Homework 2 - Group 03

Planning, Learning and Intelligent Decision Making

Duarte Almeida
95565

Martim Santos
95638

Exercise 1.

(a) Let Π be the set of possible stops that the truck can visit and Γ set of possible unordered combinations of stops where the trash has been collected so far. We then have

$$\Pi = \{RP, A, B, C, D, E, F\} \quad \text{and} \quad \Gamma = \{\{\}, B, C, D, BC, BD, CD, BCD\}$$

We can define the state space X for the MDP using the Cartesian product between these two sets as follows:

$$\mathcal{X} = \Pi \times \Gamma = \{(RP, \{\}), (RP, B), (RP, C), (RP, D) \cdots (F, CD), (F, BCD)\}$$

For simplicity and readability, we represent the state $(s, v) \mid s \in \Pi, v \in \Gamma$ using s_v. Therefore

$$\boldsymbol{\mathcal{X}} = \{RP_{\{\}}, RP_B, RP_C, RP_D, \cdots F_{BD}, F_{CD}, F_{BCD}\}$$

The action space A for the MDP is defined as follows:

$$\boldsymbol{\mathcal{A}} = \{Collect, Drop, U(p), D(own), L(eft), R(ight)\}$$

(b) Let $\mathcal{C}_i \mid i \in \Gamma$ be a cost submatrix for all possible actions in every the states that have i as its subscript. For instance, we have:

$$\mathcal{C}_{\{\}} = \begin{array}{c} \\ RP_{\{\}} \\ A_{\{\}} \\ B_{\{\}} \\ C_{\{\}} \\ D_{\{\}} \\ E_{\{\}} \\ F_{\{\}} \end{array} \begin{array}{c} \begin{array}{cccccc} Collect & Drop & U & D & L & R \end{array} \\ \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 0.15 \\ 1 & 1 & 0.35 & 0.275 & 0.15 & 0.2 \\ 0.05 & 1 & 1 & 1 & 0.2 & 0.4 \\ 0.05 & 1 & 1 & 1 & 0.275 & 0.275 \\ 0.05 & 1 & 1 & 1 & 0.35 & 0.35 \\ 1 & 1 & 1 & 1 & 0.275 & 0.1 \\ 1 & 1 & 0.35 & 0.1 & 0.4 & 1 \end{bmatrix} \end{array}$$

Note that performing an ”invalid action” has maximum cost (1) and a successful garbage drop has no cost (0). The costs associated with the remaining valid actions are proportional to the time that the action takes to execute. For a valid action $a \in \mathcal{A}$ in state $x \in \mathcal{X}$, its cost is defined as:

$$c(x, a) = \frac{time}{max\ time} \cdot 0.4$$

where $time$ is the time in minutes that a takes to execute in x and $max\ time$ is maximum time needed to execute an action (i.e., 80 minutes), which we associated with an intermediate cost of 0.4 that is smaller than the maximum cost.

To define the remaining cost submatrices, we note that, if the agent has collected the garbage present in the locations that appear in subscript i, then collecting garbage becomes an invalid action in those locations. Moreover, if $i = BCD$, then dropping garbage at the recycling plant becomes a valid action. With this in mind, we have:

- $[\mathcal{C}_B]_{B,Collect} = 1$ and $[\mathcal{C}_B]_{x,a} = \left[\mathcal{C}_{\{\}}\right]_{x,a}$ if $(x,a) \neq (B, Collect)$
- $[\mathcal{C}_C]_{C,Collect} = 1$ and $[\mathcal{C}_C]_{x,a} = \left[\mathcal{C}_{\{\}}\right]_{x,a}$ if $(x,a) \neq (C, Collect)$
- $[\mathcal{C}_D]_{D,Collect} = 1$ and $[\mathcal{C}_D]_{x,a} = \left[\mathcal{C}_{\{\}}\right]_{x,a}$ if $(x,a) \neq (D, Collect)$
- $[\mathcal{C}_{BC}]_{C,Collect} = 1$ and $[\mathcal{C}_{BC}]_{x,a} = [\mathcal{C}_B]_{x,a}$ if $(x,a) \neq (C, Collect)$
- $[\mathcal{C}_{BD}]_{D,Collect} = 1$ and $[\mathcal{C}_{BD}]_{x,a} = [\mathcal{C}_B]_{x,a}$ if $(x,a) \neq (D, Collect)$
- $[\mathcal{C}_{CD}]_{D,Collect} = 1$ and $[\mathcal{C}_{CD}]_{x,a} = [\mathcal{C}_C]_{x,a}$ if $(x,a) \neq (D, Collect)$
- $[\mathcal{C}_{BCD}]_{D,Collect} = 1$, $[\mathcal{C}_{BCD}]_{RP,Drop} = 0$ and $[\mathcal{C}_{BCD}]_{x,a} = [\mathcal{C}_{BC}]_{x,a}$

  if $(x,a) \notin \{(D, Collect)\ ,\ (RP, Drop)\}$

The cost function for the MDP can thus be defined through a matrix $\boldsymbol{\mathcal{C}}$ consisting in the vertical concatenation of the aforementioned cost submatrices:

$$\boldsymbol{\mathcal{C}} = \begin{bmatrix} \mathcal{C}_{\{\}} & \mathcal{C}_B & \mathcal{C}_C & \mathcal{C}_D & \mathcal{C}_{BC} & \mathcal{C}_{BD} & \mathcal{C}_{CD} & \mathcal{C}_{BCD} \end{bmatrix}^\top$$

where $\mathcal{C}_i$ contains the rows named RP_i, A_i, B_i, C_i, D_i, E_i and F_i.

(c) For a given discount factor $\gamma \in (0,1)$ and a policy π, the cost-to-go function is defined as:

$$J^\pi(x) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t \mathrm{c}_t \mid \mathrm{x}_0 = x\right]$$

Given that the cost function in this MDP satisfies:

$$c(x,a) \geq 0\ ,\ \forall x \in \mathcal{X}\ ,\ \forall a \in \mathcal{A}$$

we have that $J^\pi(x) \geq 0$, for any policy $\pi \in \Pi^{\mathrm{HR}}$ and any state $x \in \mathcal{X}$. Thus, to prove or disprove the claim that $J^*(x) > 0$, $\forall x \in \mathcal{X}$, it suffices to assess if $J^*(x^*) = 0$ for some state $x^* \in \mathcal{X}$, where J^* is the cost-to-go function associated with the optimal policy.
By inspection of the *Bellman optimality equation*:

$$J^*(x) = \min_{a \in \mathcal{A}}\left[c(x,a) + \gamma \sum_{y \in \mathcal{X}} \boldsymbol{P}(y \mid x,a) J^*(y)\right]$$

we conclude that in order for $J^*(x^*) = 0$ to be true for some state x^* it must be the case that $c(x^*, a^*) = 0$ for some action $a^* \in \mathcal{A}$, given that the cost function and the optimal cost-to-go function terms are all non-negative. Hence, x^* surely cannot be any one of the states in $\mathcal{X} \setminus \{RP_{BCD}\}$, since all actions that can be performed in those states are associated with a positive cost. Thus, the only possible state that fulfills the aforementioned condition is RP_{BCD} (rest of proof in the following page).

Now, we have:

$$\begin{aligned}
J^*(RP_{BCD}) &= \min_{a\in\mathcal{A}} \left[c(RP_{BCD}, a) + \gamma \sum_{y\in\mathcal{X}} \boldsymbol{P}(y \mid RP_{BCD}, a) J^*(y) \right] \\
&= \min \left(\min_{a\in\{Collect, U, D, L\}} \left[c(RP_{BCD}, a) + \gamma \sum_{y\in\mathcal{X}} \boldsymbol{P}(y \mid RP_{BCD}, a) J^*(y) \right], \right. \\
&\qquad\qquad c(RP_{BCD}, R) + \gamma \sum_{y\in\mathcal{X}} \boldsymbol{P}(y \mid RP_{BCD}, R) J^*(y)\ , \\
&\qquad\qquad \left. c(RP_{BCD}, Drop) + \gamma \sum_{y\in\mathcal{X}} \boldsymbol{P}(y \mid RP_{BCD}, Drop) J^*(y) \right) \\
&= \min \left(1 + \gamma J^*(RP_{BCD}),\ c(RP_{BCD}, A) + \gamma J^*(A)\ ,\ \gamma J^*(RP_{\{\}})\right) > 0
\end{aligned}$$

Therefore, we have that the statement is **true**, as every state x verifies $J^*(x) > 0$ following the reasoning made above.