Prof. Frank Hutter, Prof. Joschka Bödecker

Dr. Marius Lindauer
Gabriel Kalweit

Reinforcement Learning
Exercise 1
Submit until **Thursday, November 16 at 2:00pm**

Next week, we will start with practical exercises. This week, after understanding the general RL setting, we are going to have a look at some very basic theory. Please push your solutions as a PDF to subdirectory `exercise-01` in your assigned git-repository. We are going to submit a `feedback.txt` in that directory.

Preliminaries

This exercise is based on Lecture 2[1] from David Silver's RL course[2]. Watch before the upcoming meeting on Friday, November 10.

1 Markov Decision Processes (12p)

In this exercise, we will deal with finite horizon problems of length N. Consider a student $\mathcal{S}$ who takes a written exam. The list of tasks includes k different tasks to be solved, each will provide r_i ($i \in \{1, \ldots, k\}$, see Table 1) points. To keep the problem easy, we assume that each task can be solved only either completely wrong or completely correct (in the first case 0 points, in the latter r_i) and that the student knows with safety whether a task was solved properly or not after working on it. Hence, the student acts deterministically.
The student $\mathcal{S}$ can try to solve each task arbitrarily often. Each try will be successful with a probability $p_i^{\mathcal{S}}$ which depends on the difficulty of the task and the knowledge of the student (these probabilities are listed in Table 1).

(a) Formalize the above described problem as a Markov Decision Process. You do not have to write out the individual elements completely (i.e. name all states explicitly or provide a full transition matrix), but define the contents in unambiguous expressions. (2p)

Of course, there is a time limit in exams. We assume that the duration of the exam allows the student a total of up to $N = 5$ tries to solve the tasks (again to simplify, we assume each try takes the same time). It is also known that to pass the exam at least 40 % of the maximum score must be achieved.

(b) How would you model the risk of failing the exam in this scenario? *Hint: You can introduce terminal rewards.* (1p)

[1] `https://youtu.be/lfHX2hHRMVQ`
[2] `http://www0.cs.ucl.ac.uk/staff/d.silver/web/Teaching.html`

Task	Points	Probability of Success $p_i^{\mathcal{S}}$
u_1	4	0.1
u_2	1	0.8
u_3	3	0.3
u_4	2	0.5

Table 1: Properties of exam tasks.

In the following, we will compare several policies to solve the exam, i.e. which unsolved exercise a student chooses to solve next. In case of a failed exam, for all subsequent calculations, we assume terminal rewards of -10 for failing states.

(c) Student $\mathcal{S}$ considers two possible policies, $\pi_A^{\mathcal{S}}$ and $\pi_B^{\mathcal{S}}$, that determine in which order the tasks will be solved. The first policy follows the tasks in increasing difficulty, i.e. with decreasing possibility of success $p_i^{\mathcal{S}}$. Using policy $\pi_B^{\mathcal{S}}$, the tasks are tackled in reversed order (reversed with respect to $\pi_A^{\mathcal{S}}$). Compare both policies by determining the path costs for both policies $\pi_A^{\mathcal{S}}$ and $\pi_B^{\mathcal{S}}$. *Remember: The student knows about the success of solving a task after dealing with it and may choose to solve a task again in the case of failing.* (4p)

(d) Derive a stationary policy $\pi_C^{\mathcal{S}}$ that is more promising than $\pi_A^{\mathcal{S}}$ and $\pi_B^{\mathcal{S}}$. In what order does this strategy choose the tasks? (2p)

(e) Suggest a method that reduces the probability of failing for each of the above policies. You can describe in non-formal sentences. *Hint: Use non-stationary policies.* (1p)

(f) Student $\mathcal{T}$ has learned selectively and did not prepare the topics of tasks 2-4 ($p_i^{\mathcal{T}} = 0.0$ for $i \in \{2, 3, 4\}$). How good must student $\mathcal{T}$ be prepared for the topic of task 1 (i.e. how high has $p_1^{\mathcal{T}}$ to be at least) in order to be at least as successful as student $\mathcal{S}$, if student $\mathcal{S}$ follows policy $\pi_A^{\mathcal{S}}$? (2p)

2 Bellman Equation (8p)

In the grid world shown in Figure 1, the cells of the grid correspond to states. In each cell, four actions can be taken (north, south, east and west) which move the agent deterministically in the respective neighboring cell. Actions that would move the agent out of the grid won't lead to cell changes, but cause direct costs of -1. All other actions are free, with the exception of the states A and B. Every action performed by the agent in A moves him in A' with a reward of 10, each action in B moves him to B' with a reward of 5. Assume that the agent following policy π selects all actions with equal probability in all states. The accompanying value function v_π with a discounting factor of $\gamma = 0.9$ is given in the right part of Figure 1.

(a) Show exemplary for state $s_{3,3}$ in the middle of the grid with $v_\pi(s_{3,3}) = 0.7$ that the Bellman equation is satisfied for all neighboring states. (2p)

(b) Explain why the value of state B is higher than the direct reward. Why does this not hold for state A? (3p)

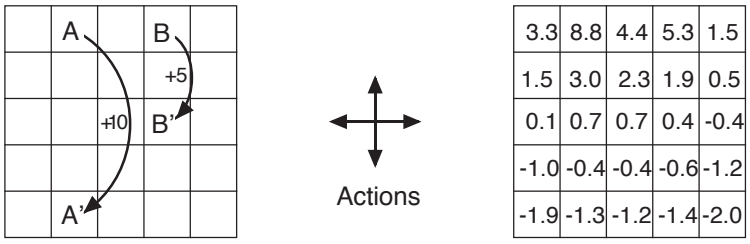

Figure 1: gridworld

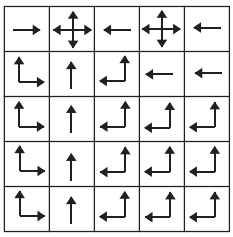

Figure 2: Optimal policy for gridworld.

The optimal policy π_* is shown in Figure 2. Now assume that we reduce the reward for jumping from A to A' to 4.

(c) How does the optimal policy change? What are the new values of states A and B? (3p)

3 Bonus: Experiences (1p)

Submit an `experiences.txt`, where you provide a brief summary of your experience with this exercise, the corresponding lecture and the last meeting. As a minimum, say how much time you invested and if you had major problems – and if yes, where.

Please push your solutions to subdirectory `exercise-01` in your assigned git-repository by **Thursday, November 16 at 2:00pm**. **Solutions after that or via email will not be accepted.**