Assignment5: Markov Decision Procedures

Dmitrii, Maksimov
dmitrii.maksimov@fau.de
ko65beyp

Ilia, Dudnik
ilia.dudnik@fau.de
ex69ahum

Aleksandr, Korneev
aleksandr.korneev@fau.de
uw44ylyz

June 26, 2022

Exercise 5.2 (MDP Example)

The world is 101 fields wide. In the *Start* state an agent has two possible actions, Up and *Down*. It cannot return to *Start* though and the cannot pass gray fields, so after the first move the only possible action is *Right*.

1. Model this world as a Markov Decision Process, i.e., give the components S, s_0, A, P, and R.

   $S = \{s_{10}\} + \{s_{ij}, \forall i, j \,:\, i \in \{0, 2\}, j \in \{0, 100\}\}$,
   $s_0 = s_{10}$,
   $A(s) = \begin{cases} \{Up, Down\} & s = s_{10}, \\ \{Right\} & \text{otherwise} \end{cases}$
   $P(s'|s, a) = 1$,
   $R(s) = \begin{cases} R_{s_{10}} & s = s_{10}, \\ 50 & s = s_{00}, \\ -50 & s = s_{20}, \\ -1 & s \in \{s_{01}, s_{02}, \dots, s_{0100}\}, \\ 1 & s \in \{s_{21}, s_{22}, \dots, s_{2100}\}, \end{cases}$

2. For what discount factor γ should the agent choose Up and for which *Down*? Compute the utility of each action as a function of γ.

   We have 2 possible sequences of states: $Path_1 = [s_{10}, s_{00}, s_{01}, \dots, s_{0100}]$ and $Path_2 = [s_{10}, s_{20}, s_{21}, \dots, s_{2100}]$. Since $U(Path) = \sum_{t=0}^{len(Path)-1} \gamma^t R(s_t)$:

   - $U(Path_1)$

     $U(Path_1) = R(s_{10}) + \sum_{t=1}^{101} \gamma^t R(s_{0(t-1)}) = R(s_{10}) + \gamma \cdot 50 - \sum_{t=2}^{101} \gamma^t =$
     $R(s_{10}) + \gamma \cdot 50 - \dfrac{\gamma^2(1 - \gamma^{100})}{1 - \gamma}$

- $U(Path_2)$

  $U(Path_2) = R(s_{10}) + \sum_{t=1}^{101} \gamma^t R(s_{2(t-1)}) = R(s_{10}) - \gamma \cdot 50 + \sum_{t=2}^{101} \gamma^t =$
  $R(s_{10}) - \gamma \cdot 50 + \dfrac{\gamma^2(1-\gamma^{100})}{1-\gamma}$

Now, let's find for what discount factor γ should the agent choose Up and for which *Down*:

(a) Up

$$U(Path_1) > U(Path_2) \,:\, R(s_{10}) + \gamma \cdot 50 - \frac{\gamma^2(1-\gamma^{100})}{1-\gamma} > R(s_{10}) - \gamma \cdot 50 + \frac{\gamma^2(1-\gamma^{100})}{1-\gamma} \Rightarrow$$
$$50 > \frac{\gamma(1-\gamma^{100})}{1-\gamma} \Rightarrow \gamma \leq 0.984 \Rightarrow$$
$$U(Path_1) = R(s_{10}) + 1.162, U(Path_2) = R(s_{10}) - 1.162$$

(b) *Down*

$$U(Path_1) < U(Path_2) \,:\, R(s_{10}) + \gamma \cdot 50 - \frac{\gamma^2(1-\gamma^{100})}{1-\gamma} < R(s_{10}) - \gamma \cdot 50 + \frac{\gamma^2(1-\gamma^{100})}{1-\gamma} \Rightarrow$$
$$50 < \frac{\gamma(1-\gamma^{100})}{1-\gamma} \Rightarrow \gamma \geq 0.985 \Rightarrow$$
$$U(Path_1) = R(s_{10}) - 0.745, U(Path_2) = R(s_{10}) + 0.745$$

3. What is the optimal policy if the upper path is better?

   The optimal policy($\pi_s^\star$): $\pi_s^\star = \arg\max_\pi U^\pi(s)$, where $U^\pi(s) = EU$. Since $P(s'|s,a) = 1$, $\pi_s^\star = Up$