Prof. Joschka Bödecker, Gabriel Kalweit

Reinforcement Learning
Exercise 1

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. Solve this exercise sheet using only pen and paper.

0 Lecture

Watch *Lecture 02: Planning by Dynamic Programming*[1] before the upcoming session on Friday, November 2.

1 Markov Decision Processes

In this exercise, we will deal with finite horizon problems of length N. Consider a student $\mathcal{S}$ who takes a written exam. The exam consists of 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 absolute certainty 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. *Hint: It can make sense to also include the next state in the reward function.*

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 $N = 3$ 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 50 % 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, i.e. rewards for getting in some terminal state that causes the episode to end.*

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.

[1] https://ilias.uni-freiburg.de/goto.php?target=xvid_1121346&client_id=unifreiburg

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

Table 1: Properties of exam tasks.

(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 tackles the tasks in increasing difficulty, i.e. with decreasing possibility of success $p_i^{\mathcal{S}}$. Following policy $\pi_B^{\mathcal{S}}$, the tasks are solved in reversed order (reversed with respect to $\pi_A^{\mathcal{S}}$). The student knows about the success of solving a task after dealing with it and sticks with a task until it is solved correctly. Compare both policies by determining the values for following $\pi_A^{\mathcal{S}}$ and $\pi_B^{\mathcal{S}}$ starting in the initial state of not having solved a task yet.

2 Markov Property

Imagine you want to apply the algorithms from this lecture on a real physical system. You get sensor input after each 0.01 seconds, but the execution of actions has a delay of 0.2 seconds.

(a) Is the Markov property fulfilled?

(b) Is the Markov property fulfilled if you consider the history of the last 0.2 seconds as part of the state space?

3 Bellman Equation

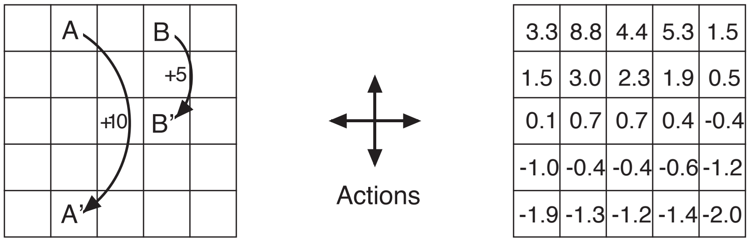
In the grid world shown in Figure 1, the cells of the grid correspond to states. In each cell, four actions can be taken (north, south, east and west) which move the agent deterministically in the respective neighboring cell. Actions that would move the agent out of the grid won't lead to cell changes, 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 it in A' with a reward of 10, each action in B moves it 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 $s_{0,0}$ with $v_\pi(s_{0,0}) = 3.3$ that the Bellman equation is satisfied.

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

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?



3.3	8.8	4.4	5.3	1.5
1.5	3.0	2.3	1.9	0.5
0.1	0.7	0.7	0.4	-0.4
-1.0	-0.4	-0.4	-0.6	-1.2
-1.9	-1.3	-1.2	-1.4	-2.0

Figure 1: gridworld

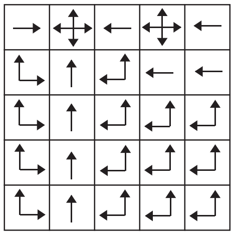


Figure 2: Optimal policy for gridworld.

4 Experiences

Make a post in thread *Week 01: Markov Decision Processes* in the forum[2], where you provide a brief summary of your experience with this exercise, the corresponding lecture and the last meeting.

[2] https://ilias.uni-freiburg.de/goto.php?target=frm_1121060&client_id=unifreiburg