Ma 661 Dynamic Programming and Reinforcement Learning
Darinka Dentcheva darinka.dentcheva@stevens.edu

Homework 6
Solutions

Problem 1. Prove Lemma 2.4 (ii) of the note on "Infinite-Horizon Average Cost Problem", that is, show that $HP^* = 0$.

Recall that H is defined as

$$H = (I - P + P^*)^{-1}(I - P^*).$$

$$\begin{aligned} HP^* &= (I - P + P^*)^{-1}(I - P^*)P^* = (I - P + P^*)^{-1}(P^* - P^*P^*) \\ &= (I - P + P^*)^{-1}(P^* - P^*) = 0. \end{aligned}$$

We have use that $P^*P^* = P^*$ according to Lemma 2.1 (i).

Problem 2. In Lemma 8.1 of the note on "Infinite-Horizon Average Cost Problem", show that the equation $(I - \alpha P)^{-1} = I + \sum_{k=1}^{\infty} \alpha^k P^k$ holds.

We denote $A = I + \sum_{k=1}^{\infty} \alpha^k P^k$. We shall show that $(I - \alpha P)A = I$, which would entail that A is the inverse matrix of $(I - \alpha P)$, that is $A = (I - \alpha P)^{-1}$. To this end, we calculate

$$\begin{aligned} (I - \alpha P)A &= (I - \alpha P)\Big(I + \sum_{k=1}^{\infty} \alpha^k P^k\Big) \\ &= I + \sum_{k=1}^{\infty} \alpha^k P^k - \alpha P - \sum_{k=1}^{\infty} \alpha^{k+1} P^{k+1} \\ &= I + \sum_{k=1}^{\infty} \alpha^k P^k - \sum_{\kappa=1}^{\infty} \alpha^{\kappa} P^{\kappa} = I. \end{aligned}$$

Problem 3. Each quarter the marketing manager of a retail store divides the customers into two groups based on their purchase behavior in the previous quarter. The classes are denoted by L and H. The manager wishes to determine to which group of customers he should sent a catalog. The cost of sending a catalog is \$ 15 per customer. If a customer from group L receives a catalog, then the expected purchase in the current quarter is \$ 20, otherwise it is \$ 10. If a customer from group H receives a catalog, then the expected purchase in the current quarter is \$ 50, otherwise it is \$ 25. Furthermore, if a customer from group L receives a catalog, then the probability that he will stay in group L for the next quarter is 0.3, otherwise, it is 0.5. If a customer from group H receives a catalog, then he probability that he will stay in group H for the next quarter is 0.8, otherwise, it is 0.4.

(a) Formulate an average reward problem to help the manager.

- State: type of customer $\mathcal{X} = \{L, H\}$.
- Control set: $\mathcal{U} = \{0, 1\}$ do not send catalog $u = 0$ or send catalog $u = 1$.
- Feasible control mapping: $U(x) = \{0, 1\}$, for both states $x = L, H$.
- Transition kernel:

$$P(1) = \begin{pmatrix} 0.3 & 0.7 \\ 0.2 & 0.8 \end{pmatrix} \quad P(0) = \begin{pmatrix} 0.5 & 0.5 \\ 0.6 & 0.4 \end{pmatrix}$$

- Reward function: setting $r_L = 10$, $r_H = 25$, we have

$$c(x, u) = r_x(u + 1) - 15u, \quad x = L, H.$$

- Dynamic Programming Equation:

$$\gamma + h_x = \max_{u=0,1} \left\{ c(x, u) + p(L|x, u)h_L + p(H|x, u)h_H \right\}, \quad x = L, H.$$

(b) Determine an optimal policy using policy iteration method.

The optimal policy is to send a catalog to everybody. The average cost is 28.33 for the bias (0,33.33).

(c) Solve the problem using linear programming.

The problem formulation is

$$\begin{aligned} &\min \ \gamma \\ &\text{s.t. } \ \gamma \geq c(L, u) + p(L|x, u)h_L + p(H|x, u)h_H - h_L, \quad u = 0, 1; \\ &\qquad \ \gamma \geq c(H, u) + p(L|x, u)h_L + p(H|x, u)h_H - h_H, \quad u = 0, 1. \end{aligned}$$

(d) For what discount factor is the discounted infinite-horizon problem equivalent to the average reward problem in this context?

Any discount factor above 0.53 leads to the same optimal policy.