TÉCNICO LISBOA

MSc in Computer Science and Engineering

Learning and Decision Making 2016-2017

Homework 2. Markov decision problems

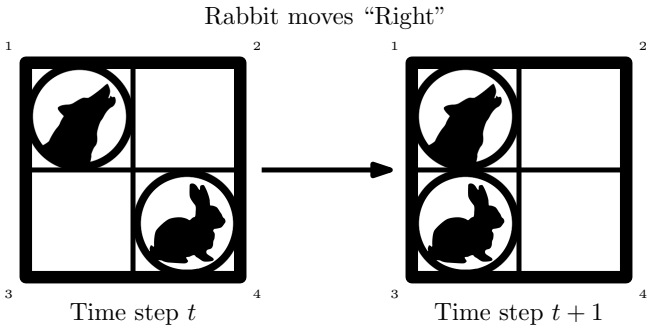

Figure 1: Environment where a predator pursues a randomly moving prey in a 2×2 toroidal world. In this situation, the prey moves right, reentering the environment on the left.

Consider a predator (wolf) pursuing a prey (hare) in a 2×2 toroidal grid. A toroidal world "wraps around", i.e., an individual exiting through any of the four sides of the grid reenters on the opposite side (see Fig. 1 for an example).

At each time step, the hare selects uniformly at random one of the four directions (up, down, left, and right) and moves to the adjacent cell in that direction with a probability 0.4. With a probability 0.6 it remains in the same cell.

The wolf, on the other hand, can select at each time step one of five actions—up (U), down (D), left (L) and right (R) or stay (S). If it selects action S, it remains in the same cell with probability 1.0. Otherwise, the other 4 actions succeed in moving the wolf to the adjacent cell in the corresponding direction with a probability 0.8 and fail with a probability 0.2.

The goal of the wolf is to catch the hare. In this homework, you will model the decision of the wolf as a Markov decision problem (MDP).

Exercise 1.

(a) Identify the state space, $\mathcal{X}$, and the action space, $\mathcal{A}$, for the MDP.

(b) Write down the transition probabilities and the 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 $a \in \mathcal{A}$.

(c) Compute the cost-to-go function associated with the policy in which the wolf always goes up, using a discount $\gamma = 0.99$. 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 wolf's decision process. Since the decision of the wolf depends on the position of the hare, the state must contain information about both animals. One possibility is to consider the relative positions of the two animals.

(a) We can represent the state as

$$\mathcal{X} = \{C, V, H, D\},$$

where C stands for "coincident", V stands for "vertically aligned", H stands for "horizontally aligned" and D stands for "diagonally positioned". The action space is still $\mathcal{A} = \{S, U, D, L, R\}$.

(b) The transition probabilities come:

$$\boldsymbol{P}_S = \begin{bmatrix} 0.60 & 0.20 & 0.20 & 0 \\ 0.20 & 0.60 & 0 & 0.20 \\ 0.20 & 0 & 0.60 & 0.20 \\ 0 & 0.20 & 0.20 & 0.60 \end{bmatrix}, \qquad \boldsymbol{P}_L = \boldsymbol{P}_R = \begin{bmatrix} 0.28 & 0.52 & 0.04 & 0.16 \\ 0.52 & 0.28 & 0.16 & 0.04 \\ 0.04 & 0.16 & 0.28 & 0.52 \\ 0.16 & 0.04 & 0.52 & 0.28 \end{bmatrix},$$

and

$$\boldsymbol{P}_U = \boldsymbol{P}_D = \begin{bmatrix} 0.28 & 0.04 & 0.52 & 0.16 \\ 0.04 & 0.28 & 0.16 & 0.52 \\ 0.52 & 0.16 & 0.28 & 0.04 \\ 0.16 & 0.52 & 0.04 & 0.28 \end{bmatrix}.$$

As for the cost function, one possibility is

$$\mathbf{C} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix},$$

which penalizes the wolf for every step that the prey is "free".

(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 $\boldsymbol{P}_\pi = \boldsymbol{P}_U$ and $\mathbf{c}_\pi = \mathbf{C}_{:,U}$. The solution is given by:

$$J^\pi = (\mathbf{I} - \gamma \boldsymbol{P}_\pi)^{-1} \mathbf{c}_\pi = \begin{bmatrix} 74.0 & 75.7 & 74.8 & 75.6 \end{bmatrix}^\top.$$

An second solution is possible, which explicitly enumerates the position of wolf and hare. Although easier to interpret, it leads to a significantly larger model. We also provide such solution below.

Solution 2:

(a) Numbering the cells from 1 to 4 as in the diagram, the state space is the set

$$\begin{aligned} \mathcal{X} = \{&(1,1),(1,2),(1,3),(1,4),(2,1),(2,2),(2,3),(2,4), \\ &(3,1),(3,2),(3,3),(3,4),(4,1),(4,2),(4,3),(4,4)\}, \end{aligned}$$

where the first component is the position of the wolf and the second is the position of the hare. The action space is $\mathcal{A} = \{S, U, D, L, R\}$.

(b) The transition probabilities come:

$$\boldsymbol{P}_S = \begin{bmatrix}
0.60 & 0.20 & 0.20 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0.20 & 0.60 & 0 & 0.20 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0.20 & 0 & 0.60 & 0.20 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0.20 & 0.20 & 0.60 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0.60 & 0.20 & 0.20 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0.20 & 0.60 & 0 & 0.20 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0.20 & 0 & 0.60 & 0.20 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0.20 & 0.20 & 0.60 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.60 & 0.20 & 0.20 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.20 & 0.60 & 0 & 0.20 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.20 & 0 & 0.60 & 0.20 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.20 & 0.20 & 0.60 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.60 & 0.20 & 0.20 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.20 & 0.60 & 0 & 0.20 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.20 & 0 & 0.60 & 0.20 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.20 & 0.20 & 0.60
\end{bmatrix}$$

$$\boldsymbol{P}_U = \boldsymbol{P}_D = \begin{bmatrix}
0.12 & 0.04 & 0.04 & 0 & 0 & 0 & 0 & 0 & 0.48 & 0.16 & 0.16 & 0 & 0 & 0 & 0 & 0 \\
0.04 & 0.12 & 0 & 0.04 & 0 & 0 & 0 & 0 & 0.16 & 0.48 & 0 & 0.16 & 0 & 0 & 0 & 0 \\
0.04 & 0 & 0.12 & 0.04 & 0 & 0 & 0 & 0 & 0.16 & 0 & 0.48 & 0.16 & 0 & 0 & 0 & 0 \\
0 & 0.04 & 0.04 & 0.12 & 0 & 0 & 0 & 0 & 0 & 0.16 & 0.16 & 0.48 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0.12 & 0.04 & 0.04 & 0 & 0 & 0 & 0 & 0 & 0.48 & 0.16 & 0.16 & 0 \\
0 & 0 & 0 & 0 & 0.04 & 0.12 & 0 & 0.04 & 0 & 0 & 0 & 0 & 0.16 & 0.48 & 0 & 0.16 \\
0 & 0 & 0 & 0 & 0.04 & 0 & 0.12 & 0.04 & 0 & 0 & 0 & 0 & 0.16 & 0 & 0.48 & 0.16 \\
0 & 0 & 0 & 0 & 0 & 0.04 & 0.04 & 0.12 & 0 & 0 & 0 & 0 & 0 & 0.16 & 0.16 & 0.48 \\
0.48 & 0.16 & 0.16 & 0 & 0 & 0 & 0 & 0 & 0.12 & 0.04 & 0.04 & 0 & 0 & 0 & 0 & 0 \\
0.16 & 0.48 & 0 & 0.16 & 0 & 0 & 0 & 0 & 0.04 & 0.12 & 0 & 0.04 & 0 & 0 & 0 & 0 \\
0.16 & 0 & 0.48 & 0.16 & 0 & 0 & 0 & 0 & 0.04 & 0 & 0.12 & 0.04 & 0 & 0 & 0 & 0 \\
0 & 0.16 & 0.16 & 0.48 & 0 & 0 & 0 & 0 & 0 & 0.04 & 0.04 & 0.12 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0.48 & 0.16 & 0.16 & 0 & 0 & 0 & 0 & 0 & 0.12 & 0.04 & 0.04 & 0 \\
0 & 0 & 0 & 0 & 0.16 & 0.48 & 0 & 0.16 & 0 & 0 & 0 & 0 & 0.04 & 0.12 & 0 & 0.04 \\
0 & 0 & 0 & 0 & 0.16 & 0 & 0.48 & 0.16 & 0 & 0 & 0 & 0 & 0.04 & 0 & 0.12 & 0.04 \\
0 & 0 & 0 & 0 & 0 & 0.16 & 0.16 & 0.48 & 0 & 0 & 0 & 0 & 0 & 0.04 & 0.04 & 0.12
\end{bmatrix}$$

$$
\boldsymbol{P}_L = \boldsymbol{P}_R = \begin{bmatrix}
0.12 & 0.04 & 0.04 & 0 & 0.48 & 0.16 & 0.16 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0.04 & 0.12 & 0 & 0.04 & 0.16 & 0.48 & 0 & 0.16 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0.04 & 0 & 0.12 & 0.04 & 0.16 & 0 & 0.48 & 0.16 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0.04 & 0.04 & 0.12 & 0 & 0.16 & 0.16 & 0.48 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0.48 & 0.16 & 0.16 & 0 & 0.12 & 0.04 & 0.04 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0.16 & 0.48 & 0 & 0.16 & 0.04 & 0.12 & 0 & 0.04 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0.16 & 0 & 0.48 & 0.16 & 0.04 & 0 & 0.12 & 0.04 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0.16 & 0.16 & 0.48 & 0 & 0.04 & 0.04 & 0.12 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.12 & 0.04 & 0.04 & 0 & 0.48 & 0.16 & 0.16 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.04 & 0.12 & 0 & 0.04 & 0.16 & 0.48 & 0 & 0.16 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.04 & 0 & 0.12 & 0.04 & 0.16 & 0 & 0.48 & 0.16 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.04 & 0.04 & 0.12 & 0 & 0.16 & 0.16 & 0.48 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.48 & 0.16 & 0.16 & 0 & 0.12 & 0.04 & 0.04 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.16 & 0.48 & 0 & 0.16 & 0.04 & 0.12 & 0 & 0.04 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.16 & 0 & 0.48 & 0.16 & 0.04 & 0 & 0.12 & 0.04 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.16 & 0.16 & 0.48 & 0 & 0.04 & 0.04 & 0.12
\end{bmatrix}.
$$

As for the cost function, one possibility is

$$
\mathbf{C} = \begin{bmatrix}
0 & 0 & 0 & 0 & 0 \\
1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 \\
0 & 0 & 0 & 0 & 0 \\
1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 \\
0 & 0 & 0 & 0 & 0 \\
1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 \\
0 & 0 & 0 & 0 & 0
\end{bmatrix},
$$

which penalizes the wolf for every step that the prey is “free”.

(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 $\boldsymbol{P}_\pi = \boldsymbol{P}_U$ and $\mathbf{c}_\pi = \mathbf{C}_{:,U}$. The solution is given by:

$$
\begin{aligned}
J^\pi &= (\mathbf{I} - \gamma \boldsymbol{P}_\pi)^{-1} \mathbf{c}_\pi \\
&= \begin{bmatrix} 74.0 & 75.7 & 74.8 & 75.6 & 75.7 & 74.0 & 75.6 & 74.8 & 74.8 & 75.6 & 74.0 & 75.7 & 75.6 & 74.8 & 75.7 & 74.0 \end{bmatrix}^\top .
\end{aligned}
$$