Solutions of Problem set 3, AIES track
Optimal control and reinforcement learning, 4SC000, TU/e

Duarte Antunes

Q2, 2022-2023

Model-based stochastic rollout, Adaptive sampling, Sequential dynamic programming

Problem 1.1

Differently from dynamic programming, in approximate dynamic programming (or stochastic rollout) we do not need to start at the end decision stage to obtain the decision.

At time $k = 0$ the decision (either $\sigma_0 = 1$ or $\sigma_0 = 2$) is determined by computing the expected cost of the following two decision sequences (note that the base policy is $\sigma_1 = 1$, $\sigma_2 = 1$)

$$\sigma_0 = 1, \sigma_1 = 1, \sigma_2 = 1$$

$$\sigma_0 = 2, \sigma_1 = 1, \sigma_2 = 1$$

Let us start by computing this expected cost for the first sequence in which case

$$x_{k+1} = A_1 x_k$$

for $k \in \{0, 1, 2\}$. The formula and the explanation to arrive at it can be found in a Live Script of Lecture9 called MPCRollout.mlx, and is given by

$$x_0^\intercal P_0 x_0 + \alpha_0$$

with P_0 the first matrix of the iteration

$$P_i = A_{\sigma_i}^\intercal P_{i+1} A_{\sigma_i} + Q \tag{1}$$

for $i \in \{2, 1, 0\}$ with $P_3 = I$ (since the terminal cost is $x_3^\intercal x_3$) and $Q = I$ (since the running cost is $x_k^\intercal x_k$) and $\alpha_0 = \text{trace}(P_1 W) + \text{trace}(P_2 W) + \text{trace}(P_3 W)$ with $W = \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix}$; here $\sigma_i = 1$ for $i \in \{2, 1, 0\}$. The same formula holds for the second sequence but then $\sigma_0 = 2$. Computing these we obtain, for the first sequence

$$P_2 = \begin{bmatrix} 1.8100 & 0.0900 \\ 0.0900 & 1.6500 \end{bmatrix} \quad P_1 = \begin{bmatrix} 2.4661 & 0.2277 \\ 0.2277 & 2.0885 \end{bmatrix} \quad P_0 = \begin{bmatrix} 2.9975 & 0.3859 \\ 0.3859 & 2.3977 \end{bmatrix}$$

For the second sequence we obtain the same except for P_0, which is now denoted by $\bar{P}_0$

$$\bar{P}_0 = \begin{bmatrix} 2.5783 & 0.3612 \\ 0.3612 & 2.7573 \end{bmatrix}$$

We no not need to compute α_0 as it is the same for both sequence and thus it will not play a role while comparing the cost of these two. We can now compare the costs of the first sequence

$$x_0^\intercal P_0 x_0 + \alpha_0$$

and of the second sequence

$$x_0^\intercal \bar{P}_0 x_0 + \alpha_0$$

and provide the following policy

$$\sigma_0 = \begin{cases} 1 \text{ if } x_0^\intercal P_0 x_0 \leq x_0^\intercal \bar{P}_0 x_0 \\ 0 \text{ otherwise} \end{cases}.$$

At time $k = 1$ the decision (either $\sigma_1 = 1$ or $\sigma_1 = 2$) is determined by computing the expected cost of the following two decision sequences (note that the base policy is $\sigma_2 = 1$)

$$\sigma_1 = 1, \sigma_2 = 1$$

$$\sigma_1 = 2, \sigma_2 = 1$$

Let us start by computing this expected cost for the first sequence in which case

$$x_{k+1} = A_1 x_k$$

for $k \in \{1, 2\}$. Similarly to the previous case, this is given by $x_1^\intercal P_1 x_1 + \alpha_1$ with $\alpha_1 = \text{trace}(P_2 W) + \text{trace}(P_3 W)$ and P_1 obtained by (1). The same formula holds for the second sequence but then $\sigma_1 = 2$. Computing these we obtain, for the first sequence the same value as before for P_1 For the second sequence we obtain $\bar{P}_1$

$$\bar{P}_1 = \begin{bmatrix} 2.1584 & 0.2096 \\ 0.2096 & 2.3708 \end{bmatrix}$$

We no not need to compute α_1 as it is the same for both sequence and thus it will not play a role while comparing the cost of these two. We can now compare the costs of the first sequence

$$x_1^\intercal P_1 x_1 + \alpha_1$$

and of the second sequence

$$x_1^\intercal \bar{P}_1 x_1 + \alpha_1$$

and provide the following policy

$$\sigma_1 = \begin{cases} 1 \text{ if } x_1^\intercal P_1 x_1 \leq x_1^\intercal \bar{P}_1 x_1 \\ 0 \text{ otherwise} \end{cases}.$$

Finally at $k = 2$ there are simply two options $\sigma_2 = 1$ or $\sigma_2 = 2$. The first option has cost

$$x_2^\intercal P_2 x_2 + \alpha_2$$

with P_2 given as above and $\alpha_2 = \text{trace}(P_3 W)$; the second option has cost

$$x_2^\intercal \bar{P}_2 x_2 + \alpha_2$$

with

$$\bar{P}_2 = A_2^\intercal P_3 A_2 + Q = \begin{bmatrix} 1.6400 & 0.0800 \\ 0.0800 & 1.8200 \end{bmatrix}$$

The policy is

$$\sigma_2 = \begin{cases} 1 \text{ if } x_2^\intercal P_2 x_1 \leq x_2^\intercal \bar{P}_2 x_2 \\ 0 \text{ otherwise} \end{cases}.$$

(ii) To compute σ_0 according to the policy computed in (i), we need to compute the costs $x_0^\intercal P_0 x_0 = 2.6720, x_0^\intercal \bar{P}_0 x_0 = 3.0050$ allowing us to conclude that the optimal decision is $\sigma_0 = 1$. Thus x_1 is given by

$$x_1 = A_1 x_0 + w_0 = \begin{bmatrix} 0.9 & 0.1 \\ 0 & 0.8 \end{bmatrix} \begin{bmatrix} 0.2 \\ 1 \end{bmatrix} + \begin{bmatrix} 0 \\ -0.5 \end{bmatrix} = \begin{bmatrix} 0.2800 \\ 0.3000 \end{bmatrix}$$

To compute σ_1 according to the policy computed in (i), we need to compute the costs $x_1^\intercal P_1 x_1 = 0.4196, x_1^\intercal \bar{P}_0 x_1 = 0.4178$ allowing us to conclude that the optimal decision is $\sigma_1 = 2$. Thus x_2 is given by

$$x_2 = A_2 x_1 + w_1 = \begin{bmatrix} 0.8 & 0.1 \\ 0 & 0.9 \end{bmatrix} \begin{bmatrix} 0.28 \\ 0.3 \end{bmatrix} + \begin{bmatrix} 0.2 \\ 0 \end{bmatrix} = \begin{bmatrix} 0.4540 \\ 0.2400 \end{bmatrix}$$

To compute σ_2 according to the policy computed in (i), we need to compute the costs $x_2^\intercal P_2 x_2 = 0.4877, x_2^\intercal \bar{P}_2 x_2 = 0.4603$ allowing us to conclude that the optimal decision is $\sigma_2 = 2$.

Problem 1.2

This is a hard exercise since we have not learned how to compute the cost of a base policy in the context of switched linear systems for discounted cost problems (for regular infinite-horizon non-discounted problems, this can be simply obtained by the Lyapunov equation). We need the discount cost when we have persistent disturbances since otherwise the cost is infinity.

In the context of this exercise we have a base policy $\sigma_k = 2$ for all k and we want to compute the following cost

$$\mathbb{E}[\sum_{k=0}^{\infty} \alpha^k x_k^\intercal Q x_k]$$

when

$$x_{k+1} = A_2 x_k.$$

This cost is given by

$$x_0^\intercal P x_0 + \frac{\alpha}{1-\alpha} \text{trace}(PW)$$

where P is the unique solution to the linear system

$$P = \alpha A_2^\intercal P A_2 + Q$$

The justification is omitted; this can be obtained combining several arguments used in the course. Solving this linear system we obtain

$$P = \begin{bmatrix} 3.0667 & 0.3252 \\ 0.3252 & 3.6900 \end{bmatrix}$$

Moreover, $\frac{\alpha}{1-\alpha}\text{trace}(PW) = 154.831$.

Since we have a rollout policy with horizon $H = 2$ we need to compute the expected cost of the following four sequences to compute the policy at time $k = 0$

$$\sigma_0 = 1, \sigma_1 = 1, \sigma_k = 2, k \geq 2$$

$$\sigma_0 = 1, \sigma_1 = 2, \sigma_k = 2, k \geq 2$$

$$\sigma_0 = 2, \sigma_1 = 1, \sigma_k = 2, k \geq 2$$

$$\sigma_0 = 2, \sigma_1 = 2, \sigma_k = 2, k \geq 2$$

These costs are summarized in the table

index i	Sequence	Expected cost	$\rho(i)$
1	$\sigma_0 = 1, \sigma_1 = 1, \sigma_k = 2, k \geq 2$	$x_0^\intercal P_1 x_0 + \alpha_1$	1
2	$\sigma_0 = 1, \sigma_1 = 2, \sigma_k = 2, k \geq 2$	$x_0^\intercal P_2 x_0 + \alpha_2$	1
3	$\sigma_0 = 2, \sigma_1 = 1, \sigma_k = 2, k \geq 2$	$x_0^\intercal P_3 x_0 + \alpha_3$	2
4	$\sigma_0 = 2, \sigma_1 = 2, \sigma_k = 2, k \geq 2$	$x_0^\intercal P_4 x_0 + \alpha_4$	2

where

$$P_1 = \alpha A1^\intercal \tilde{P} A_1 + Q, \quad P_2 = \tilde{P}, \quad P_3 = \alpha A2^\intercal \tilde{P} A_2 + Q, \quad P4 = P, \quad \tilde{P} = \alpha A_1^\intercal P_2 A_1 + Q$$

and

$$\alpha_1 = \alpha \text{trace}(\tilde{P}W) + \alpha\alpha_4 \quad \alpha_2 = \alpha_4 \quad \alpha_3 = \alpha_1 \quad \alpha_4 = \frac{\alpha}{1-\alpha}\text{trace}(PW).$$

The numerical values are

$$P_1 = \begin{bmatrix} 2.4188 & 0.4490 \\ -0.0747 & 0.2587 \end{bmatrix}, \; P_2 = \begin{bmatrix} 6.3436 & -0.1711 \\ -0.1711 & 1.8303 \end{bmatrix}, \; P_3 = \begin{bmatrix} 2.4188 & 0.4490 \\ -0.0747 & 0.2587 \end{bmatrix} \; P_4 = \begin{bmatrix} 2.4188 & 0.4490 \\ -0.0747 & 0.2587 \end{bmatrix},$$

Again these can be obtained by combining arguments learned in the course, and in particular using arguments very similar to the ones in the Live Script of Lecture9 called MPCRollout.mlx. In the table, the first decision of each cost sequence is summarized by ρ. The optimal policy is then

$$\sigma_k = \rho(\text{argmin}_{i\in\{1,2,3,4\}} x_k^\intercal P_i x_k)$$

Problem 1.3 We need to evaluate the cost for $j \in \{1, 2, 3, 4\}$ and pick the best option. For $j = 1$ we have from the values in the table

$$1.1 - \sqrt{\frac{\log(10)}{1}} = -0.4174$$

For $j = 2$

$$(0.9 + 0.25 + 0.4)/3 - \sqrt{\frac{\log(10)}{3}} = -0.3594$$

For $j = 3$:

$$(0.5 + 0.33 + 0.4)/3 - \sqrt{\frac{\log(10)}{3}} = -0.4661$$

For $j = 4$:

$$(0.2 + 1 + 0.8)/3 - \sqrt{\frac{\log(10)}{3}} = -0.2094$$

Therefore,

$$a_{11} = 3$$

since it corresponds to the smallest cost.

Problem 1.4 For $c = 0$. For $j = 1$ we have from the values in the table that the cost is

$$1.1$$

For $j = 2$

$$(0.9 + 0.25 + 0.4)/3 = 0.5167$$

For $j = 3$:

$$(0.5 + 0.33 + 0.4)/3 = 0.41$$

For $j = 4$:

$$(0.2 + 1 + 0.8)/3 = 0.6667$$

Therefore, again

$$a_{11} = 3$$

since it corresponds to the smallest cost. For $c = 2$ For $j = 1$ we have from the values in the table

$$1.1 - 2\sqrt{\frac{\log(10)}{1}} = -1.9349$$

For $j = 2$

$$(0.9 + 0.25 + 0.4)/3 - 2\sqrt{\frac{\log(10)}{3}} = -1.2355$$

For $j = 3$:

$$(0.5 + 0.33 + 0.4)/3 - 2\sqrt{\frac{\log(10)}{3}} = -1.3422$$

For $j = 4$:

$$(0.2 + 1 + 0.8)/3 - 2\sqrt{\frac{\log(10)}{3}} = -1.0855$$

Therefore,

$$a_{11} = 1$$

since it corresponds to the smallest cost.

When $c = 0$ we pick simply based on the cost. When $c = 1$ there is an incentive to explore but not yet sufficient to change the optimal decision decision $a_{11} = 3$. However, when $c = 2$ this incentive is larger and leads to a change in decision to $a_{11} = 1$.

Problem 1.5 Since the approximating cost-to-go function is linear in the parameters one can just apply the formula of slide 25 of lecture 10 (this formula can be obtained by differentiating the cost function with respect to the parameters and setting it to zero).

$$r_{h-1} = [\sum_{s=1}^{q} \phi_{h-1}\left(x_{h-1}^s\right) \phi_{h-1}\left(x_{h-1}^s\right)^\top]^{-1} \sum_{s=1}^{q} \phi_{h-1}\left(x_{h-1}^s\right) \beta^s$$

For the provided basis functions

$$\phi_{h-1}(x) = \begin{bmatrix} 1 & x & x^2 \end{bmatrix}^\mathsf{T}$$

and data

s	1	2	3	4	5
x_{h-1}^s	-4	-2	0	1	3
β_{h-1}^s	17.49	6.87	0.73	-0.55	-1.19

we get

$$\begin{aligned}\sum_{s=1}^{q} \phi_{h-1}\left(x_{h-1}^s\right) \phi_{h-1}\left(x_{h-1}^s\right)^\top &= \begin{bmatrix} 1 & -4 & 16 \\ -4 & 16 & -64 \\ 16 & -64 & 256 \end{bmatrix} + \begin{bmatrix} 1 & -2 & 4 \\ -2 & 4 & -8 \\ 4 & -8 & 16 \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} + \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \\ &+ \begin{bmatrix} 1 & 3 & 9 \\ 3 & 9 & 27 \\ 9 & 27 & 81 \end{bmatrix} \\ &= \begin{bmatrix} 5 & -2 & 30 \\ -2 & 30 & -44 \\ 30 & -44 & 354 \end{bmatrix}\end{aligned}$$

$$\sum_{s=1}^{q} \phi_{h-1}\left(x_{h-1}^s\right) \beta^s = \begin{bmatrix} 1 \\ -4 \\ 16 \end{bmatrix} 17.49 + \begin{bmatrix} 1 \\ -2 \\ 4 \end{bmatrix} 6.87 + \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} 0.73 + \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} (-0.55) + + \begin{bmatrix} 1 \\ 3 \\ 9 \end{bmatrix} (-1.19) = \begin{bmatrix} 23.3443 \\ -87.8124 \\ 296.0432 \end{bmatrix}$$

So that

$$\hat{r}_{h-1} = \begin{bmatrix} 5 & -2 & 30 \\ -2 & 30 & -44 \\ 30 & -44 & 354 \end{bmatrix}^{-1} \begin{bmatrix} 23.3443 \\ -87.8124 \\ 296.0432 \end{bmatrix} = \begin{bmatrix} 0.7925 \\ -2.1355 \\ 0.5037 \end{bmatrix}$$

Problem 1.6 Following a similar procedure to the previous exercise

$$\sum_{s=1}^{q} \phi_{h-1}\left(x_{h-1}^{s}\right) \phi_{h-1}\left(x_{h-1}^{s}\right)^{\top} = \begin{bmatrix} 0.4272 & -0.4947 \\ -0.4947 & 0.5728 \end{bmatrix} + \begin{bmatrix} 0.1732 & 0.3784 \\ 0.3784 & 0.8268 \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 0.2919 & 0.4546 \\ 0.4546 & 0.7081 \end{bmatrix}$$
$$+ \begin{bmatrix} 0.9801 & -0.1397 \\ -0.1397 & 0.0199 \end{bmatrix} = \begin{bmatrix} 2.8724 & 0.1987 \\ 0.1987 & 2.1276 \end{bmatrix}$$

$$\sum_{s=1}^{q} \phi_{h-1}\left(x_{h-1}^{s}\right) \beta^{s} = \begin{bmatrix} -0.6536 \\ 0.7568 \end{bmatrix} 17.49 + \begin{bmatrix} -0.4161 \\ -0.9093 \end{bmatrix} 6.87 + \begin{bmatrix} 1 \\ 0 \end{bmatrix} 0.73 + \begin{bmatrix} 0.5403 \\ 0.8415 \end{bmatrix} (-0.55) + + \begin{bmatrix} -0.9900 \\ 0.1411 \end{bmatrix} (-1.19) = \begin{bmatrix} -12.6802 \\ 6.3589 \end{bmatrix}$$

So that

$$\hat{r}_{h-1} = \begin{bmatrix} 2.8724 & 0.1987 \\ 0.1987 & 2.1276 \end{bmatrix}^{-1} \begin{bmatrix} -12.6802 \\ 6.3589 \end{bmatrix} = \begin{bmatrix} -4.6512 \\ 3.4231 \end{bmatrix}$$

Infinite-horizon problems, Policy iteration

Problem 2.1 (i) We can write this problem as

$$\min \sum_{k=0}^{\infty} \mathbb{E}\left[\bar{g}\left(\bar{x}_k, u_k, w_k\right)\right]$$

for

$$\bar{x}_{k+1} = \bar{f}\left(\bar{x}_k, u_k, w_k\right) = \min(\max(x_{1,k} + u_k + w_k, 0), 3)$$

with $\bar{x}_k = (x_{1,k}, x_{2,k}) = (x_k, k)$,

$$\bar{g}\left(\bar{x}_k, u_k, w_k\right) = \begin{cases} g\left(x_{1,k}, u_k\right) \text{ if } 0 \leq x_{2,k} \leq 1 \\ g\left(x_{1,2}, u_2\right) + g_3\left(\min(\max(x_{1,2} + u_2 + w_2, 0), 3)\right) \text{ if } x_{2,k} = 2 \\ 0 \text{ if } x_{2_k} \geq 3 \end{cases}$$

(jj) The Bellman equation for this shortest path problem is

$$J(\bar{x}) = \min_{u \in \{-1,0,1\}} \mathbb{E}[\bar{g}(\bar{x}, u, w) + J(\bar{f}(\bar{x}, u, w))]$$

The trivial solution for $k \geq 3$ is $J(\bar{x}) = J(x, k) = 0$. Specializing this equation to $\bar{x} = (x_0, 0)$, $\bar{x} = (x_1, 1)$, $\bar{x} = (x_2, 2)$ and using the notation $J_k(x) = J(\bar{x})$ when $\bar{x} = (x, k)$, we have

$$\begin{aligned} J_0(x_0) &= \min_{u \in \{-1,0,1\}} g(x_0, u) + \mathbb{E}[J_1(\min(\max(x_0 + u_0 + w_0, 0), 3))] \\ J_1(x_1) &= \min_{u \in \{-1,0,1\}} g(x_1, u) + \mathbb{E}[J_2(\min(\max(x_1 + u_1 + w_1, 0), 3))] \\ J_2(x_2) &= \min_{u \in \{-1,0,1\}} \mathbb{E}[\bar{g}(x_2, u_2, w_2)] \\ &= \min_{u \in \{-1,0,1\}} g(x_2, u_2) + \mathbb{E}[g_3(\min(\max(x_2 + u_2 + w_2, 0), 3))] \end{aligned}$$

These are the equations of the stochastic dynamic programming algorithm which provide the optimal costs-to-go.

Problem 2.2 (i) Let $\bar{w}_k = (w_k, w_{2,k})$ with $w_{2,k} \in \{0, 1\}$, $\text{Prob}[w_{2,k} = 1] = \alpha$, $k \in \mathbb{N})_0$. Consider the following stochastic shortest path problem

$$\sum_{k=0}^{\infty} \mathbb{E}[\bar{g}(\bar{x}_k, u_k)]$$

$$\bar{x}_{k+1} = \bar{f}(\bar{x}_k, u_k, \bar{w}_k) = \begin{cases} \min(\max(x_k + u_k + w_k, 0), 3) \text{ if } w_{2,k} = 1 \\ -1 \text{ if } w_{2,k} = 0 \text{ or } \bar{x}_k = -1 \end{cases}$$

with $\bar{g}(\bar{x}_k, u_k) = g(\bar{x}_k, u_k)$ if $\bar{x}_k \neq -1$ and $\bar{g}(-1, u_k) = 0$ for every u_k. The state -1 corresponds to an additional state typically labelled $n + 1$. Note that

$$\begin{aligned} \mathbb{E}[\bar{g}(x_k, u_k)] =& \mathbb{E}[\bar{g}(\bar{x}_k, u_k) | w_{2,0} = 1, w_{2,1} = 1, \ldots, w_{2,k-1} = 1] \underbrace{\text{Prob}[w_{2,0} = 1, w_{2,1} = 1, \ldots, w_{2,k-1} = 1]}_{\alpha^k} \\ &+ \underbrace{\mathbb{E}[\bar{g}(\bar{x}_k, u_k) | w_{2,\ell} = 0 \text{ for some } \ell \in \{0, \ldots, k-1\}]}_{=0} \text{Prob}[w_{2,\ell} = 0 \text{ for some } \ell \in \{0, \ldots, k-1\}] \end{aligned}$$

Note that if $w_{2,0} = 1, w_{2,1} = 1, \dots, w_{2,k-1}$ then the system evolves as in the definition of the discounted cost problem

$$x_{k+1} = \min(\max(x_k + u_k + w_k, 0), 3), \quad k \in \{0, 1, \dots\},$$

Thus, for this problem and in this event, $\mathbb{E}[\bar{g}(\bar{x}_k, u_k)] = \alpha^k \mathbb{E}[g(x_k, u_k)]$

$$\sum_{k=0}^{\infty} \mathbb{E}[\bar{g}(\bar{x}_k, u_k)] = \sum_{k=0}^{\infty} \alpha^k \mathbb{E}[g(x_k, u_k)]$$

$$x_{k+1} = \min(\max(x_k + u_k + w_k, 0), 3)$$

which is the original discounted cost problem.

(ii) The Bellman equation for this shortest path problem is

$$J(\bar{x}) = \min_{u \in \{-1,0,1\}} \bar{g}(\bar{x}, u) + \mathbb{E}[J(\bar{f}(\bar{x}, u, w))]$$

we have $J(-1) = 0$ (since after reaching state -1 the cost is zero always, thus this is a good candidate for a solution which we can confirm by uniqueness) and $\mathbb{E}[J(\bar{f}(\bar{x}, u, w))] = \mathbb{E}[J(f(x, u, w))|w_2 = 1] \underbrace{\text{Prob}[w_2 = 1]}_{=\alpha} + \underbrace{\mathbb{E}[J(-1)|w_2 = 0]}_{=0} \text{Prob}[w_2 = 0]$ so that we can write

$$J(x) = \min_{u \in \{-1,0,1\}} g(x, u) + \alpha \mathbb{E}[J(f(x, u, w))], \quad x \in \{0, 1, 2, 3\}$$

which is the Bellman equation for the original discounted cost problem.

Problem 2.3 We can convert the finite horizon problem into a shortest path problem by considering the augmented state (x_k, k). However, here we still use the usual notation in the context of Dynamic programming. Policy evaluation of policy μ_k^0 is the same as running DP but assuming for each state there is only one action for each state and stage pair (x_k, k), which is $\mu_k^0(x_k)$. Then we obtain the recursion

$$J_k(x_k) = \mathbb{E}[g(x_k, \mu_k^0(x_k), d_k) + J_{k+1}(\max\{x_k + \mu_k^0(x_k) - d_k, 0\}))]$$

and since $\mu_k^0(x_k) = 0$ for every state and stage, this boils down to

$$J_k(x_k) = \mathbb{E}[g(x_k, 0, d_k) + J_{k+1}(\max\{x_k - d_k, 0\}))]$$

or

$$J_k(x_k) = \mathbb{E}[x_k - 9\min\{x_k, d_k\} + J_{k+1}(\max\{x_k - d_k, 0\}))]$$

Finally, evaluating at $k \in \{0, 1, 2\}$ we obtain

$$J_0(x_0) = x_0 - 9\min\{x_0, 1\} + J_1(\max\{x_0 - 1, 0\}))]$$

$$J_1(x_1) = x_1 + \frac{1}{3}\big((-9\min\{x_1, 0\} + J_2(\max\{x_1, 0\})) + (-9\min\{x_1, 1\} + J_2(\max\{x_1 - 1, 0\})) \\ + (-9\min\{x_1, 2\} + J_2(\max\{x_1 - 2, 0\}))\big)$$

and

$$J_2(x_2) = x_2 - 9\min\{x_2, 1\} + g_3(\max\{x_2 - 1, 0\}))]$$

First we compute the values of $J_2(0)$, $J_2(1)$, $J_2(2)$ leading to the values depicted in the figure at stage $k = 2$. From these values we can obtain the values of $J_1(0)$, $J_1(1)$, $J_1(2)$. It is clear that $J_1(0) = 0$ and the two other values are

$$J_1(1) = 1 - \frac{1}{3}\big((-9\min\{1, 0\} + \underbrace{J_2(\max\{1, 0\}))}_{=-8} + (-9\min\{1, 1\} + J_2(\max\{1 - 1, 0\})) \\ + (-9\min\{1, 2\} + J_2(\max\{1 - 2, 0\}))\big) = -7.66$$

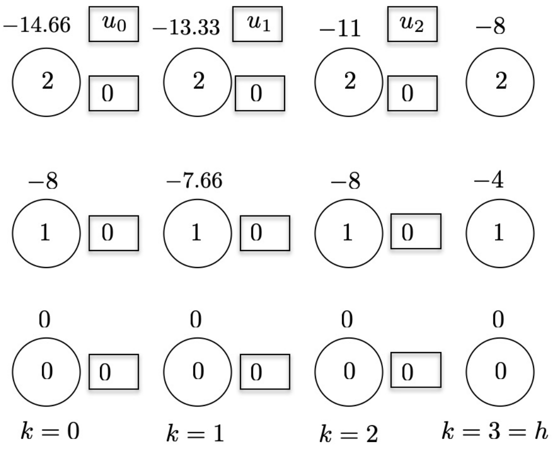

Figure 1: Policy evaluation of μ^0

$$J_1(2) = 2 + \frac{1}{3}\big((-9\min\{2,0\} + \underbrace{J_2(\max\{2,0\})}_{-11}) + (-9\min\{2,1\} + \underbrace{J_2(\max\{2-1,0\})}_{-8})$$
$$+ (-9\min\{2,2\} + J_2(\max\{2-2,0\}))\big) = -13.33$$

Finally $J_0(0) = 0$,

$$J_0(1) = 1 - 9\min\{1,1\} + J_1(\max\{1-1,0\}))] = -8$$
$$J_0(2) = 2 - 9\min\{2,1\} + J_1(\max\{2-1,0\}))] = 2 - 9 - 7.66 = -14.66$$

These costs are shown in Figure 1 and are henceforth denoted by $J_k^{\mu_0}(x_k)$.

For policy improvement we have for stage $k = 2$

$$\mu_2^1(x_2) = \arg\min_{u_2\in\{0,\ldots,2-x_2\}} g(x_2,u_2,1) + g_3(\max\{x_2+u_2-1,0\})$$

and after replacing the expressions of g and g_3,

$$\mu_2^1(x_2) = \arg\min_{u_2\in\{0,\ldots,2-x_2\}} \big(x_2 + c_2(u_2) - 9\min\{x_2+u_2,1\}\big) + g_3(\max\{x_2+u_2-1,0\})$$

For $x_2 = 2$, we have $\mu_2^1(2) = 0$.

For $x_2 = 1$,

$$u_2 = 0 \to \big(1 + -9\min\{1,1\}\big) + g_3(\max\{1+0-1,0\}) = -8$$
$$u_2 = 1 \to \big(1 + 5.9 - 9\min\{1+1,1\}\big) + \underbrace{g_3(\max\{1+1-1,0\})}_{-4} = -6.1$$

so that $\mu_2^1(1) = 0$.

For $x_2 = 0$,

$$u_2 = 0 \to \big(0 + 0 - 9\min\{0+0,1\}\big) + g_3(\max\{0+0-1,0\}) = 0$$
$$u_2 = 1 \to \big(0 + 5.9 - 9\min\{0+1,1\}\big) + g_3(\max\{0+1-1,0\}) = --3.1$$
$$u_2 = 2 \to \big(0 + 10.9 - 9\min\{0+2,1\}\big) + g_3(\max\{0+2-1,0\}) = -2.1$$

so that $\mu_2^1(0) = 1$.

At decision stage $k = 1$,

$$\mu_1^1(x_1) = \arg\min_{u_1\in\{0,\ldots,2-x_1\}} \mathbb{E}[g(x_1,u_1,d_1) + J_2^{\mu_0}(\max\{x_2+u_2-d_1,0\})]$$

for state $x_1 = 2$ there is only one option, so that $\mu_1^1(2) = 0$.

For $x_1 = 1$ there are two options. For $u_1 = 0$ the costs are summarized in the next table

$d_1 = 0$	$d_1 = 1$	$d_1 = 2$
$1 - 8 = -7$	$-9 + 1 = -8$	$-9 + 1 = -8$

leading to an expected cost $-7/3 - 8/3 - 8/3 = -7.66$ and for $u_1 = 1$ the costs are summarized in the next table

$d_1 = 0$	$d_1 = 1$	$d_1 = 2$
$1 + 5.9 - 11 = -4.1$	$1 + 5.9 - 9 - 8 = -10.1$	$1 + 5.9 - 9 \times 2 = -11.1$

leading to an expected cost -8.43. Therefore the optimal decision is

$$\mu_1^1(1) = 1$$

When $x_1 = 0$, there are three options $u_1 \in \{0, 1, 2\}$. For $u_1 = 0$ the costs are

$d_1 = 0$	$d_1 = 1$	$d_1 = 2$
0	0	0

For $u_1 = 1$ the costs are

$d_1 = 0$	$d_1 = 1$	$d_1 = 2$
$5.9 - 8 = -2.1$	$5.9 - 9 = -3.1$	$5.9 - 9 = -3.1$

For $u_1 = 2$ the costs are

$d_1 = 0$	$d_1 = 1$	$d_1 = 2$
$10.9 - 11 = -0.1$	$10.9 - 9 - 8 = -6.1$	$10.9 - 9 \times 2 = -7.1$

The expected costs of $u_0 = 0$, $u_0 = 1$, $u_0 = 2$, are 0, -2.766, -4.433, respectively. Therefore the optimal decision is

$$\mu_1^1(0) = 2$$

At stage $k = 0$

$$\mu_0^1(x_0) = \arg \min_{u_0 \in \{0, \dots, 2 - x_0\}} g(x_0, u_0, 1) + J_1^{\mu_0}(\max\{x_0 + u_0 - 1, 0\})$$

and after replacing the expressions of g and $J_1^{\mu_0}$,

$$\mu_0^1(x_0) = \arg \min_{u_0 \in \{0, \dots, 2 - x_0\}} \big(x_0 + c_0(u_0) - 9 \min\{x_0 + u_0, 1\}\big) + J_1^{\mu_0}(\max\{x_0 + u_0 - 1, 0\})$$

For $x_0 = 2$, we have $\mu_0^1(2) = 0$.

For $x_0 = 1$,

$$u_0 = 0 \to \big(1 + -9 \min\{1, 1\}\big) + J_1^{\mu_0}(\max\{1 + 0 - 1, 0\}) = -8$$

$$u_2 = 1 \to \big(1 + 5.9 - 9 \min\{1 + 1, 1\}\big) + \underbrace{J_1^{\mu_0}(\max\{1 + 1 - 1, 0\})}_{-7.66} = -9.76$$

so that $\mu_0^1(1) = 1$.

For $x_0 = 0$,

$$u_2 = 0 \to \big(0 + 0 - 9 \min\{0 + 0, 1\}\big) + J_1^{\mu_0}(\max\{0 + 0 - 1, 0\}) = 0$$

$$u_2 = 1 \to \big(0 + 5.9 - 9 \min\{0 + 1, 1\}\big) + J_1^{\mu_0}(\max\{0 + 1 - 1, 0\}) = -3.1$$

$$u_2 = 2 \to \big(0 + 10.9 - 9 \min\{0 + 2, 1\}\big) + J_1^{\mu_0}(\max\{0 + 2 - 1, 0\}) = -5.76$$

so that $\mu_0^0(0) = 2$.

The policy is summarized in Figure 2.

Problem 2.4 $J_{\mu_0}(1)$, $J_{\mu_0}(2)$, $J_{\mu_0}(3)$ are described by

$$J_{\mu_0}(i) = \mathbb{E}[\sum_{k=0}^{\infty} \alpha^k \big(c_1(x_k) + c_2(\mu_0(x_k)) - p \min\{x_k + \mu_0(x_k), d_k\}\big) | x_0 = i], \quad i \in \{0, 1, \dots, N\},$$

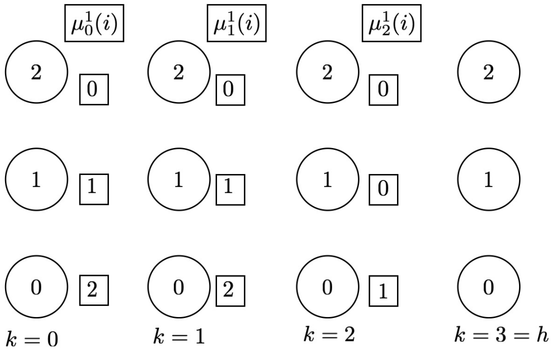

Figure 2: Policy μ^1

can be obtained by solving the following linear system

$$J_{\mu_0}(i) = \mathbb{E}[c_1(i) + c_2(\mu_0(i)) - p\min\{i+\mu_0(i), d_k\} + \alpha J_{\mu_0}(\max\{i+\mu_0(i)-d_k, 0\})], i \in \{0,1,2\}$$

$$\begin{bmatrix} J_{\mu_0}(0) \\ J_{\mu_0}(1) \\ J_{\mu_0}(2) \end{bmatrix} = \mathbb{E}[\begin{bmatrix} c_1(0) + c_2(\mu_0(0)) - p\min\{0+\mu_0(0), d_k\} \\ c_1(1) + c_2(\mu_0(1)) - p\min\{1+\mu_0(1), d_k\} \\ c_1(2) + c_2(\mu_0(2)) - p\min\{2+\mu_0(2), d_k\} \end{bmatrix}] + \alpha \begin{bmatrix} 1 & 0 & 0 \\ 2/3 & 1/3 & 0 \\ 1/3 & 1/3 & 1/3 \end{bmatrix} \begin{bmatrix} J_{\mu_0}(0) \\ J_{\mu_0}(1) \\ J_{\mu_0}(2) \end{bmatrix} \quad (2)$$

or equivalently

$$\begin{bmatrix} J_{\mu_0}(0) \\ J_{\mu_0}(1) \\ J_{\mu_0}(2) \end{bmatrix} = \begin{bmatrix} 0 \\ 1-(9\times 1/3 + 9 \times 1/3) \\ 2-(1/3\times 0 + 1/3\times 9 + 18\times 1/3) \end{bmatrix}] + \alpha \begin{bmatrix} 1 & 0 & 0 \\ 2/3 & 1/3 & 0 \\ 1/3 & 1/3 & 1/3 \end{bmatrix} \begin{bmatrix} J_{\mu_0}(0) \\ J_{\mu_0}(1) \\ J_{\mu_0}(2) \end{bmatrix} \quad (3)$$

Solving this linear system results in

$$\begin{bmatrix} J_{\mu_0}(0) \\ J_{\mu_0}(1) \\ J_{\mu_0}(2) \end{bmatrix} = \begin{bmatrix} 0.0000 \\ -7.1429 \\ -13.0612 \end{bmatrix}$$

For the policy improvement step we simply have to compute

$$\mu_1(i) = \arg\min_{u\in\{0,\dots,2-i\}} \mathbb{E}[g(i,u,d_k) + \alpha J_{\mu_0}(\max\{i+u-d_k, 0\})]$$

When $i=2$ there is only one option, so that $\mu_1(2)=0$.

For $i=1$ there are two options. For $u=0$ the costs are summarized in the next table

$d_k=0$	$d_k=1$	$d_k=2$
$1-0.9\times 7.1429 = -5.4286$	$-9+1=-8$	$-9+1=-8$

leading to an expected cost $-5.4286/3 - 8/3 - 8/3 = -7.1429$ and for $u_1 = 1$ the costs are summarized in the next table

$d_k=0$	$d_k=1$	$d_k=2$
$1+5.9-0.9\times 13.0612 = -4.8551$	$1+5.9-9-0.9\times 7.1429 = -8.5286$	$1+5.9-9\times 2 = -11.1$

leading to an expected cost -8.1612. Therefore the optimal decision is

$$\mu_1(1) = 1$$

When $i=0$, there are three options $u\in\{0,1,2\}$. For $u=0$ the costs are

$d_k=0$	$d_k=1$	$d_k=2$
0	0	0

For $u=1$ the costs are

$d_k = 0$	$d_k = 1$	$d_k = 2$
$5.9 - 0.9 \times 7.1429 = -0.5286$	$5.9 - 9 = -3.1$	$5.9 - 9 = -3.1$

For $u = 2$ the costs are

$d_k = 0$	$d_k = 1$	$d_k = 2$
$10.9 - 0.9 \times 13.06122 = -0.8551$	$10.9 - 9 - 0.9 \times 7.1429 = -4.5286$	$10.9 - 9 \times 2 = -7.1$

The expected costs of $u = 0$, $u = 1$, $u = 2$, are 0, -2.2429, -4.1612, respectively. Therefore the optimal decision is

$$\mu_1(0) = 2$$

$$\begin{bmatrix} \mu_1(0) \\ \mu_1(1) \\ \mu_1(2) \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix}$$

Problem 2.5 Solution not provided.

Q-Learning

Problem 3.1

ℓ	1	2	3	4	5	6	7	8	9	10
k	0	1	2	3	0	1	2	3	0	1
x_k	1	0	2	2	0	0	1	0	0	1
u_k	1	2	0	-	1	2	1	-	1	1
d_k	2	0	0	-	1	1	2	-	0	1
$g(x_k, u_k, d_k)$	-11.1	10.9	2	-	-3.1	1.9	-11.1	-	5.9	-2.1
$Q_{\ell+1}(k, x_k, u_k)$	-5.55	5.45	-3	-	-1.55	3.675	-5.55	-	2.175	-3.825

The running cost is $g(x_k, u_k, d_k) = \big(x_k + c_2(u_k) - 9\min\{x_k + u_k, d_k\}\big)$.
At $\ell = 1$

$$Q_2(0,1,1) = 0.5 \times g(1,1,2) + 0.5 \min_{v\in\{0,1,2\}} Q_1\left(1,0,v\right)) = 0.5(1 + 5.9 - 18) = -5.55$$

At $\ell = 2$

$$Q_3(1,0,2) = 0.5 \times g(0,2,0) + 0.5 \min_{v\in\{0\}} Q_2\left(2,2,v\right) = 0.5(0 + 10.9 - 0) = 5.45$$

At $\ell = 3$

$$Q_4(2,2,0) = 0.5 \times g(2,0,0) + 0.5\underbrace{g_3(2)}_{-8} = 0.5 \times 2 - 4 = -3$$

At $\ell = 4$, the process reaches the terminal state, there is not decision and no need to update any (x,u) pair. At $\ell = 5$

$$Q_6(0,0,1) = 0.5 \times g(0,1,1) + 0.5 \min_{v\in\{0,1,2\}} Q_5\left(1,0,v\right) = 0.5(-3.1) = -1.55$$

Note that although at iteration $\ell = 2$ the estimate of $Q(1,0,2)$ was updated to a value larger than zero, the fact that the estimates $Q(1,0,1) = 0$ and $Q(1,0,0) = 0$ are zero, results in $\min_{v\in\{0\}} Q_5\left(1,0,v\right) = 0$
At $\ell = 6$

$$Q_7(1,0,2) = 0.5 \times \underbrace{Q_6(1,0,2)}_{=Q_3(1,0,2)=5.45} + 0.5 \times \underbrace{g(0,2,1)}_{1.9} + 0.5 \min_{v\in\{0,1\}} Q_6\left(2,1,v\right) = 0.5(5.45 + 1.9 - 0) = 3.675$$

At $\ell = 7$

$$Q_8(2,1,1) = 0.5 \times \underbrace{g(1,1,2)}_{-11.1} + 0.5 g_3(0) = -5.55$$

At $\ell = 8$, the process reaches the terminal state, there is not decision and no need to update any (x,u) pair.
At $\ell = 9$

$$Q_{10}(0,0,1) = 0.5 \underbrace{Q_9(0,0,1)}_{Q_6(0,0,1)=-1.55} + 0.5 \times \underbrace{g(0,1,0)}_{=5.9} + 0.5 \min_{v\in\{0,1\}} Q_9\left(1,1,v\right) = -0.5 \times 1.55 + 0.5 \times 5.9 = 2.175$$

At $\ell = 10$ (note that although x_{11} is not in the table we can compute it $x_{11} = x_{10} + u_{10} - d_{10} = 1 + 1 - 1 = 1$)

$$Q_{11}(1,1,1) = 0.5 \underbrace{\times g(1,1,1)}_{-2.1} + 0.5 \min_{v\in\{0,1\}} Q_{10}\left(2,1,v\right) = 0.5(-2.1 - 5.55) = -3.825$$

Note that we use the fact that at iteration $\ell = 7$ the estimate of $Q(2,1,1)$ was updated to a value small than zero, and the fact that the estimate $Q(2,1,0)$ is zero, results in $\min_{v\in\{0,1\}} Q\left(2,1,v\right) = -5.55$.

The states $(1,0,2)$ and $(0,0,1)$ were updated twice and the states $(0,1,1)$, $(2,1,1)$, $(2,2,0)$, $(1,1,1)$ were updated once. For these states

$$\begin{aligned}
Q_{11}(1,0,2) &= 3.675\\
Q_{11}(0,0,1) &= 2.175\\
Q_{11}(0,1,1) &= -5.55\\
Q_{11}(2,1,1) &= -5.55\\
Q_{11}(2,2,0) &= -3\\
Q_{11}(1,1,1) &= -3.825
\end{aligned}$$

For all the other states $Q_{11}(.,.,.) = 0$.

Problem 3.2 Solution not provided.

Problem 3.3 Solution not provided.