EN.520.637: Foundations of Reinforcement Learning Homework 2

Ha Manh Bui (CS Department)
hbui13@jhu.edu

Fall 2023

1 Problem 1

A driver is looking for inexpensive parking on the way to his destination. The parking area contains N spaces, and a garage at the end. The driver starts at space 0 and traverses the parking spaces sequentially, i.e., from space k he goes next to space $k+1$, etc. Each parking space k costs $c(k)$ and is free with probability $p(k)$ independently of whether other parking spaces are free or not. If the driver reaches the last parking spaces $N-1$ and does not park there, he must park at the garage, which costs C. The driver can observe whether a parking space is free only when he reaches it, and then, if it is free, he makes a decision to park in that space or not to park and check the next space. The problem is to find the minimum expected cost parking policy.

(a) Now formulate this problem as a DP problem with N stages (Fixed-Horizon DP). What is the set of all possible states $\mathcal{S}_k$ at stage k? What are the set of all possible actions $\mathcal{A}_k(s_k)$ at stage k given a particular state?

This problem can be formulated as a Dynamic Programming with Fixed-Horizon with N stages corresponding to N parking spaces and artificial terminal state T corresponding to having parked.

Let S_k be the state random variable at stage k, $S_k = F$ if the parking space k is free, and $S_k = \bar{F}$ otherwise. Let A_k be the action random variable at stage k, $A_k = P$ if the driver decides to park, and $A_k = \bar{P}$ otherwise. Since $k \in [0, N-1]$, the set of all possible states at stage k is

$$\mathcal{S}_k = \{F, \bar{F}\}.$$

Because the driver can only decide to park if the parking space k is free, the set of corresponding actions are

$$\mathcal{A}_k(s_k = F) = \{P, \bar{P}\} \text{ and } \mathcal{A}_k(s_k = \bar{F}) = \{\bar{P}\}.$$

(b) Under your DP problem, specify the state-reward distribution

$$p_k(s', r|s, a) = \mathbb{P}(S_{k+1} = s', R_{k+1} = r|S_k = s, A_k = a)$$

(Here, you can either specify the reward as the negative cost and maximize the expected return, or specify the reward as the cost and minimize the expected return)

Let R_{k+1} be the reward as the cost of action A_k at stage k. If $k = N-1$ and the driver does not park there, he must park at the final garage, so

$$\mathbb{P}(S_{k+1} = T, R_{k+1} = C|S_k = \bar{F}, A_k = \bar{P}) = \mathbb{P}(S_{k+1} = T, R_{k+1} = C|S_k = F, A_k = \bar{P}) = 1.$$

If $k \leq N-1$ and the driver decides to park, then

$$\mathbb{P}(S_{k+1} = T, R_{k+1} = c(k)|S_k = F, A_k = P) = 1.$$

If $k < N-1$ and the driver does not park there, then he must check the next space, so

$$\begin{cases} \mathbb{P}(S_{k+1} = F, R_{k+1} = 0|S_k = \bar{F}, A_k = \bar{P}) = p(k+1) \\ \mathbb{P}(S_{k+1} = \bar{F}, R_{k+1} = 0|S_k = \bar{F}, A_k = \bar{P}) = 1 - p(k+1) \\ \mathbb{P}(S_{k+1} = F, R_{k+1} = 0|S_k = F, A_k = \bar{P}) = p(k+1) \\ \mathbb{P}(S_{k+1} = \bar{F}, R_{k+1} = 0|S_k = F, A_k = \bar{P}) = 1 - p(k+1). \end{cases}$$

(c) Write down the DP Algorithm in terms of $N, C, c(k), p(k)$.

Let the optimal value function at stage k be $v_k^*(s)$. By definition of the DP Algorithm, we have

$$v_k^*(s) = \min_{a\in\{P,\bar{P}\}} \left\{ r_{k+1}(s,a) + \sum_{s'\in\{F,\bar{F}\}} p(s'|s,a) v_{k+1}^*(s') \right\}. \tag{1}$$

Consider at stage $k \in [0, N-1]$, s_k can be free or not, combining with Eq. 1, we get

$$v_k^*(s=F) = \begin{cases} \min\{c(k), C\} & \text{if } k = N-1 \\ \min\{c(k), p(k+1)v_{k+1}^*(s'=F) + [1-p(k+1)]v_{k+1}^*(s'=\bar{F})\} & \text{if } k < N-1, \end{cases}$$

and

$$v_k^*(s=\bar{F}) = \begin{cases} C & \text{if } k = N-1 \\ p(k+1)v_{k+1}^*(s'=F) + [1-p(k+1)]v_{k+1}^*(s'=\bar{F}) & \text{if } k < N-1. \end{cases}$$

Let $V_k^* := p(k)v_k^*(s=F) + [1-p(k)]v_k^*(s=\bar{F})$, then we have the DP Algorithm is

Algorithm 1 the DP Algorithm in terms of $N, C, c(k), p(k)$

Initialize: $V_{N-1}^* = p(N-1)\min\{c(N-1), C\} + [1-p(N-1)]C$;
for $k = (N-2) \to 0$ **do**
  $V_k^* = p(k)\min\{c(k), V_{k+1}^*\} + [1-p(k)]V_{k+1}^*$;
end for
Return: the optimal policy $\pi^* = \arg\min_{\pi\in\Pi} V^*$;

(d) Suppose $c(t)$ is monotonically decreasing as t increases. Show that the optimal policy is a threshold policy: the driver should first travel to some space (without parking) and then to park at the first available space.

From Algo. 1, we have

$$\begin{aligned} V_k^* &= p(k)\min\{c(k), V_{k+1}^*\} + [1-p(k)]V_{k+1}^* \\ &= p(k)\left[\min\{c(k), V_{k+1}^*\} - V_{k+1}^*\right] + V_{k+1}^*. \end{aligned}$$

This yields if $c(k) \geq V_{k+1}^*$, then $V_k^* = V_{k+1}^*$. Combining with the fact that $c(t)$ is monotonically decreasing as t increases, i.e., $c(k) \geq c(k+1), \forall k \in [0, N-1]$, we achieves

$$V_k^* = V_{k+1}^*, \forall k \in [0, N-1],$$

i.e., V_k^* is non-decreasing in k. Therefore, the optimal policy is a threshold policy: the driver should first travel to some space (without parking) and then to park at the first available space.

2 Problem 2

Suppose we have a machine that is either running or is broken down. If it runs throughout one week, it makes a gross profit of \$120. If it fails during the week, gross profit is zero. If it is running at the start of the week and we perform preventive maintenance, the probability that it will fail during the week is 0.4. If we do not perform such maintenance, the probability of failure is 0.7. However, maintenance will cost \$20. When the machine is broken down at the start of the week (it failed during the week before), it may either be repaired at a cost of \$60, in which case it will fail during the week with a probability of 0.4, or it may be replaced at a cost of \$120 by a new machine that is guaranteed to run through its first week of operation.

(a) Find the optimal repair, replacement, and maintenance policy that maximizes total profit over three weeks, assuming that:

- at the start of the first week, we have a working machine with failure probability 0.4 if maintenance is performed and failure probability 0.7 otherwise.

- We do not care if the machine is broken or running at the end of the horizon.

We can formalize this problem as a DP with Fixed-Horizon $T = N = 3$ stages corresponding to three weeks. Then $\forall k \in [0, N-1]$, let

- $S_k \in \mathcal{S}$ be the state random variable at stage k, $S_k = G$ if the machine is in a good condition, and $S_k = \bar{G}$ otherwise,
- $A_k \in \mathcal{A}$ be the action random variable at stage k, then we have $\mathcal{A} = \{R, \bar{R}, M, \bar{M}\}$, where R represents for replace, $\bar{R}$ represents for repair, $M, \bar{M}$ represent for maintain or not.
- R_{k+1} be the reward of action A_k at stage k.

Then, we have the expected reward at stage k is

$$r_{k+1}(s, a) = \mathbb{E}[R_{k+1}|S_k = s, A_k = a], \tag{2}$$

this yields $\forall k \in [0, N-1]$, we have

$$\begin{cases} r_{k+1}(s = G, a = M) = (120 - 20) * (1 - 0.4) + (0 - 20) * 0.4 = 52 \\ r_{k+1}(s = G, a = \bar{M}) = 120 * (1 - 0.7) + 0 * 0.7 = 36 \\ r_{k+1}(s = \bar{G}, a = R) = (120 - 120) * 1 + 120 * 0 = 0 \\ r_{k+1}(s = \bar{G}, a = \bar{R}) = (120 - 60) * (1 - 0.4) + (0 - 60) * 0.4 = 12. \end{cases}$$

Let the optimal value function at stage k be $v_k^*(s)$. By definition of the DP Algorithm, we have

$$v_k^*(s) = \max_{a \in \{R, \bar{R}, M, \bar{M}\}} \left\{ r_{k+1}(s, a) + \sum_{s' \in \{G, \bar{G}\}} p(s'|s, a) v_{k+1}^*(s') \right\}. \tag{3}$$

Consider at stage 2, we have

$$\begin{aligned} v_2^*(s = G) &= \max_{a \in \{M, \bar{M}\}} \left\{ r_3(s = G, a) + \sum_{s' \in \{G, \bar{G}\}} p(s'|s = G, a) v_3^*(s') \right\} \\ &= \max\{r_3(s = G, a = M) + p(s' = R|s = G, a = M) v_3^*(s' = R) + p(s' = \bar{R}|s = G, a = M) v_3^*(s' = \bar{R}), \\ &\quad r_3(s = G, a = \bar{M}) + p(s' = R|s = G, a = \bar{M}) v_3^*(s' = R) + p(s' = \bar{R}|s = G, a = \bar{M}) v_3^*(s' = \bar{R})\} \\ &= \max\{52, 36\} = 52 \Rightarrow a_2^*(s = G) = M, \end{aligned}$$

and

$$\begin{aligned} v_2^*(s = \bar{G}) &= \max_{a \in \{R, \bar{R}\}} \left\{ r_3(s = \bar{G}, a) + \sum_{s' \in \{G, \bar{G}\}} p(s'|s = \bar{G}, a) v_3^*(s') \right\} \\ &= \max\{r_3(s = \bar{G}, a = R) + p(s' = R|s = \bar{G}, a = R) v_3^*(s' = R) + p(s' = \bar{R}|s = \bar{G}, a = R) v_3^*(s' = \bar{R}), \\ &\quad r_3(s = \bar{G}, a = \bar{R}) + p(s' = R|s = \bar{G}, a = \bar{R}) v_3^*(s' = R) + p(s' = \bar{R}|s = \bar{G}, a = \bar{R}) v_3^*(s' = \bar{R})\} \\ &= \max\{0, 12\} = 12 \Rightarrow a_2^*(s = \bar{G}) = \bar{R}. \end{aligned}$$

Consider at stage 1, we have

$$\begin{aligned} v_1^*(s = G) &= \max_{a \in \{M, \bar{M}\}} \left\{ r_2(s = G, a) + \sum_{s' \in \{G, \bar{G}\}} p(s'|s = G, a) v_2^*(s') \right\} \\ &= \max\{r_2(s = G, a = M) + p(s' = R|s = G, a = M) v_2^*(s' = R) + p(s' = \bar{R}|s = G, a = M) v_2^*(s' = \bar{R}), \\ &\quad r_2(s = G, a = \bar{M}) + p(s' = R|s = G, a = \bar{M}) v_2^*(s' = R) + p(s' = \bar{R}|s = G, a = \bar{M}) v_2^*(s' = \bar{R})\} \\ &= \max\{52 + (1 - 0.4) * 52 + 0.4 * 12, 36 + (1 - 0.7) * 52 + 0.7 * 12\} = \max\{88, 60\} = 88 \Rightarrow a_1^*(s = G) = M, \end{aligned}$$

and

$$v_1^*(s=\bar{G}) = \max_{a\in\{R,\bar{R}\}}\left\{r_2(s=\bar{G},a) + \sum_{s'\in\{G,\bar{G}\}} p(s'|s=\bar{G},a)v_2^*(s')\right\}$$
$$= \max\{r_2(s=\bar{G},a=R) + p(s'=R|s=\bar{G},a=R)v_2^*(s'=R) + p(s'=\bar{R}|s=\bar{G},a=R)v_2^*(s'=\bar{R}),$$
$$r_2(s=\bar{G},a=\bar{R}) + p(s'=R|s=\bar{G},a=\bar{R})v_2^*(s'=R) + p(s'=\bar{R}|s=\bar{G},a=\bar{R})v_2^*(s'=\bar{R})\}$$
$$= \max\{0+1*52+0*12, 12+(1-0.4)*52+0.4*12\} = \max\{52,48\} = 52 \Rightarrow a_1^*(s=\bar{G}) = R.$$

Consider at stage 0, we have

$$v_0^*(s=G) = \max_{a\in\{M,\bar{M}\}}\left\{r_1(s=G,a) + \sum_{s'\in\{G,\bar{G}\}} p(s'|s=G,a)v_1^*(s')\right\}$$
$$= \max\{r_1(s=G,a=M) + p(s'=R|s=G,a=M)v_1^*(s'=R) + p(s'=\bar{R}|s=G,a=M)v_1^*(s'=\bar{R}),$$
$$r_1(s=G,a=\bar{M}) + p(s'=R|s=G,a=\bar{M})v_1^*(s'=R) + p(s'=\bar{R}|s=G,a=\bar{M})v_1^*(s'=\bar{R})\}$$
$$= \max\{52+(1-0.4)*88+0.4*52, 36+(1-0.7)*88+0.7*52\}$$
$$= \max\{125.6, 98.8\} = 125.6 \Rightarrow a_0^*(s=G) = M.$$

As a result, we obtain the possible optimal sequence of action $a_0^*(s=G)=M$, $a_1^*(s=G)=M$, $a_1^*(s=\bar{G})=R$, $a_2^*(s=\bar{G})=\bar{R}$, $a_2^*(s=G)=M$, i.e., the optimal repair, replacement, and maintenance policy that maximizes total profit over three weeks as follows

1. In the first week, we will maintain the machine.
2. In the second week, if the machine still running, we will maintain it. Otherwise, we will replace the new one.
3. In the third week, if the machine still running, we will maintain it. Otherwise, we will repair it.

(b) Now, assume that there is a penalty of \$100 if the machine is broken at the end of the third week. Re-compute the optimal policy for the last stage of the problem. Does your answer change with respect to (a)?
If there is a penalty of \$100 if the machine is broken at the end of the third week, then consider the last stage 2, we have

$$v_2^*(s=G) = \max_{a\in\{M,\bar{M}\}}\left\{r_3(s=G,a) + \sum_{s'\in\{G,\bar{G}\}} p(s'|s=G,a)v_3^*(s')\right\}$$
$$= \max\{r_3(s=G,a=M) + p(s'=R|s=G,a=M)v_3^*(s'=R) + p(s'=\bar{R}|s=G,a=M)v_3^*(s'=\bar{R}),$$
$$r_3(s=G,a=\bar{M}) + p(s'=R|s=G,a=\bar{M})v_3^*(s'=R) + p(s'=\bar{R}|s=G,a=\bar{M})v_3^*(s'=\bar{R})\}$$
$$= \max\{52+(1-0.4)*0+0.4*(-100), 36+(1-0.7)*0+0.7*(-100)\}$$
$$= \max\{12,-34\} = 12 \Rightarrow a_2^*(s=G) = M,$$

and

$$v_2^*(s=\bar{G}) = \max_{a\in\{R,\bar{R}\}}\left\{r_3(s=\bar{G},a) + \sum_{s'\in\{G,\bar{G}\}} p(s'|s=\bar{G},a)v_3^*(s')\right\}$$
$$= \max\{r_3(s=\bar{G},a=R) + p(s'=R|s=\bar{G},a=R)v_3^*(s'=R) + p(s'=\bar{R}|s=\bar{G},a=R)v_3^*(s'=\bar{R}),$$
$$r_3(s=\bar{G},a=\bar{R}) + p(s'=R|s=\bar{G},a=\bar{R})v_3^*(s'=R) + p(s'=\bar{R}|s=\bar{G},a=\bar{R})v_3^*(s'=\bar{R})\}$$
$$= \max\{0+1*0+0*(-100), 12+(1-0.4)*0+0.4*(-100)\}$$
$$= \max\{0,-28\} = 0 \Rightarrow a_2^*(s=\bar{G}) = R.$$

Therefore, we obtain the optimal policy for the last stage is $a_2^*(s=G)=M$ and $a_2^*(s=\bar{G})=R$, i.e., in the third week, if the machine still running, we will maintain it. Otherwise, we will replace it. As a result, my answer will change w.r.t. (a) in the optimal action of the last stage.

3 Problem 3

Consider the system

$$x_{k+1} = x_k + u_k, k = 0, 1, 2, 3$$

with initial state $x_0 = 1$, and the return function

$$\sum_{k=0}^{3} x_k^2 - 2u_k^2.$$

Given the control constraint

$$u_k \in \{u : 0 \leq x_k + u \leq 5, u \text{ is integer}\},$$

apply the DP algorithm to find the optimal control sequence u_0^*, u_1^*, u_2^* that maximizes the return function. We can formalize this problem as a DP with Fixed-Horizon $T = N = 4$ stages. Since the return function is $\sum_{k=0}^{3} x_k^2 - 2u_k^2$, we obtain the reward at stage k is

$$R_{k+1} := x_k^2 - 2u_k^2, \forall k \in \{0, 1, 2, 3\}.$$

In addition, since $x_{k+1} = x_k + u_k, u_k \in \{u : 0 \leq x_k + u \leq 5, u \text{ is integer}\}$, we obtain

$$x_{k+1} \in \{0, 1, 2, 3, 4, 5\}, \forall k \in \{0, 1, 2, 3\}.$$

Let the optimal value function at stage k be $v_k^*(x)$. By definition of the DP Algorithm, $\forall k \in [0, N-1]$, we have

$$v_k^*(x) = \max_{u \in \{0 \leq x+u \leq 5, u \text{ is integer}\}} \{r_{k+1}(x, u) + \underbrace{\sum_{x' \in [0,5]} p(x'|x, u) v_{k+1}^*(x')}_{v_{k+1}^*(x+u)}\}.$$

Apply DP-Algorithm, at stage 3, we have

$$\begin{cases} v_3^*(x=0) = \max_{u \in [0,5]}\{-2u^2\} = 0 \Rightarrow u_3^*(x=0) = 0 \\ v_3^*(x=1) = \max_{u \in [-1,4]}\{1 - 2u^2\} = 1 \Rightarrow u_3^*(x=1) = 0 \\ v_3^*(x=2) = \max_{u \in [-2,3]}\{2^2 - 2u^2\} = 4 \Rightarrow u_3^*(x=2) = 0 \\ v_3^*(x=3) = \max_{u \in [-3,2]}\{3^2 - 2u^2\} = 9 \Rightarrow u_3^*(x=3) = 0 \\ v_3^*(x=4) = \max_{u \in [-4,1]}\{4^2 - 2u^2\} = 16 \Rightarrow u_3^*(x=4) = 0 \\ v_3^*(x=5) = \max_{u \in [-5,0]}\{5^2 - 2u^2\} = 25 \Rightarrow u_3^*(x=5) = 0, \end{cases}$$

hence, $v_3^*(x) = 25$ correspond to $x_3 = 5$ and $u_3 = 0$. Continue to apply for stage 2, we have

$$\begin{cases} v_2^*(x=0) = \max_{u \in [0,5]}\{-2u^2 + v_3^*(x'=u)\} = 0 \Rightarrow u_2^*(x=0) = 0 \\ v_2^*(x=1) = \max_{u \in [-1,4]}\{1 - 2u^2 + v_3^*(x'=1+u)\} = 3 \Rightarrow u_2^*(x=1) = 1 \\ v_2^*(x=2) = \max_{u \in [-2,3]}\{2^2 - 2u^2 + v_3^*(x'=2+u)\} = 12 \Rightarrow u_2^*(x=2) = 2 \\ v_2^*(x=3) = \max_{u \in [-3,2]}\{3^2 - 2u^2 + v_3^*(x'=3+u)\} = 26 \Rightarrow u_2^*(x=3) = 2 \\ v_2^*(x=4) = \max_{u \in [-4,1]}\{4^2 - 2u^2 + v_3^*(x'=4+u)\} = 39 \Rightarrow u_2^*(x=4) = 1 \\ v_2^*(x=5) = \max_{u \in [-5,0]}\{5^2 - 2u^2 + v_3^*(x'=5+u)\} = 50 \Rightarrow u_2^*(x=5) = 0, \end{cases}$$

hence, $v_2^*(x) = 50$ correspond to $x_2 = 5$ and $u_2 = 0$. Continue to apply for stage 1, we have

$$\begin{cases} v_1^*(x=0) = \max_{u \in [0,5]}\{-2u^2 + v_2^*(x'=u)\} = 8 \Rightarrow u_1^*(x=0) = 3 \\ v_1^*(x=1) = \max_{u \in [-1,4]}\{1 - 2u^2 + v_2^*(x'=1+u)\} = 22 \Rightarrow u_1^*(x=1) = 3 \\ v_1^*(x=2) = \max_{u \in [-2,3]}\{2^2 - 2u^2 + v_2^*(x'=2+u)\} = 50 \Rightarrow u_1^*(x=2) = 3 \\ v_1^*(x=3) = \max_{u \in [-3,2]}\{3^2 - 2u^2 + v_2^*(x'=3+u)\} = 51 \Rightarrow u_1^*(x=3) = 2 \\ v_1^*(x=4) = \max_{u \in [-4,1]}\{4^2 - 2u^2 + v_2^*(x'=4+u)\} = 64 \Rightarrow u_1^*(x=4) = 1 \\ v_1^*(x=5) = \max_{u \in [-5,0]}\{5^2 - 2u^2 + v_2^*(x'=5+u)\} = 75 \Rightarrow u_1^*(x=5) = 0, \end{cases}$$

hence, $v_1^*(x) = 75$ correspond to $x_1 = 5$ and $u_1 = 0$. Consider at stage 0, we have the initial state $x_0 = 1$, combining with the result at stage 1, $x_1 = 5$, then $u_0 = x_1 - x_0 = 4$ and the optimal of return function is $\sum_{k=0}^{3} x_k^2 - 2u_k^2 = 44$. As a consequence, the optimal control sequence is $u_0^* = 4, u_1^* = 0, u_2^* = 0, u_3^* = 0$.