn containing 30 red balls and 60 black and yellow balls. The player does not know how many of the 60 are black and how many are yellow. The balls are well mixed in the urn, so that no one ball is more likely to be drawn than any other. The player should select one of the following bets:

> *A*: It draws one ball at random from the urn. If the ball is red, the player wins 100 EUR and nothing otherwise.
>
> *B*: It draws one ball at random from the urn. If the ball is black, the player wins 100 EUR and nothing otherwise.

The expected utility theory allows both alternatives to be possible picks, depending on the player's personal preferences.

a) The study reveals that people strictly prefer alternative A to alternative B. Show that, according to the expected utility theory, this means that people tend to estimate that there are more red balls than black balls in the urn.

Consider now the following variation of the previous game. Again the player is presented with an urn containing 30 red balls and 60 black and yellow balls. The proportion of black and yellow balls is unknown, and the balls are well-mixed in the urn. The player should now select one of the following bets:

> *C*: It draws one ball at random from the urn. If the ball is red *or yellow*, the player wins 100 EUR and nothing otherwise.
>
> *D*: It draws one ball at random from the urn. If the ball is black *or yellow*, the player wins 100 EUR and nothing otherwise.

b) The study reveals that, in this second situation, people strictly prefer alternative D to alternative C. Show that, according to the expected utility theory, this means that people tend to estimate that there are more black balls than red balls in the urn.

Note that the information available to the player does not change from the first to the second game, which means that the theory, in itself, is insufficient to explain the differences observed in both games. On the other hand, note that in situations A and D the player knows the exact chances of winning, unlike situations B and C. This difference may suggest that humans may have some form of aversion to the lack of information, which is not contemplated in expected utility theory.

5 MDPs

Exercise 18.
Consider the 2-state system depicted in Fig. 4.

a) Represent this system as a Markov decision process $(\mathcal{X}, \mathcal{A}, \{\boldsymbol{P}_a\}, c)$, explicitly indicating

    - The state space, $\mathcal{X}$;
    - The action space, $\mathcal{A}$;

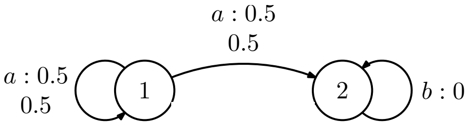


Figure 4: 2-state system of Question 18. The agent has available two actions, a and b. Action b always leads to state 2, while action a has a 0.5 probability of ending up in either state, 1 or 2.

- The transition probabilities, $\{\boldsymbol{P}_a, a \in \mathcal{A}\}$;
- The cost function, c.

b) Using policy iteration, determine the optimal policy for the resulting MDP when $\gamma = 0.99$.

c) Using value iteration, determine the optimal policy for the resulting MDP when $\gamma = 0.9$.

Exercise 19.

Consider the map depicted in Fig. 5(a), depicting several locations in a neighborhood, as well as the routes between them. A driver agent must take a customer from location A to location B, for which it must select the best possible route.

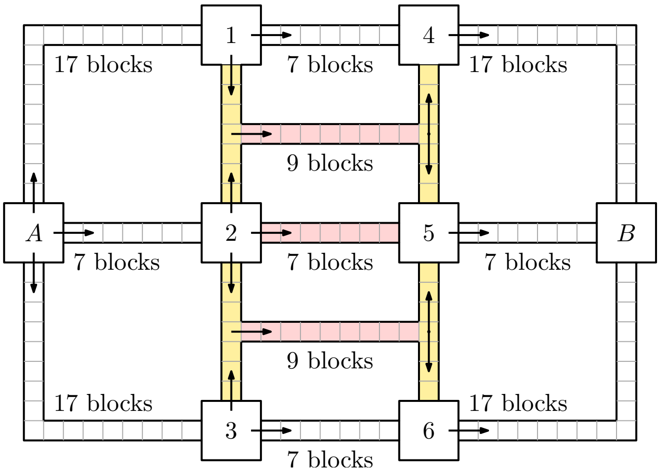
The white routes take 1 time unit per block to traverse (1 block corresponds to one of the small squares delimited in the map). Yellow routes, due to traffic, take 1.5 time units per block to traverse with a 0.5 probability; with a probability 0.5, they take 2 time units per block. Finally, due to heavy traffic, red routes take 2 time units per block to traverse with a probability 0.3; they take 3 time units per block with a probability 0.4; and, with a probability 0.3, they take 4 time units per block to traverse.

a) Represent this system as a Markov decision process $(\mathcal{X}, \mathcal{A}, \{\boldsymbol{P}_a\}, c)$, explicitly indicating

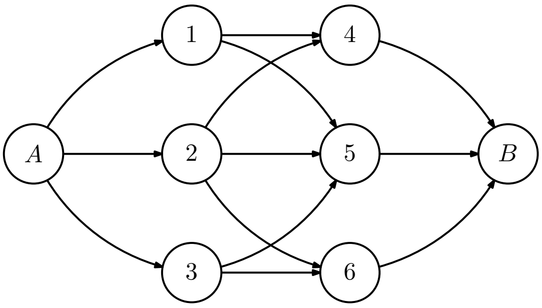
- The state space, $\mathcal{X}$;
- The action space, $\mathcal{A}$;
- The transition probabilities, $\{\boldsymbol{P}_a, a \in \mathcal{A}\}$;
- The cost function, c.

b) Determine the optimal policy for the resulting MDP when $\gamma = 0.99$.

c) Determine the optimal policy for the resulting MDP when $\gamma = 0.9$.



(a) Map of a city neighborhood, used in Question 19. Small grey blocks each take 1 time unit to traverse. Yellow blocks take 1.5 or 2 time units to traverse, due to medium traffic. Red blocks take 2 to 4 time units to traverse. A driver agent must drive a customer from location A to location B, where the price of the ride is fixed upfront.



(b) Abstract representation of the map in (a) with the possible transitions between locations.

Figure 5: Domain from Exercise 19.

Exercise 20.
Consider an agent whose job is to observe a process evolving as a Markov chain $(\mathcal{X}, \boldsymbol{P})$, with $\mathcal{X} = \{1, 2, 3, 4\}$ and

$$\boldsymbol{P} = \begin{bmatrix} 0.3 & 0.4 & 0.2 & 0.1 \\ 0.2 & 0.3 & 0.5 & 0.0 \\ 0.1 & 0.0 & 0.8 & 0.1 \\ 0.4 & 0.0 & 0.0 & 0.6 \end{bmatrix}.$$

At each time step, the agent may choose to quit as an observer, for which it is granted a payment $P_{\text{quit}} = 20$ EUR, or remain as an observer, for which it receives an amount corresponding to the state of the system—if $\mathrm{x}_t = k$, then the agent receives k EUR.

a) Represent this system as a Markov decision process $(\mathcal{X}, \mathcal{A}, \{\boldsymbol{P}_a\}, c)$, explicitly indicating

- The state space, $\mathcal{X}$;
- The action space, $\mathcal{A}$;
- The transition probabilities, $\{\boldsymbol{P}_a, a \in \mathcal{A}\}$;
- The cost function, c.

Make sure that your cost function preserves the relation between the payoffs received by the agent.

b) Letting $\gamma = 0.9$, use policy iteration to determine the optimal policy for the MDP $(\mathcal{X}, \mathcal{A}, \{\boldsymbol{P}_a\}, c, \gamma)$.

c) Find the smallest amount P_{quit} for which it is optimal to quit when $\mathrm{x}_t = 2$.

Exercise 21.
Consider once again the hiring problem in the book a number of candidates $N = 10$.

a) Formulate this problem as an MDP $(\mathcal{X}, \mathcal{A}, \{\boldsymbol{P}_a\}, c, \gamma)$.

b) Compute the optimal policy for the MDP, using $\gamma = 0.99$.

Exercise 22.
Consider the MDP for the epidemic control scenario of the book. Determine the optimal policy for this problem.

Exercise 23.
Let $J : \mathcal{X} \to \mathbb{R}$ denote an arbitrary cost-to-go function, and π_g^J the associated greedy policy. Show that it is possible that $J^{\pi_g^J} > J$. Why does this not contradict the policy improvement result?

Exercise 24.
Consider the operator T_π^T defined for a policy π, $T > 0$, and a cost-to-go function $\boldsymbol{J} \in \mathbb{R}^{|\mathcal{X}|}$ as

$$\mathsf{T}_\pi^T \boldsymbol{J} = \sum_{t=0}^{T} \gamma^t \boldsymbol{P}_\pi^t \boldsymbol{c}_\pi + \gamma^{T+1} \boldsymbol{P}_\pi^{T+1} \boldsymbol{J}.$$

Show that T_π^T is a contraction in the L_2-norm.

Exercise 25.
Let $\mathcal{M} = (\mathcal{X}, \mathcal{A}, \{\boldsymbol{P}_a\}, c, \gamma)$ denote a finite MDP and π a stationary policy for $\mathcal{M}$. Given a a function $J : \mathcal{X} \to \mathbb{R}$, define the operator

$$(\boldsymbol{\Pi} J)(x) = \boldsymbol{\phi}^\top(x) \boldsymbol{\Phi}^{-1} \sum_{y \in \mathcal{X}} \mu(y) \boldsymbol{\phi}(y) J(y).$$

where μ is some probability distribution over $\mathcal{X}$ such that $\mu(x) > 0$ for all $x \in \mathcal{X}$, and $\boldsymbol{\Phi}$ is the matrix

$$\boldsymbol{\Phi} = \mathbb{E}_\mu \left[\boldsymbol{\phi}(\mathrm{x}) \boldsymbol{\phi}^\top(\mathrm{x}) \right] = \sum_{x \in \mathcal{X}} \mu(x) \boldsymbol{\phi}(x) \boldsymbol{\phi}^\top(x).$$

a) Show that the operator $\boldsymbol{\Pi}$ is a *non-expansion* in the μ-weighted L_2-norm, defined for a function $J : \mathcal{X} \to \mathbb{R}$ as

$$\|J\|_\mu = \left(\sum_{x \in \mathcal{X}} \mu(x) J^2(x) \right)^{\frac{1}{2}}.$$

b) Show that the operator T_π is a contraction in the μ-weighted L_2-norm.

c) Show that the composed operator $\boldsymbol{\Pi}\mathsf{T}_\pi$ is a contraction in the μ-weighted L_2-norm. Use the results from a) and b).

d) Show that the function $\hat{J}^\pi$, defined as the fixed point

$$\hat{J}^\pi(x) = (\boldsymbol{\Pi}\mathsf{T}_\pi \hat{J}^\pi)(x),$$

verifies

$$\left\| \hat{J}^\pi - J^\pi \right\|_\mu \leq \frac{1}{1-\gamma} \left\| \boldsymbol{\Pi} J^\pi - J^\pi \right\|_\mu.$$

6 Partially Observable MDPs

Exercise 26.
Consider an agent moving in the environment depicted in Fig. 6.

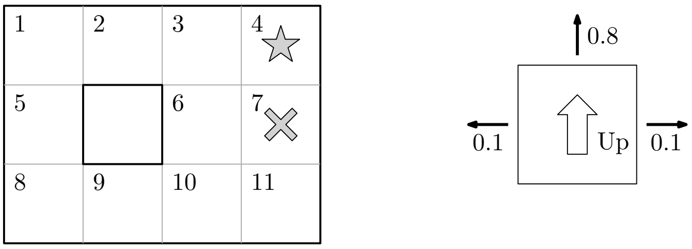


Figure 6: 4×3 environment with obstacle. The agent must reach the star while avoiding the cross. When performing an action (for example, action Up), it will succeed with probability 0.8 and fail with probability 0.2, moving the agent to one of the two adjacent cells, orthogonally to the intended motion.

The agent has available a total of 4 actions, corresponding to motions in the four directions: U (Up), D (Down), L (Left) and R (Right). Each action succeeds with probability 0.8 and fails with a probability 0.2, moving the agent to one of the two adjacent cells, orthogonally to the intended motion (see Fig. 6 for an example). Each action has a cost of 0.5.

The environment includes two terminal states, one containing a star and the other containing a cross. Upon reaching one of these two states, the agent gets to stay for free and pay a cost of 1, respectively, and remains there forever. When in a non-terminal state, the agent is unable to perceive the current state, and sees only the number of walls of its present cell. However, with a 0.1 probability, the agent observes the wrong number of walls.

This problem can be modeled as a POMDP $\mathcal{M} = (\mathcal{X}, \mathcal{A}, \mathcal{Z}, \{\boldsymbol{P}_a\}, \{\boldsymbol{O}_a\}, c, \gamma)$ with $\mathcal{X} = \{1, \ldots, 11\}$, $\mathcal{A} = \{U, D, L, R\}$, and $\mathcal{Z} = \{1, 2, S, C\}$, where S corresponds to the star and C to the bomb (the two terminal states are fully observable).

a) Write down the remaining POMDP parameters, namely the transition probabilities $\boldsymbol{P}_a, a \in \mathcal{A}$, the observation probabilities $\{\boldsymbol{O}_a, a \in \mathcal{A}\}$ and the costs c.

b) Suppose that the initial belief state corresponds to the uniform distribution over non-terminal states, i.e.,

$$\boldsymbol{b}_0 = \begin{bmatrix} \frac{1}{9} & \frac{1}{9} & \frac{1}{9} & 0 & \frac{1}{9} & \frac{1}{9} & 0 & \frac{1}{9} & \frac{1}{9} & \frac{1}{9} & \frac{1}{9} \end{bmatrix}$$

Compute the belief state after the agent selects the action L and observes one adjacent wall.

Exercise 27.
Let $\mathcal{M} = (\mathcal{X}, \mathcal{A}, \mathcal{Z}, \{\boldsymbol{P}_a\}, \{\boldsymbol{O}_a\}, c, \gamma)$ denote a 3-state, 2-action POMDP, and suppose that $\mathcal{M}$ was solved using value iteration. Further suppose that, upon convergence, the algorithm returned the following set of α-vectors:

$$\Gamma^* = \Big\{ \begin{bmatrix} 0.5 & 0.7 & 0.9 \end{bmatrix}^\top, \begin{bmatrix} 0.6 & 0.7 & 0.9 \end{bmatrix}^\top, \\ \begin{bmatrix} 1.4 & 0.3 & 0.5 \end{bmatrix}^\top, \begin{bmatrix} 0.0 & 0.0 & 1.2 \end{bmatrix} \Big\}.$$

The first and last α-vectors correspond to action 1, while the second and third correspond to action 2.

a) Does the above set correspond to a parsimonious representation of J^* for the POMDP? Explain.

b) What is the optimal action for $\boldsymbol{b} = \begin{bmatrix} 0.05 & 0.9 & 0.05 \end{bmatrix}^\top$?

Exercise 28.
Let $\mathcal{M} = (\mathcal{X}, \mathcal{A}, \mathcal{Z}, \{\boldsymbol{P}_a\}, \{\boldsymbol{O}_a\}, r, \gamma)$ denote a 2-state, 2-action POMDP, and suppose that $\mathcal{M}$ was solved using value iteration. Further suppose that, upon convergence, the algorithm returned the following set of α-vectors:

$$\Gamma^* = \Big\{ \begin{bmatrix} 0.5 & 1.125 \end{bmatrix}^\top, \begin{bmatrix} 0.8 & 0.8 \end{bmatrix}^\top, \begin{bmatrix} 2.8 & 0.4 \end{bmatrix} \Big\}.$$

The first and last α-vectors correspond to action 1, while the second corresponds to action 2.

a) Graphically depict the value function J^* corresponding to the set Γ^* above.

b) What is the optimal action for $\boldsymbol{b} = \begin{bmatrix} 0.25 & 0.75 \end{bmatrix}^\top$?

Exercise 29.
Consider a 2-state POMDP $\mathcal{M} = (\mathcal{X}, \mathcal{A}, \mathcal{Z}, \{\boldsymbol{P}_a\}, \{\boldsymbol{O}_a\}, c, \gamma)$ with

- $\mathcal{X} = \{0, 1\}$;
- $\mathcal{A} = \{Stay, Move\}$;
- $\mathcal{Z} = \{0, 1\}$;
- The transition probabilities are given by

$$\boldsymbol{P}_{Stay} = \begin{bmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{bmatrix} \qquad \boldsymbol{P}_{Move} = \begin{bmatrix} 0.1 & 0.9 \\ 0.9 & 0.1 \end{bmatrix}.$$

- The observation probabilities are given by

$$\boldsymbol{O}_{Stay} = \boldsymbol{O}_{Move} = \begin{bmatrix} 0.6 & 0.4 \\ 0.4 & 0.6 \end{bmatrix}$$

- The cost function is given by

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

- The discount is $\gamma = 0.99$.

Suppose that, after 2 iterations of VI, the cost-to-go function is represented by the set of α-vectors:

$$\Gamma^{(2)} = \left\{ \begin{bmatrix} 2.69 \\ 0.28 \end{bmatrix}, \begin{bmatrix} 1.28 \\ 1.69 \end{bmatrix}, \begin{bmatrix} 2.30 \\ 0.51 \end{bmatrix}, \begin{bmatrix} 1.51 \\ 1.30 \end{bmatrix} \right\}$$

a) Graphically depict the cost-to-go function $J^{(2)}$ corresponding to the set $\Gamma^{(2)}$ above.

b) Compute the value $J^{(3)}(\boldsymbol{b})$, where $\boldsymbol{b}$ is the belief $\boldsymbol{b} = [0.7, 0.3]^\top$ and $J^{(3)}$ is the updated value after 3 iterations.

Exercise 30.
Consider a POMDP $\mathcal{M} = (\mathcal{X}, \mathcal{A}, \mathcal{Z}, \{\boldsymbol{P}_a\}, \{\boldsymbol{O}_a\}, c, \gamma)$ where

- $\mathcal{X} = \{1, 2, 3, 4\}$;
- $\mathcal{A} = \{a, b, c\}$;
- $\mathcal{Z} = \{1, I, 4\}$;
- The transition probabilities are given by

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

- The observation probabilities are given by

$$\boldsymbol{O}_a = \boldsymbol{O}_b = \boldsymbol{O}_c = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

- The cost function is given by

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

- The discount is $\gamma = 0.9$.

a) Compute the optimal Q-function for the underlying MDP.

b) Indicate the action prescribed by the MLS heuristic for $\boldsymbol{b} = [0, 0.55, 0.45, 0]^\top$.

c) Indicate the action prescribed by the Q-MDP heuristic for $\boldsymbol{b} = [0, 0.55, 0.45, 0]^\top$.

d) Do you think that any of the two heuristics above is optimal for $\boldsymbol{b} = [0, 0.55, 0.45, 0]^\top$? Explain.

7 Supervised Learning

Exercise 31.
Consider the following problem:

- An agent must select a prediction $\hat{a} \in \mathcal{A}$, following some policy π (note that there is no state);
- At the same time, the environment "selects" a random outcome $\mathrm{a} \in \mathcal{A}$, following some unknown distribution p.
- After the previous steps, the agent and the environment simultaneously disclose the prediction $\hat{a}$ and the outcome a, respectively.
- The policy π is evaluated using a cost function $Q^\pi : \mathcal{A} \to \mathbb{R}$ that "scores" π given the action a.

Assume that the agent is provided with a data set

$$\mathcal{D} = \{a_1, \ldots, a_N\},$$

containing examples of previous outcomes selected by the environment. For simplicity, assume that $\mathcal{A} = \{0, 1\}$ and let

$$N_0 = \sum_{n=1}^{N} \mathbb{I}(a_n = 0) \qquad\qquad N_1 = \sum_{n=1}^{N} \mathbb{I}(a_n = 1).$$

We write

$$\pi_1 = \mathbb{P}_\pi [\hat{a} = 1] \qquad\qquad \pi_0 = \mathbb{P}_\pi [\hat{a} = 0] = 1 - \pi_1.$$

The value of the policy π is then given by

$$J^\pi = \mathbb{E}_{\mu_D} [Q^\pi(\mathrm{a})],$$

where the expectation is on a, the random action selected by nature, and the goal of the agent is to determine a policy π minimizing J^π. Since the underlying distribution μ_D is unknown, the agent must instead rely on the data set $\mathcal{D}$, letting

$$J^\pi \approx \hat{J}^\pi_N = \frac{1}{N} \sum_{n=1}^{N} Q^\pi(a_n).$$

a) Suppose that the cost Q^π is given by

$$Q^\pi(a) = (a - \pi_1)^2.$$

Show that the best policy (i.e., the policy minimizing J_N^π) is obtained by setting

$$\pi_1 = \frac{N_1}{N_0 + N_1}.$$

b) Suppose that the cost is given by

$$Q^\pi(a) = |a - \pi_1|.$$

Show that the best policy is obtained by setting $\pi_1 = \text{median}(\mathcal{D})$.

Exercise 32.
Consider the problem of learning the XOR function. The XOR function can be summarized as

	$\phi_1(x)$	$\phi_2(x)$	Action
x_1	-1	-1	-1
x_2	-1	$+1$	$+1$
x_3	$+1$	-1	$+1$
x_4	$+1$	$+1$	-1

and is represented in Fig. 7.

a) Suppose that you want to learn the XOR function using a linear classifier such as logistic regression or an SVM. Is the data linearly separable? Explain.

b) Show that it is possible to replace the feature ϕ_2 by an alternative feature ϕ_2', computed from ϕ_1 and ϕ_2, and such that the XOR function is learnable by a linear classifier. **Suggestion:** Determine one such feature.

c) Using ϕ_1 and the feature ϕ_2' computed in (b), represent the new XOR data in a scatter plot.

d) Sketch in your plot the decision boundary obtained from the SVM classifier for the representation in (c). In your plot, indicate the margin and the support vectors.

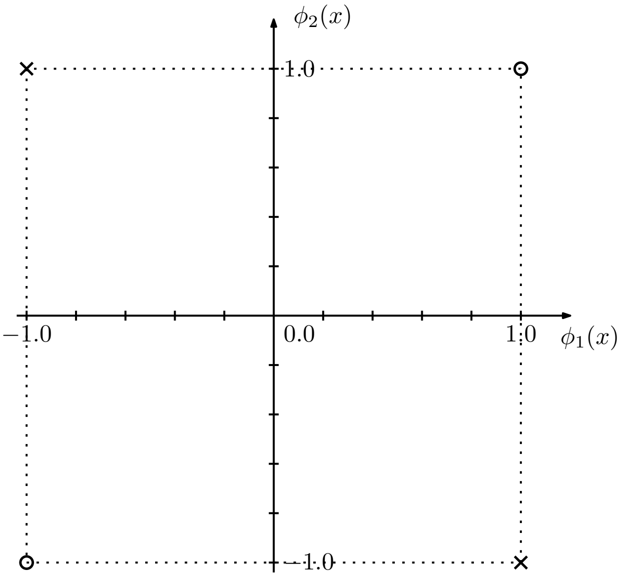


Figure 7: The XOR function.

Exercise 33.
Suppose that, unlike all situations considered so far, we want to learn a continuous function $f^*(x)$ from a dataset

$$\mathcal{D} = \{(x_1, y_1), \ldots, (x_N, y_N)\},$$

where each y_n is the image of x_n by the target function, eventually perturbed by noise, i.e.,

$$y_n = f^*(x_n) + \varepsilon.$$

Further suppose that, in our approach, we again represent the state x by a vector of features, $\boldsymbol{\phi}(x) = [\phi_1(x), \ldots, \phi_M(x)]^\top$. We can approximate the value of f^* by a *linear combination of the features* ϕ_m, i.e.,

$$f^*(x) \approx \boldsymbol{\phi}^\top(x)\boldsymbol{w} = \sum_{m=1}^{M} \phi_m(x) w_m.$$

Adopting such representation, the task of learning f^* now reduces to the task of learning the vector of parameters $\boldsymbol{w}$ that yields the best representation for f^*.

Suppose that we want to select the parameter $\boldsymbol{w}^*$ that minimizes

$$J(\boldsymbol{w}) = \frac{1}{N} \sum_{n=1}^{N} \big(y_n - \boldsymbol{\phi}^\top(x_n)\boldsymbol{w}\big)^2.$$

Show that

$$\boldsymbol{w}^* = \left(\sum_{n=1}^{N} \boldsymbol{\phi}^\top(x_n)\boldsymbol{\phi}(x_n)\right)^{-1} \sum_{n=1}^{N} y_n \boldsymbol{\phi}(x_n).$$

Exercise 34.
Consider the following data

	$\phi_1(x)$	$\phi_2(x)$	Action
x_1	-0.75	0.94	1
x_2	0.40	-0.51	0
x_3	-0.60	0.65	1
x_4	0.94	0.91	0
x_5	0.45	0.30	0
x_6	-0.50	0.74	1
x_7	0.60	-0.90	0
x_8	-0.80	0.61	1
x_9	0.22	0.75	0
x_{10}	-0.11	-0.20	1

a) Represent the data above in a *scatter plot.* Data points corresponding to action $a = 1$ should be represented with the symbol "$\circ$", and the remaining data points with the symbol "$\times$".

b) Suppose that you want to train a logistic regression classifier with the data above using Newton's method. Suppose that, after a number of iterations, the weight vector is

$$\boldsymbol{w} = \begin{bmatrix} 0.1379 \\ -5.7409 \\ -0.7640 \end{bmatrix},$$

where the first component corresponds to the weight associated with the bias term. Represent the decision boundary corresponding to the vector above in your scatter plot.

c) Starting with the vector $\boldsymbol{w}$ in (b), compute the updated vector $\boldsymbol{w}'$ obtained after one update of Netwon's method.

d) Is the data linearly separable? Explain.

e) Sketch in your scatter plot the decision boundary obtained from the SVM classifier for the representation in (c). Indicate the margin and the support vectors.

8 Reinforcement Learning

Exercise 35.

Suppose that an agent is modeled as a 3-state, 2-action MDP. Suppose that the agent follows a fixed policy π but, since the MDP parameters are unknown, the agent is unable to evaluate the quality of the policy analytically. Instead, it executes the policy for a number of steps, and experienced the following trajectories:

$$\begin{aligned}
T_1 &= \{(1, b, 0.5, 3), (3, a, 1.0, 3)\}\\
T_2 &= \{(1, a, 1.0, 1), (1, a, 1.0, 1), (1, a, 1.0, 1), (1, a, 1.0, 1),\\
&\qquad (1, a, 1.0, 1), (1, b, 0.5, 1), (1, a, 1.0, 2), (2, a, 0.0, 2)\}\\
T_3 &= \{(1, a, 1.0, 2), (2, a, 0.0, 2)\}\\
T_4 &= \{(1, a, 1.0, 1), (1, a, 1.0, 2), (2, b, 0.0, 2)\},
\end{aligned}$$

where each 4-tuple corresponds to a transition. Using a Monte Carlo approach (as done in Lab 2), compute the value of J^π for every state. To do so, approximate the value of J^π at each state $x \in \mathcal{X}$ as the average total discounted value over all trajectories starting in x.

Exercise 36.

Consider an agent moving in a 7-room environment. The agent has available, at each time-step, three actions, a, b and c. The agent is following a random policy and using a model-based approach to compute Q^*. After 20 time steps, you have collected the following data:

$$\hat{\boldsymbol{P}}_a = \begin{bmatrix}
0.11 & 0.61 & 0.28 & 0.00 & 0.00 & 0.00 & 0.00\\
0.00 & 0.15 & 0.00 & 0.00 & 0.00 & 0.00 & 0.85\\
0.00 & 0.00 & 0.17 & 0.00 & 0.00 & 0.83 & 0.00\\
0.60 & 0.00 & 0.00 & 0.40 & 0.00 & 0.00 & 0.00\\
0.60 & 0.00 & 0.00 & 0.00 & 0.40 & 0.00 & 0.00\\
0.80 & 0.00 & 0.00 & 0.00 & 0.00 & 0.20 & 0.00\\
0.75 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.25
\end{bmatrix}$$

$$\hat{\boldsymbol{P}}_b = \begin{bmatrix}
0.15 & 0.46 & 0.39 & 0.00 & 0.00 & 0.00 & 0.00\\
0.00 & 0.17 & 0.00 & 0.00 & 0.00 & 0.83 & 0.00\\
0.00 & 0.00 & 0.29 & 0.00 & 0.00 & 0.00 & 0.71\\
0.33 & 0.00 & 0.00 & 0.67 & 0.00 & 0.00 & 0.00\\
0.50 & 0.00 & 0.00 & 0.00 & 0.50 & 0.00 & 0.00\\
0.75 & 0.00 & 0.00 & 0.00 & 0.00 & 0.25 & 0.00\\
0.60 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.40
\end{bmatrix}$$

$$\hat{\boldsymbol{P}}_c = \begin{bmatrix}
0.10 & 0.43 & 0.47 & 0.00 & 0.00 & 0.00 & 0.00\\
0.00 & 0.22 & 0.00 & 0.78 & 0.00 & 0.00 & 0.00\\
0.00 & 0.00 & 0.18 & 0.00 & 0.82 & 0.00 & 0.00\\
0.00 & 0.60 & 0.00 & 0.40 & 0.00 & 0.00 & 0.00\\
0.00 & 0.00 & 0.67 & 0.00 & 0.33 & 0.00 & 0.00\\
0.75 & 0.00 & 0.00 & 0.00 & 0.00 & 0.25 & 0.00\\
0.78 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.22
\end{bmatrix},$$

$$\hat{c} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ 0 & 0 & 0 \\ 1 & 1 & 1 \end{bmatrix}$$

and

$$N = \begin{bmatrix} 17 & 12 & 20 \\ 12 & 11 & 8 \\ 11 & 6 & 10 \\ 4 & 2 & 4 \\ 4 & 3 & 5 \\ 9 & 7 & 7 \\ 7 & 4 & 8 \end{bmatrix} \qquad \hat{Q} = \begin{bmatrix} 2.63 & 2.72 & 2.83 \\ 2.49 & 2.54 & 1.47 \\ 2.55 & 2.51 & 2.54 \\ 2.03 & 0.54 & 1.13 \\ 2.35 & 1.73 & 2.59 \\ 1.86 & 1.98 & 1.92 \\ 2.71 & 1.78 & 2.75 \end{bmatrix}.$$

Suppose that the agent is in state 1, executes action c and moves to state 3, and pays a cost of 1 in the process. Assuming that $\gamma = 0.95$, compute a step of model-based reinforcement learning, updating the estimates $\hat{\boldsymbol{P}}$, $\hat{c}$ and $\hat{Q}$.

Exercise 37.

Consider the fully observable version of the problem introduced in Question 26. Suppose that, after a number of steps of a reinforcement learning algorithm, the agent has the following estimate for Q^*:

$$\hat{\boldsymbol{Q}} = \begin{bmatrix} 2.19 & 2.43 & 1.80 & 2.36 \\ 1.66 & 1.77 & 1.20 & 2.24 \\ 1.11 & 2.27 & 0.66 & 1.76 \\ 0 & 0 & 0 & 0 \\ 2.27 & 2.97 & 2.60 & 2.65 \\ 2.00 & 3.06 & 5.18 & 2.21 \\ 6.28 & 6.13 & 6.37 & 6.22 \\ 2.69 & 3.01 & 3.26 & 2.97 \\ 3.17 & 3.12 & 2.96 & 2.97 \\ 2.57 & 3.01 & 3.29 & 3.15 \\ 5.79 & 3.28 & 3.58 & 3.08 \end{bmatrix}$$

a) Suppose that the agent experiences a transition $(1, S, 0.5, 5)$. Perform a Q-learning update to the corresponding value, using $\alpha = 0.1$.

b) Suppose that the agent experiences a transition $(5, N, 0.5, 1)$ while following an ε-greedy policy, with $\varepsilon = 0.01$. Perform a SARSA update to the corresponding value, assuming that the action taken in the resulting state after the transition (i.e., state 1) is not exploratory.

Exercise 38.

Consider an agent moving in a 4-state, 3-action MDP following some policy π. Suppose that, after a number of steps of TD-learning, the agent has the following estimate for J^π.

$$\hat{\boldsymbol{J}} = \begin{bmatrix} 0.99 \\ 2.38 \\ 2.21 \\ 0.61 \end{bmatrix}.$$

Suppose that the agent experiences the transitions $(1, 1.0, 2)$, $(2, 0.0, 2)$ and $(2, 1.0, 1)$, where each transition is a triplet (x, c, y), with $x, y \in \mathcal{X}$ and $c \in \mathbb{R}$. Perform the corresponding TD-learning updates.

Exercise 39.

Explain the role of λ in the TD(λ) algorithm.

9 Exploration × Exploitation

Exercise 40.
Consider a multi-armed bandit setup where $\mathcal{A} = \{a_1, a_2\}$. Further suppose that, after some attempts, an agent estimates that $\hat{c}(a_1) = 0.9$. On the other hand, it has never experienced action a_2, so it is unsure if a_2 awards any reward at all. Explain the exploration vs exploitation tradeoff in the context of this problem, and explain how UCB addresses such tradeoff.

Exercise 41.
Again consider the situation in Question 40, and suppose that, after t attempts, $\hat{c}(a_1) = 0.7$ and $\hat{c}(a_2) = 0.6$. Further suppose that

$$\sqrt{\frac{\log t}{N_t(a_1)}} = 0.1 \qquad\qquad \sqrt{\frac{\log t}{N_t(a_2)}} = 0.3,$$

where we denote by $N_t(a)$ the number of times that action a was played after the t attempts. If the agent is following UCB, what action will it select next? Explain.

Exercise 42.
Explain the need for non-deterministic policies in the adversarial bandit setup.

Exercise 43.
Consider an agent that must, at each time-step t, select one of three possible actions: a_1, a_2 and a_3. After selecting the action, the agent pays a cost c_t that depends on the action selected at time-step t. Suppose that, for $t = 1, \ldots, 10$, the agent selected the actions:

$$a_3 \qquad a_1 \qquad a_3 \qquad a_1 \qquad a_3 \qquad a_2 \qquad a_1 \qquad a_1 \qquad a_2 \qquad a_2$$

and incurred the costs

a_1	0.15	0.11	0.24	0.00	0.13	0.22	0.06	0.32	0.15	0.23
a_2	0.43	0.74	0.73	1.00	0.81	0.82	0.52	0.66	0.59	0.74
a_3	1.00	0.99	0.00	0.00	0.39	0.15	0.40	0.11	0.38	0.84

a) Indicate the action selected by the agent at time-step $t = 11$ if the agent follows the UCB algorithm.

b) Estimate the regret incurred by the agent up to $T = 10$.

c) Compare the regret computed in (b) with the theoretical bound for UCB. Comment.

Exercise 44.
Consider an agent that must, at each time-step t, select one of three possible actions: a_1, a_2 and a_3. After selecting the action, the agent pays a cost c_t that depends on the action selected at time-step t. Suppose that, for $t = 1, \ldots, 10$, the agent selected the actions:

$$a_1 \quad a_2 \quad a_3 \quad a_2 \quad a_1 \quad a_3 \quad a_2 \quad a_2 \quad a_1 \quad a_3$$

and incurred the costs

a_1	0	1	0	0	1	1	1	0	0	0
a_2	1	0	1	0	0	0	1	1	0	1
a_3	1	0	1	1	0	0	0	0	0	1
a_4	1	1	1	1	0	1	1	1	0	1

a) Indicate the probability of each action at time-step $t = 11$ if the agent follows the EXP3 algorithm, with $\eta = 0.1$.

b) Indicate the probability of each action at time-step $t = 11$ if the agent follows the EWA with $\eta = 0.9$. Compare the resulting policy with that obtained in (a).

c) Compute the regret incurred by the agent up to $T = 10$.

Exercise 45.
Explain the main differences between the EWA and EXP3, indicating in which situations each of the two algorithms is more adequate.