Online and Reinforcement Learning

2025-2026

Home Assignment 3

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

The deadline for this assignment is **25 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 Empirical evaluation of algorithms for adversarial environments `(0 points) [Yevgeny]`

Solve Exercise 7.7 in (Seldin, 2025).

You do not need to submit your solution to this question, but we expect that you think seriously about it and "solve it for yourself".

2 A tighter analysis of the Hedge algorithm (20 points) [Yevgeny]

Solve Exercise 7.8 in (Seldin, 2025).

3 Empirical comparison of UCB1 and EXP3 algorithms (30 points) [Yevgeny]

Solve Exercise 7.12 in (Seldin, 2025).

4 Off-Policy Evaluation (28 points) [Sadegh]

In this exercise, we study off-policy evaluation for an episodic game in the RiverSwim MDP. The setting is as follows: the agent starts in state 1 (left-most state). she receives a terminal reward *once*, regardless of the action taken, and the episode terminates. Rewards within each episode are discounted at rate γ. When a new episode begins, the discounting process resets.

We have run this game for 200 episodes using actions generated by a behavior policy π_{b} defined as follows:

$$\pi_{\mathrm{b}}(s) = \begin{cases} \mathsf{right} & \text{w.p. } 0.65 \\ \mathsf{left} & \text{w.p. } 0.35 \end{cases}, \qquad \text{for all } s \in \mathcal{S}.$$

The resulting data are recorded in a dataset $\mathcal{D}$ (provided as `dataset0.csv`). Specifically, $\mathcal{D}$ contains 200 episodes, each corresponding to a trajectory τ_i (following the notation from the slides). However, the dataset is stored as a single long trajectory in which all episodes are concatenated. Your task is to reconstruct the individual trajectories $\{\tau_i\}$; however, note that in $\mathcal{D}$ states are encoded as $0, 1, \ldots, L-1$.

Consider the policy with π defined as:

$$\pi(s) = \begin{cases} \mathsf{right} & \text{w.p. } 0.95 \\ \mathsf{left} & \text{w.p. } 0.05 \end{cases}, \qquad \text{for all } s \in \mathcal{S}.$$

Our goal is to estimate the value of π at the initial state, $V^{\pi}(1)$, using the dataset $\mathcal{D}$. The discount factor is $\gamma = 0.97$.

(i) Estimate $V^{\pi}(1)$ using the following methods: `CE-OPE`, `IS`, `wIS`, and `PDIS`. Report the four resulting value estimates: $V^{\pi}_{\texttt{CE-OPE}}(1)$, $V^{\pi}_{\texttt{IS}}(1)$, $V^{\pi}_{\texttt{wIS}}(1)$, $V^{\pi}_{\texttt{PDIS}}(1)$.

(ii) For each method, compute and report the estimation error $|V^\pi(1) - \widehat{V}^\pi(1)|$, where $V^\pi = (I - \gamma P^\pi)^{-1} r^\pi$ denotes the true value of π, and where $\widehat{V}^\pi$ denotes a value estimate.

(iii) Now consider 10 additional datasets `dataset1.csv`,...,`dataset9.csv` generated in a similar fashion. For a given OPE method, each dataset yields an estimate of $V^\pi(1)$. Report the empirical variance of these estimates across the 10 datasets for each of the methods mentioned in Part (i).

(iv) Briefly compare the 4 methods above in terms of empirical error and variance.

(Hint: I will discuss some hints in the class.)

5 Bellman Operators `(16 points) [Sadegh]`

The optimal Bellman operator (for Q-values) is denoted by $\mathcal{T}$ and defined as follows: For $f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, for all (s, a), for all (s, a),

$$\mathcal{T} f(s, a) := R(s, a) + \gamma \sum_{x \in \mathcal{S}} P(x|s, a) \max_{b \in \mathcal{A}} f(x, b) \,.$$

(i) Show that $\mathcal{T}$ is a contraction mapping with respect to $\|\cdot\|_\infty$.

(ii) What can be said about the properties of the fixed-point of $\mathcal{T}$? *(Hint: See Banach's fixed-point theorem.)*

6 Short Questions `(6 points) [Sadegh]`

(i) The model-based off-policy evaluation method introduced in the course is based on a key principle. State the name of this principle and briefly explain its main idea.

(ii) Briefly discuss whether the following statement is true or false. *Under the coverage assumption, the Weighted Importance Sampling Estimator $\widehat{V}_{wIS}$ converges to V^π with probability* 1.

> 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.

7 The doubling trick `(0 points) [Yevgeny]` (Optional)

Solve Exercise 7.10 in (Seldin, 2025).

8 Rewards vs. losses (0 points) [Yevgeny] (Optional)

Solve Exercise 7.13 in (Seldin, 2025).

9 Regularization by relative entropy and the Gibbs distribution (0 points) [Yevgeny] (Optional)

Solve Exercise 7.11 in (Seldin, 2025).

References

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