Reinforcement Learning
Prof. Milica Gašić

Release: Thursday, 16 October 2025
Discussion: Wednesday, 22 October 2025

Exercise set #1

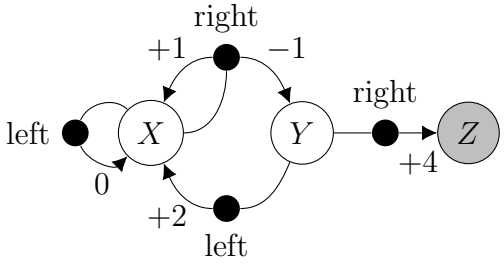
You do not have to hand in your solutions to the exercises and they will **not** be graded. However, there will be four short tests during the semester. You need to reach ≥40% of the total points in order to be admitted to the final exam (Klausur). The tests are held at the start of a lecture (room 25.22.U1.74) at the following dates:

Test 1: Thursday, 6 November 2025, 10:30-10:45
Test 2: Thursday, 27 November 2025, 10:30-10:45
Test 3: Thursday, 11 December 2025, 10:30-10:45
Test 4: Thursday, 15 January 2026, 10:30-10:45

Please ask questions in the RocketChat
The exercises are discussed every Wednesday, 14:30-16:00 in room 25.12.02.33.

1. **Three state MDP**[1]

   Consider the MDP below, in which there are three states, $\mathcal{S} = \{X, Y, Z\}$, two actions, $\mathcal{A} = \{\text{left}, \text{right}\}$, and the rewards on each transition are as indicated by the numbers. Note that if action *right* is taken in state X, then the transition may be either to X with a reward of $+1$ or to Y with a reward of -1. These two possibilities occur with probabilities 0.75 (for the transition to X) and 0.25 (for the transition to state Y). The state Z is a terminal state, i.e., all transitions from Z are back to Z with a reward of 0. The initial state is always X.



   (a) Write down the initial state distribution $\mathcal{P}_0$.

   (b) For what combinations of inputs $s, s' \in \mathcal{S}$, $a \in \mathcal{A}$, $r \in \{4, 2, 1, 0, -1\}$ is the dynamics distribution $p(s', r|s, a)$ of this MDP non-zero? Note that the distribution is discrete since the states, actions, and rewards are discrete. Write down the probabilities for these combinations.

   Hint: There should be seven combinations with non-zero probability.

   (c) Write down $\mathcal{P}(s'|s, a)$ and $\mathcal{R}(s, a)$ for all $s, s' \in \mathcal{S}, a \in \mathcal{A}$. The reward function can be derived from the dynamics distribution considered in part (b) using the formula from the lecture.

Exercises by Stefan Harmeling, used with permission

[1]MDP adopted from Richard Sutton's CMPUT 609 course: `http://www.incompleteideas.net/rlai.cs.ualberta.ca/RLAI/RLAIcourse/2009.html`

(d) Consider the two deterministic policies π_1 and π_2:

$$\pi_1(X) = \text{right} \qquad \pi_2(X) = \text{left}$$
$$\pi_1(Y) = \text{right} \qquad \pi_2(Y) = \text{right}$$

Write down a typical trajectory for policy π_1, i.e., make up a sequence of states, actions, and rewards that is likely to occur. What happens if you do this for π_2?

(e) Implement this MDP as a `gym` environment (use `import gymnasium as gym`)[2]. We provide a starting point in the Jupyter notebook[3] `three-state-mdp.ipynb`. Next, implement the deterministic policy π_1 from part (d) and the stochastic policy π_3:

$$\pi_3(\text{left}|X) = 0 \qquad \pi_3(\text{left}|Y) = 0.9$$
$$\pi_3(\text{right}|X) = 1 \qquad \pi_3(\text{right}|Y) = 0.1$$

If you sum all rewards of an episode and average this over many episodes, what values do you get for π_1 and π_3?

[2]For more information on `gym`, visit `https://gymnasium.farama.org`

[3]You can install jupyter notebook as explained here