Reinforcement Learning
Prof. Milica Gašić

Release: Thursday, 10 October 2024
Discussion: Wednesday, 16 October 2024

Exercise set #1

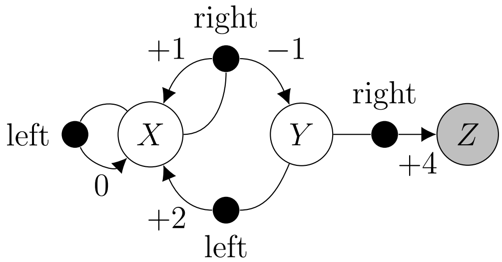
You do not have to hand in your solutions to the exercises and they will **not** be graded. However, there will be four short tests during the semester. You need to reach ≥50% of the total points in order to be admitted to the final exam (Klausur). The tests are held at the start of a lecture (room 2522.U1.74) at the following dates:

Test 1: Thursday, 31 October 2024, 10:30-10:45
Test 2: Thursday, 21 November 2024, 10:30-10:45
Test 3: Thursday, 5 December 2024, 10:30-10:45
Test 4: Thursday, 9 January 2025, 10:30-10:45

Please ask questions in the RocketChat
The exercises are discussed every Wednesday, 14:30-16:00 in room 2512.02.33.

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 $\mathcal{P}_0$.

Answer: $\mathcal{P}_0(X) = 1,\ \mathcal{P}_0(Y) = 0,\ \mathcal{P}_0(Z) = 0$

(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.

Answer:

$$p(X, 0|X, \text{left}) = 1 \qquad p(X, 1|X, \text{right}) = 0.75$$
$$p(Y, -1|X, \text{right}) = 0.25 \qquad p(X, 2|Y, \text{left}) = 1$$
$$p(Z, 4|Y, \text{right}) = 1 \qquad p(Z, 0|Z, \text{left}) = 1$$
$$p(Z, 0|Z, \text{right}) = 1$$

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`

(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.

Answer:

$$\begin{aligned} \mathcal{P}(X|X,\text{left}) &= 1 & \mathcal{P}(X|X,\text{right}) &= 0.75 \\ \mathcal{P}(Y|X,\text{left}) &= 0 & \mathcal{P}(Y|X,\text{right}) &= 0.25 \\ \mathcal{P}(Z|X,\text{left}) &= 0 & \mathcal{P}(Z|X,\text{right}) &= 0 \end{aligned}$$

$$\begin{aligned} \mathcal{P}(X|Y,\text{left}) &= 1 & \mathcal{P}(X|Y,\text{right}) &= 0 \\ \mathcal{P}(Y|Y,\text{left}) &= 0 & \mathcal{P}(Y|Y,\text{right}) &= 0 \\ \mathcal{P}(Z|Y,\text{left}) &= 0 & \mathcal{P}(Z|Y,\text{right}) &= 1 \end{aligned}$$

$$\begin{aligned} \mathcal{P}(X|Z,\text{left}) &= 0 & \mathcal{P}(X|Z,\text{right}) &= 0 \\ \mathcal{P}(Y|Z,\text{left}) &= 0 & \mathcal{P}(Y|Z,\text{right}) &= 0 \\ \mathcal{P}(Z|Z,\text{left}) &= 1 & \mathcal{P}(Z|Z,\text{right}) &= 1 \end{aligned}$$

$$\mathcal{R}(s,a) = \sum_r r \sum_{s'} p(s', r|s,a)$$

$$\begin{aligned} \mathcal{R}(X,\text{left}) &= 0 \cdot 1 = 0 \\ \mathcal{R}(X,\text{right}) &= 0.75 \cdot 1 + 0.25 \cdot (-1) = 0.5 \end{aligned}$$

$$\begin{aligned} \mathcal{R}(Y,\text{left}) &= 1 \cdot 2 = 2 \\ \mathcal{R}(Y,\text{right}) &= 1 \cdot 4 = 4 \end{aligned}$$

$$\begin{aligned} \mathcal{R}(Z,\text{left}) &= 0 \\ \mathcal{R}(Z,\text{right}) &= 0 \end{aligned}$$

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

$$\begin{aligned} \pi_1(X) &= \text{right} & \pi_2(X) &= \text{left} \\ \pi_1(Y) &= \text{right} & \pi_2(Y) &= \text{right} \end{aligned}$$

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?

Answer:

$$\begin{aligned} &\pi_1 : X, \text{right}, 1, X, \text{right}, 1, X, \text{right}, 1, X, \text{right}, -1, Y, \text{right}, 4, Z \\ &\pi_2 : X, \text{left}, 0, X, \text{left}, 0, X, \text{left}, 0, X, \ \ldots \ \text{(we are stuck in a loop)} \end{aligned}$$

(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:

$$\begin{aligned} \pi_3(\text{left}|X) &= 0 & \pi_3(\text{left}|Y) &= 0.9 \\ \pi_3(\text{right}|X) &= 1 & \pi_3(\text{right}|Y) &= 0.1 \end{aligned}$$

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