Technische Universität Darmstadt
FG Kommunikationstechnik
Prof. Dr.-Ing. Anja Klein
Dr. rer. nat. Sabrina Klos and Dr.-Ing. Andrea Ortiz

Fundamentals of Reinforcement Learning

Programming Exercise 4

Task 1 - Programming: Policy Evaluation

Consider a grid world as shown in Fig. 1.

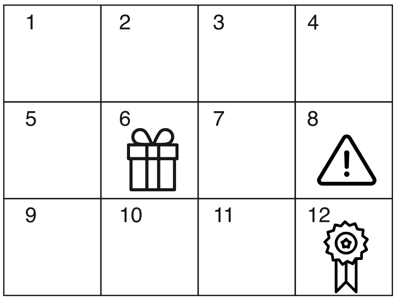

Figure 1: Grid world

The agent always starts at state 1 and its goal is to reach state 12. State 8 is a trap. The agent is equipped with four movement actions, i.e., up, down, left and right. Actions that go beyond the limits of the grid world do not change the state. For example, going up in state 1, forces the agent to stay in state 1. If the agent falls into state 8, it receives a reward of -10 points and the episode terminates. Similarly, if it reaches state 12, it receives a reward of $+10$ and the episode terminates. The environment dynamics are deterministic, except for state 6. This means that when taking actions in state 6, the agent moves to the intended direction with probability of 0.8, and to any of the two perpendicular directions with probability 0.1 each. The movement cost is -1.0 for each step and assume a discount factor $\gamma = 0.9$ is considered.

1.1 Determine the transition matrices $\mathbf{P}_a$ for each of the four possible actions $a \in \mathcal{A}$, $\mathcal{A} = \{\text{up, down, left, right}\}$.

1.2 Determine the reward vectors $\mathbf{r}_a$ for each of the four possible actions $a \in \mathcal{A}$, $\mathcal{A} = \{\text{up, down, left, right}\}$.

1.3 Using the Iterative Policy Evaluation algorithm, determine the value $v(s)$ of each state for a policy that assigns the same probability to the selection of any of the actions.

Task 2 - Programming: Policy Iteration and Value Iteration

Consider the same grid world as in Task 5.

2.1 Determine the optimal policy π_* using the policy iteration algorithm and the value iteration algorithm.

2.2 Compare the value function v_* of the optimal policy π_* with the one obtained in Task (5.3). Explain your findings in note form.

2.3 Find the optimal policy π_* for the cases when $\gamma = 0.5$, $\gamma = 0.1$ and $\gamma = 0$. What do you observe from the results?