Reinforcement Learning — Release: Thursday, 17 October 2024
Prof. Milica Gašić — Discussion: Wednesday, 23 October 2024

Exercise set #2

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. **Discounted returns**
   (a) Assume you observe a sequence of five rewards

   $$R_1 = -1, R_2 = 2, R_3 = 6, R_4 = 3, R_5 = 2$$

   until you reach a terminal state, i.e. a state that always transitions back to itself with a reward of 0. Calculate the returns $G_0, \ldots, G_5$ for a discount factor of $\gamma = 0.5$.

   (b) Assume an MDP produces an infinite sequence of rewards of 5, i.e.

   $$R_1 = 5, R_2 = 5, R_3 = 5, \ldots$$

   Calculate the return G_0 for the discount factors $\gamma \in \{0, 0.2, 0.5, 0.9, 0.95, 0.99, 0.999\}$. What would happen if the discount factor was $\gamma = 1$?

   Hint: You can use the closed form of a special case of the power series.

   (c) Assume you observe a sequence of $T > 1$ rewards

   $$R_1 = 0, R_2 = 0, \ldots, R_{T-1} = 0, R_T$$

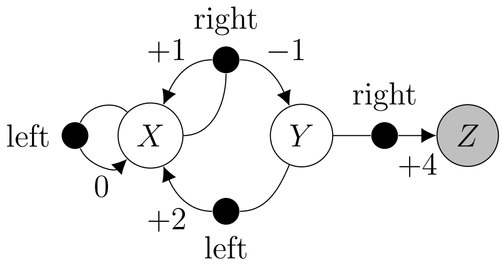
   until you reach a terminal state. Note that all rewards except R_T are zero. How can you choose γ such that the initial return G_0 is equal to $\epsilon > 0$? Calculate these γ for the following situations:

   i. $\epsilon = 0.1$, $R_T = 1$, $T = 10$
   ii. $\epsilon = 0.1$, $R_T = 1$, $T = 50$
   iii. $\epsilon = 0.01$, $R_T = 1$, $T = 50$

2. **Value functions**
   (a) For any given MDP, policy π and *terminal state* E, what is $v_\pi(E)$? All transitions from a terminal state are back to itself with a reward of 0.

Exercises by Stefan Harmeling, used with permission

(b) Consider the MDP and policy π_1 from the previous exercise set. Note that if action *right* is taken in state X, then the transitions to X and Y occur with probabilities 0.75 and 0.25, respectively. The deterministic policy π_1 is defined as $\pi_1(X) = \text{right}$, $\pi_1(Y) = \text{right}$.



Calculate the values of states X and Y under policy π_1, i.e. $v_{\pi_1}(X)$ and $v_{\pi_1}(Y)$, using a discount factor of $\gamma = 0.9$.

Hint: Start with the value of Y.

3. **Policy iteration**

   Implement policy iteration and apply it to the Maze environment from the lecture. Follow the instructions in the Jupyter notebook `policy-iteration.ipynb`.