Reinforcement Learning
Prof. Milica Gašić

Release: Thursday, 24 October 2024
Discussion: Wednesday, 30 October 2024

Exercise set #3

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. **Recursive Bellman equations**

   Prove the recursive Bellman expectation equations for the value function v_π and the action value function q_π using the state transition function $\mathcal{P}$ and the reward function $\mathcal{R}$. You are allowed to use the equations from Theorem 1 in Section 4.

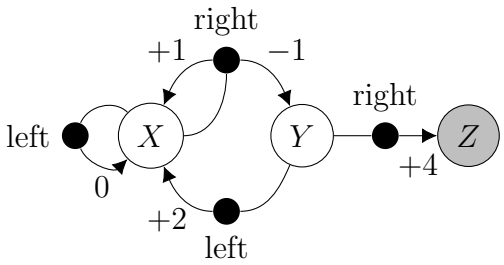
   (a) $v_\pi(s) = \sum_a \pi(a|s) \left(\mathcal{R}(s,a) + \gamma \sum_{s'} \mathcal{P}(s'|s,a) v_\pi(s') \right)$

   (b) $q_\pi(s,a) = \mathcal{R}(s,a) + \gamma \sum_{s'} \mathcal{P}(s'|s,a) \sum_{a'} \pi(a'|s') q_\pi(s',a')$

2. **Action value functions**

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

   (b) Consider the MDP and policy π_1 from the previous exercise sets. 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 $\pi_1(X) = \text{right}$, $\pi_1(Y) = \text{right}$.



   Compute the action value of state X and action *left* under policy π_1, i.e. $q_{\pi_1}(X, \text{left})$, using *only* the action value function (don't use the values from the last exercise set). The discount factor is $\gamma = 0.9$.

Exercises by Stefan Harmeling, used with permission

(c) In the lecture we defined the policy iteration algorithm to find the optimal policy using value functions. Write down a modified version of policy iteration that finds the optimal policy using action value functions (known as Q-Policy iteration).

3. **Value iteration**

(a) Perform two steps of value iteration for the MDP from exercise 2 (b), i.e. calculate $v_1(s)$ and $v_2(s)$ for $s \in \{X, Y\}$. Initialize the values with $v_0(X) = 0$ and $v_0(Y) = 0$. You can assume that the value of the terminal state Z is zero in each step.

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