Online and Reinforcement Learning

2025-2026

Home Assignment 2

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

The deadline for this assignment is **18 February 2026, 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 Short Questions (8 points) [Sadegh]

Determine whether each statement below is True or False and provide a very brief justification.

1. In a finite discounted MDP, every possible policy induces a Markov Reward Process.

   □ True □ False

   Justification:

2. Consider a finite discounted MDP, and assume that π is an optimal policy. Then, the action(s) output by π does not depend on history other than the current state (i.e., π is necessarily stationary).

   □ True □ False

   Justification:

3. In a finite discounted MDP, a greedy policy with respect to optimal action-value function, $Q^\star$, corresponds to an optimal policy.

   □ True □ False

   Justification:

4. Policy Iteration (`PI`) may return a near-optimal policy.

   □ True □ False

   Justification:

2 Introduction of New Products `(25 points) [Yevgeny]`

Solve Exercise 7.6 in (Seldin, 2025).

3 Empirical comparison of FTL and Hedge `(25 points) [Yevgeny]`

Solve Exercise 7.9 in (Seldin, 2025).

4 Value Function Bounds `(22 points) [Sadegh]`

In this exercise, we study a classical result that concerns the difference in value functions between two MDPs that are defined on the same state-action space, and whose transition and reward functions are close in some sense.

Consider two finite discounted MDPs $M_1 = (\mathcal{S}, \mathcal{A}, P_1, R_1, \gamma)$ and $M_2 = (\mathcal{S}, \mathcal{A}, P_2, R_2, \gamma)$. Assume the two reward functions take values in the range $[0, 1]$. Suppose that for all state-action pairs (s, a),

$$|R_1(s,a) - R_2(s,a)| \le \alpha, \qquad \|P_1(\cdot|s,a) - P_2(\cdot|s,a)\|_1 \le \beta$$

for some numbers $\alpha > 0$ and $\beta > 0$. Then, for all stationary deterministic policy π and state-action pair (s, a),

$$|Q_1^\pi(s,a) - Q_2^\pi(s,a)| \leq \frac{\alpha}{1-\gamma} + \frac{\gamma\beta}{(1-\gamma)^2}.$$

(i) Prove the inequality.

(ii) Briefly interpret this result in words. What does it tell us qualitatively?

Hint: By using the properties of value functions, one may arrive at

$$|Q_1^\pi(s,a) - Q_2^\pi(s,a)| \leq \ldots + \square \max_{s',a'} |Q_1^\pi(s',a') - Q_2^\pi(s',a')|,$$

where all elements in the right-hand side is independent of (s, a), and where $\square$ hides a problem parameter. Then, taking the maximum in the left-hand side over (s, a) could be helpful to conclude the proof.

5 Solving a Discounted Grid-World (20 points) [Sadegh]

In this exercise, we model a grid-world game as a discounted MDP, and solve it using PI and VI.

Consider the 4-room Grid-World MDP shown in Figure 1. It is made of a grid of size 7×7, which has $S = 20$ accessible states (after removing walls). The agent starts in the upper-left corner (shown in red). A reward of 1 is placed in the lower-right corner (shown in yellow), and the rest of the states give no reward. Once in the rewarding state (in yellow), the agent stays there forever (and continues receiving the reward). The agent can perform the 4 compass actions going up, left, down, or right (of course, when away from walls). However, the floor is slippery and brings stochasticity to the next-state. Specifically, under each of the four aforementioned actions, she moves in the chosen direction (with probability 0.7), stays in the same state (with probability 0.1), or goes in each of the two perpendicular directions (each with probability 0.1) —this environment is sometimes referred to as the *frozen lake* MDP. Walls act as reflectors, i.e., they cause moving back to the current state. A Python implementation of this MDP is provided in HA2_gridworld.py. Rewards are discounted with rate $\gamma = 0.97$.

We can model this task as a discounted MDP.

(i) Solve the grid-world task above using PI. (You may use the Python implementation of PI provided in the same file.) Report an optimal policy along with the optimal value function $V^\star$. Furthermore, visualize the derived optimal policy using arrows in the figure or by arranging it using a suitably defined matrix.

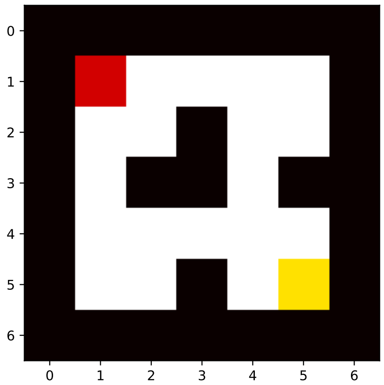

Figure 1: The 4-Room Grid-World MDP

(ii) Implement `VI` and use it to solve the grid-world task above. Your implementation should receive an MDP M and an accuracy parameter ε as input, and output a policy and the corresponding value. (Note that to ensure that `VI` returns an optimal policy, ε must be sufficiently small; here, $\varepsilon = 10^{-6}$ suffices.)

(iii) Repeat Part (ii) with $\gamma = 0.998$ and discuss how this new discount affects the convergence speed of `VI`.

> 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 Improved Parametrization of UCB1 `(0 points)` `[Yevgeny]` (Optional, but highly recommended)

Solve Exercise 7.5 Part 1 in (Seldin, 2025).

References

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