MSc in Computer Science and Engineering

Planning, Learning and Decision Making

Homework 2. Markov decision problems

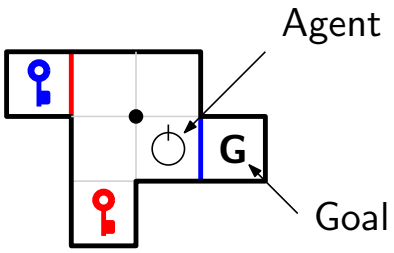

Figure 1: Grid world where an agent must reach the cell marked with "**G**".

Consider an agent moving in the grid-world environment of Fig. 1. The agent must reach the goal cell, marked with "**G**".

At each step, the agent may move in any of the four directions—up, down, left and right. Movement across a *grey* cell division succeeds with a 0.8 probability and fails with a 0.2 probability. Movements across colored cell divisions (blue or red) succeed with a 0.8 probability *only if the agent has the corresponding colored key*. Otherwise, they fail with probability 1. When the movement fails, the agent remains in the same cell.

To get a colored key, the agent simply needs to stand in the corresponding cell. In other words, as soon as the agent stands on the cell of a colored key, you may consider that it holds that key thereafter.

Exercise 1.

(a) Identify the state space, $\mathcal{X}$, and the action space, $\mathcal{A}$, for the MDP. Assume that the agent never has the blue key without the red key and never reaches the goal without both keys.

(b) Write down the transition probability matrix for the action "right" and a (possible) cost function for the MDP. Make sure that the cost function is as simple as possible and verifies $c(x, a) \in [0, 1]$ for all states $x \in \mathcal{X}$ and actions $a \in \mathcal{A}$. Note, in particular, that the cost should depend only on the agent *standing* in the goal cell.

(c) Compute the cost-to-go function associated with the policy in which the agent always goes right, using a discount $\gamma = 0.9$. You can use any software of your liking for the harder computations, but should indicate all other computations.

Solution 1:

The MDP model describes the agent's decision process. Note, in particular, that the agent's decision depends on whether the agent has (or not) the red key and whether it has (or not) the blue key.

(a) Numbering the cells from 1 to 7, starting from left to right and from top to bottom, we have

$$\mathcal{X} = \big\{(1,b,r),(2,\bar{b},\bar{r}),(2,\bar{b},r),(2,b,r),(3,\bar{b},\bar{r}),(3,\bar{b},r),(3,b,r),$$
$$(4,\bar{b},\bar{r}),(4,\bar{b},r),(4,b,r),(5,\bar{b},\bar{r}),(5,\bar{b},r),(5,b,r),(6,b,r),(7,\bar{b},r),(7,b,r)\big\},$$

where we assume that the agent can never get to the blue key without the red key and can never get to the goal without both keys. The action space comes directly $\mathcal{A} = \{u,d,l,r\}$.

(b) As for the transition probabilities, from the problem description we have, for $\boldsymbol{P}_r$,

$$\boldsymbol{P}_r = \begin{bmatrix}
0.2 & 0.0 & 0.0 & 0.8 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\
0.0 & 0.2 & 0.0 & 0.0 & 0.8 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\
0.0 & 0.0 & 0.2 & 0.0 & 0.0 & 0.8 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\
0.0 & 0.0 & 0.0 & 0.2 & 0.0 & 0.0 & 0.8 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\
0.0 & 0.0 & 0.0 & 0.0 & 1.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\
0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 1.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\
0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 1.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\
0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.2 & 0.0 & 0.0 & 0.8 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\
0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.2 & 0.0 & 0.0 & 0.8 & 0.0 & 0.0 & 0.0 & 0.0 \\
0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.2 & 0.0 & 0.0 & 0.8 & 0.0 & 0.0 & 0.0 \\
0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 1.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\
0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 1.0 & 0.0 & 0.0 & 0.0 & 0.0 \\
0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.2 & 0.8 & 0.0 & 0.0 \\
0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 1.0 & 0.0 & 0.0 \\
0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 1.0 & 0.0 \\
0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 1.0
\end{bmatrix},$$

As for the cost function, it depends only on the position of the agent, so it should be

constant across actions. Therefore, a possible cost function is:

$$\mathbf{C} = \begin{bmatrix} 1.0 & 1.0 & 1.0 & 1.0 \\ 1.0 & 1.0 & 1.0 & 1.0 \\ 1.0 & 1.0 & 1.0 & 1.0 \\ 1.0 & 1.0 & 1.0 & 1.0 \\ 1.0 & 1.0 & 1.0 & 1.0 \\ 1.0 & 1.0 & 1.0 & 1.0 \\ 1.0 & 1.0 & 1.0 & 1.0 \\ 1.0 & 1.0 & 1.0 & 1.0 \\ 1.0 & 1.0 & 1.0 & 1.0 \\ 1.0 & 1.0 & 1.0 & 1.0 \\ 1.0 & 1.0 & 1.0 & 1.0 \\ 1.0 & 1.0 & 1.0 & 1.0 \\ 1.0 & 1.0 & 1.0 & 1.0 \\ 1.0 & 1.0 & 1.0 & 1.0 \\ 0.0 & 0.0 & 0.0 & 0.0 \\ 1.0 & 1.0 & 1.0 & 1.0 \\ 1.0 & 1.0 & 1.0 & 1.0 \end{bmatrix}.$$

(c) To compute the cost to go function associated with that policy, we solve the linear system $J^\pi = \mathbf{c}_\pi + \gamma \boldsymbol{P}_\pi J^\pi$, where $\boldsymbol{P}_\pi = \boldsymbol{P}_r$ and $\mathbf{c}_\pi = \mathbf{C}_{:,r}$. The solution is given by:

$$\boldsymbol{J}^{\pi} = (\mathbf{I} - \gamma \boldsymbol{P}_{\pi})^{-1}\mathbf{c}_{\pi} = \begin{bmatrix} 100 \\ 100 \\ 100 \\ 100 \\ 100 \\ 100 \\ 100 \\ 100 \\ 100 \\ 2.29 \\ 100 \\ 100 \\ 1.22 \\ 0 \\ 100 \\ 100 \end{bmatrix}.$$