Online and Reinforcement Learning

2025-2026

Home Assignment 1

Yevgney Seldin **Sadegh Talebi**
Department of Computer Science
University of Copenhagen

The deadline for this assignment is **11 February 2025, 20:59**. You must submit your *individual* solution electronically via the Absalon home page.

A solution consists of:

- A PDF file with detailed answers to the questions, which may include graphs and tables if needed. Do *not* include your full source code in the PDF file, only selected lines if you are asked to do so.
- A .zip file with all your solution source code with comments about the major steps involved in each question (see below). Source code must be submitted in the original file format, not as PDF. The programming language of the course is Python.

Important Remarks:

- IMPORTANT: Do NOT zip the PDF file, since zipped files cannot be opened in *SpeedGrader*. Zipped PDF submissions will not be graded.
- Your PDF report should be self-sufficient. I.e., it should be possible to grade it without opening the .zip file. We do not guarantee opening the .zip file when grading.
- Your code should be structured such that there is one main file (or one main file per question) that we can run to reproduce all the results presented in your report. This main file can, if you like, call other files with functions, classes, etc.
- Handwritten solutions will not be accepted.

1 Find an online learning problem from real life `(0 points) [Yevgeny]`

Solve Exercise 7.1 in "Machine Learning. The science of selection under uncertainty" (Seldin, 2025).

*** We would like to save your time on writing down the solution. So you do not need to submit your answer to the question, but we do expect that you think about it seriously and “solve it for yourself”. (In contrast to “Optional” questions at the end of the assignment, this question is not “optional”, it is just not for reporting.) ***

2 Follow The Leader (FTL) algorithm for i.i.d. full information games (25 points) [Yevgeny]

Solve Exercise 7.2 in “Machine Learning. The science of selection under uncertainty” (Seldin, 2025).

3 Improved Parametrization of UCB1 (25 points) [Yevgeny]

Exercise 7.5, Part 2 [“Write a simulation ...”] in “Machine Learning. The science of selection under uncertainty” (Seldin, 2025). (Part 1 of the exercise will be given later.)

4 Example of Policies in RiverSwim (24 points) [Sadegh]

Part 1. Consider the following policies defined in the RiverSwim MDP (Figure 1). For each case, determine to which class the policy belongs (i.e., $\Pi^{\mathsf{SD}}, \Pi^{\mathsf{SR}}, \Pi^{\mathsf{HD}}, \Pi^{\mathsf{HR}}$). Provide a short explanation and state any assumptions that you may make.

(i) π_1 defined as: Swim to the right if the current state is not 1; otherwise swim to the left.

(ii) π_2 defined as: If time slot t is an even integer, swim to the right; otherwise, flip a fair coin, then swim to the right (resp. left) if the outcome is ‘Head’ (resp. ‘Tail’).

(iii) π_3 defined as: At $t = 1$, swim to the left. For $t > 1$, swim to the right if the index of the previous state is odd; otherwise swim to the left. (For $t = 1$, swim to the left.)

(iv) π_4 defined as: Flip a fair coin. If the outcome is ‘Head’ and the current state is either $L - 1$ or L, then swim to the right; otherwise swim to the left.

(v) π_5 defined as: If it rains, swim to the right; otherwise, swim to the left. (It rains independently of the agent’s swimming direction and position.)

Part 2. Make an arbitrary example of a policy in Π^{SR} in RiverSwim. The policy must not be deterministic.

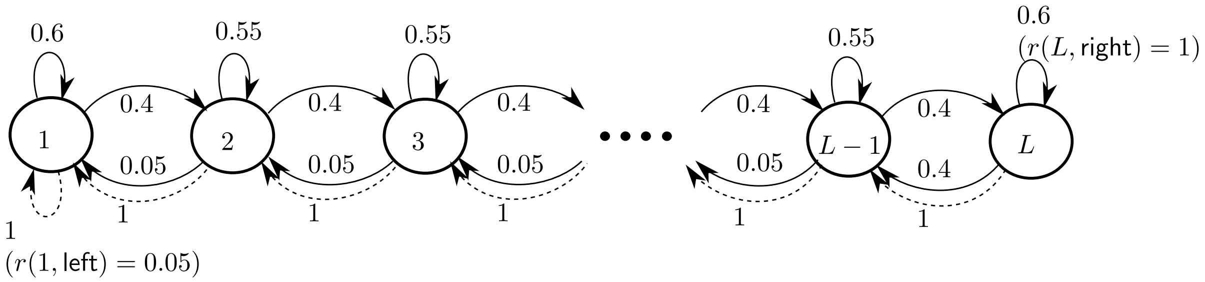


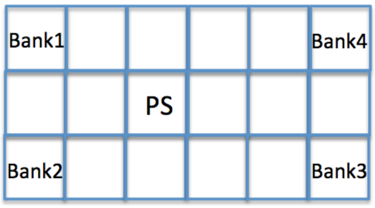
Figure 1: The L-state RiverSwim MDP (Strehl and Littman, 2008)

5 Robbing Banks (26 points) [Sadegh]

In this exercise, we model a robber chasing game using the MDP framework.

At time 1, an agent is robbing Bank 1 (see Figure 2). Then, the police gets alerted and starts chasing her from the point PS (Police Station). The agent observes where the police is, and decides in each step either to move up, left, right, down or to stay where she is. Each time the agent is at a bank and the police is not there, she collects a reward of DKK 100,000. If the police catches her, she will lose DKK 10,000, and the game will be restarted: She is brought back to Bank 1, and the police goes back to the PS. The rewards are discounted at a rate $\gamma \in (0, 1)$.

The police always chases the agent, but moves randomly as follows. When the police is on the same line or column as the agent, it moves with probability 1/3 toward the agent, and with probability 1/3 in each of the two orthogonal directions (up or down in case of the same line; left or write in case of the same column), but never in the opposite direction from the agent. When the police and the agent are neither on the same line, nor on the same column, then the police moves up, down, right, or left with probability 1/4. (We assume that walls act as reflectors, similarly to the grid-world.) The agent's objective is to maximize her expected cumulative discounted reward.



Bank1					Bank4
		PS			
Bank2					Bank3

Figure 2: The city

We would like to formulate this problem as an MDP.

(i) Indicate a suitable notion of state (i.e., one that could be used to define an MDP), and indicate the corresponding state-space. Indicate the corresponding action-space.

(ii) Specify the reward function for the chosen notions of state and action.

(iii) For the chosen state-action notion, specify the transition probabilities (as a function of the agent's action) when the agent is at Bank 1 and the police is at Bank 4.

The questions below are optional. You are very welcome to work on them, but we do not expect that everyone does it, and you should not put them in your report.

6 The worst case gap for a fixed T (0 points) [Yevgeny] (Optional)

Solve Exercise 7.4 in "Machine Learning. The science of selection under uncertainty" (Seldin, 2025).

7 Decoupling exploration and exploitation in i.i.d. multiarmed bandits (0 points) [Yevgeny] (Optional)

Solve Exercise 7.3 in "Machine Learning. The science of selection under uncertainty" (Seldin, 2025).

References

Yevgeny Seldin. Machine Learning. The science of selection under uncertainty. `https://arxiv.org/abs/2509.21547`, 2025.

Alexander L Strehl and Michael L Littman. An analysis of model-based interval estimation for Markov decision processes. *Journal of Computer and System Sciences*, 74(8): 1309–1331, 2008.