ADI
Planning, Learning and Decision Making

Erik Hamberg
Matej Rojec

Homework 2

Question 1

Write down the state space $\mathcal{X}$ and the action space $\mathcal{A}$ for the MDP describing the decision process of the truck driver. Consider that a new time step occurs whenever the driver takes an action at one of the seven dotted locations (Recycling plant and stops A to F).

Solution: **Action Space**

The action space for the MDP is:

$$\mathcal{A} = \{\text{Collect Garbage}, \text{Drop Garbage}, \text{Move Up}, \text{Move Down}, \text{Move Left}, \text{Move Right}\}$$

Solution: **State Space**

The available states are given by the current location of the truck and which loads it has picked up already (for example BD means that the truck has already picked up the trash at locations B and D). We will illustrate this as a grid.

$$\mathcal{X} = \begin{array}{c} \\ \text{A} \\ \text{B} \\ \text{C} \\ \text{D} \\ \text{E} \\ \text{F} \\ \text{R} \end{array} \begin{array}{c} \begin{array}{cccccccc} \text{None} & \text{B} & \text{C} & \text{D} & \text{BC} & \text{BD} & \text{CD} & \text{BCD} \end{array} \\ \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \end{array}$$

Formally if we define $\mathcal{P} = \{A, B, C, D, E, F, R\}$ (positions) and $\mathcal{L} = \{\text{None}, \text{B}, \text{C}, \text{DBC}, \text{BD}, \text{CD}, \text{BCD}\}$ (locations where the truck has already picked up the garbage), then the state space $\mathcal{X}$ is the cartesian product of $\mathcal{P}$ and $\mathcal{L}$:

$$\mathcal{X} = \mathcal{P} \times \mathcal{L}.$$

Question 2

Write down the cost function for the MDP.

Solution:

For the cost function we have the following criteria:

1. The cost should scale linearly with the time it takes to complete an action.
2. Be maximized for all actions that have no effect. These are the invalid actions:
   (a) Moving the truck in a direction where there is no stop
   (b) Attempting to pick up garbage where there is no garbage to pick up (because it has been visited or does the not have garbage).
   (c) Attempting to drop off garbage when not in state $s = (\text{R}, \text{BCD})$.
3. Be zero when $s = (R, BCD)$ & $a =$ Drop Garbage
4. Actions that are not invalid are valid.

To comply with all of these criteria we define the following cost function f:

$$f(s,a) = \begin{cases} 0 & \text{if } a = \text{Drop Garbage and } s = (\text{R}, \text{BCD}) \\ t \cdot k & \text{if } a \in \{\text{valid actions for } s\} \\ c_{\max} & \text{if } a \in \{\text{invalid actions for } s\} \end{cases}$$

where t is the time it takes to complete the valid action that is a movement or garbage collection ($t = 10$ for collecting garbage) and:

$$c_{max} > 80 \cdot k > 0$$

This holds because 80 is the maximum time needed to complete an action. If c_{max} and k comply with the above-mentioned criteria, the function will be a valid cost function. For instance

$$\begin{cases} c_{max} = 1.0 \\ k = 0.01 \end{cases}$$

For example $f((\text{R}, \text{BCD}), \text{Drop Garbage}) = 0$ by definition of f and $f((\text{R}, \text{BD}), \text{Drop Garbage}) = c_{max}$ because the drop off was not successful, $f((\text{R}, \text{BD}), \text{Move Left}) = c_{max}$ since the action is not valid, $f((\text{B}, \text{D}), \text{Collect Garbage}) = 10 \cdot k$, since the action is valid and $f((\text{B}, \text{BD}), \text{Collect Garbage}) = c_{max}$ since the action is not valid.

Question 3

Comment the following statement: *"For the MDP above, the cost-to-go function associated with the optimal policy is strictly positive."*

Solution:

The cost function is valued in the range $0 \leqslant f(s,a) \leqslant c_{max}$. Seeing as the cost-to-go function is the expected sum of costs and the cost function is non-negative with no repeating sequence of actions incurring 0 costs, we know that the sum is larger than 0. We know that no such sequence exists since the only 0 cost action is available in state (R, BCD), and that action results in the state (R, None) (since the world restarts). From the state (R, None) the next action is strictly positive hence the cost-to-go policy function associated with the optimal policy is strictly positive.