Prof. Dr. Leif Döring
Sara Klein, Benedikt Wille

Reinforcement Learning

5. Exercise Sheet

1. Bellman expecation operator.

Recall the Bellman expectation operator for a stationary policy $\pi \in \Pi_s$:

$$\begin{aligned}(T^\pi u)(s) &= \sum_{a\in\mathcal{A}} r(s,a)\pi(a\,;\,s) + \gamma \sum_{s'\in\mathcal{S}} \mathbb{P}^\pi(S_1 = s' | S_0 = s, A_0 = a) u(s') \\ &= \sum_{a\in\mathcal{A}} \pi(a\,;\,s)\Big(r(s,a) + \gamma \sum_{s'\in\mathcal{S}} p(s';\, s,a) u(s')\Big)\end{aligned}$$

Show that we can rewrite the fixed point equation in vector notation, i.e. check that indeed $T^\pi V = V$ is equivalent to $r_\pi + \gamma P_\pi V = V$, where

$$\begin{aligned}P_\pi &= \Big(\sum_{a\in\mathcal{A}} \pi(a\,;\,s) p(s'\,;\,s,a)\Big)_{(s,s')\in\mathcal{S}\times\mathcal{S}} \\ r_\pi &= \Big(\sum_{a\in\mathcal{A}} \pi(a\,;\,s) r(s,a)\Big)_{s\in\mathcal{S}}.\end{aligned}$$

2. Markov Decision Process 1

A rental car agency serves two cities with one office in each. The agency has M cars in total. At the start of each day, the manager must decide how many cars to move from one office to the other to balance stock. Let $f_i(q)$ for $i = 1, 2$ denote the probability that the daily demand for cars to be picked up at city i and returned to city i equals q. Let $g_i(q)$, $i = 1, 2$, denote the probability that the daily demand for “one-way” rentals from city i to the other equals q. Assume that all rentals are for one day only, that it takes one day to move cars from city to city to balance stock, and, if demand exceeds availability, customers go elsewhere. The economic parameters include a cost of K per car for moving a car from city to city to balance stock, a revenue of R per car for rentals returned to the rental location and R per car for one-way rentals. Formulate this problem as discounted infinite-horizon Markov decision model.

3. Markov Decision Process 2

Two identical machines are used in a manufacturing process. From time to time, these machines require maintenance which takes three weeks. Maintenance on one machine costs c, per week while maintenance on both machines costs c_2 per week. Assume $c_2 > 2c$. The probability that a machine breaks down if it has been i periods since its last maintenance is p_i,with p_i non-decreasing in i. Maintenance begins at the start of the week immediately following breakdown, but preventive maintenance may be started at the beginning of any week. The decision maker must choose when to carry out preventive maintenance, if ever, and, if so, how many machines to repair. Assume the two machines fail independently.

a) What is the significance of the condition $c_2 > 2c$?

b) Formulate this problem as discounted infinite-horizon Markov decision model.

c) Provide an educated guess about the form of the optimal policy when the decision maker's objective is to minimize expected operating cost.

4. Banach fixed-point Theorem

Prove Banach fixed-point theorem (Theorem 3.1.17 in the lecture notes).