Stochastic Optimisation - Problem Sheet 5

Dom Hutchinson

December 4, 2020

Question 3)
Consider a queueing system with room for n customers which operates over N time periods (denoted by $0, \dots, N-1$). Regarding this system, we assume the following

- Only one customer can arrive during a period. A new customer arrives during a period with probability p.
- Customer arrivals in different periods are independent.
- A new customer is either allowed to join the queue or be rejected. Rejected customers depart without attempting to re-enter.
- The service of a customer can start/end at the beginning/end of a period.
- A customer in service at the beginning of a period terminates the service at the end of the period with probability q, with $q < p$. This is independent of the number of periods the customer has been in service and of the number of customers in the system.
- The cost of rejecting a customer is C. The cost of maintaining i customers in the queue over a single period is ci. The cost of admitting a new customer is c.

The problem is to decide, at each time period, whether to accept or reject a new customer so as the expected total cost is minimal. Each decision should be based on the current number of customers in the queue.

Question 3) (a)
Formulate the problem of minimizing the total expected cost as a *Markov Decision Problem*. Derive the optimality equations for the formulated decision problem.

Answer 3) (a)
Let X_t be the system state at the start of time-period t and Y_t be the action the agent takes at the start of time-period t.

Let W_t represent the number of people wishing to join the queue in time-period t and Z_t represent the number of customers who leave the queue in time-period t. From the question, W_t, Z_t have the following distributions

$$\begin{aligned}\mathbb{P}(W_t = w) &= \begin{cases} p & w = 1 \\ 1-p & w = 0 \end{cases} \\ \mathbb{P}(Z_t = z) &= \begin{cases} q & z = 1 \\ 1-q & z = 0 \end{cases}\end{aligned}$$

- *Decision Epochs* - At the start of each period.

- *Time-Horizon* - $T = \{0, \dots, N-1\}$.
- *Agent Action.*

  At the start of each time-period, let the agent decide whether to allow customers to join the queue or not. This can be encoded into Y_t as

  $$Y_t = \begin{cases} 1 & \text{if accepting customers} \\ 0 & \text{if \underline{not} accepting customers} \end{cases}$$

  As at most 1 customer may wish to join the queue in a single time-period, Y_t is equivalent to the maximum amount of customer being allowed to join the queue in time-period t.
- *Action-Space.*

  Given the specification of Y_t the action-space is $A = \{0, 1\}$.
- *System States.*

  Knowing the number of customers currently in the queue is sufficient for the agent to determine which action to take. This knowledge can be encoded into X_t, by setting X_t to be the number of people in the queue at the start of time-period t.

  This gives us the state-equation $X_{t+1} = X_t + Y_t W_t - Z_t$.
- *State-Space*

  By the definition of X_t and the fact that the maximum queue capacity is n, the state-space is $S = \{0, 1, \dots, n\}$.
- *Admissible Action-Space.*

  The agent is not allowed to let customers join the queue if the queue is already full.

  $$A(s) = \begin{cases} \{0, 1\} & \text{if } s \in [0, \dots, n-1] \\ \{0\} & \text{if } s = n \end{cases}$$
- *Transition Probabilities.*

  The definition of transition probabilities states

  $$p_t(s'|s, a) := \mathbb{P}^\pi(X_{t+1} = s' | X_t = s, Y_t = a)$$

  By considering the state-equation for our system, we can derive the following

  $$\begin{array}{rcl} p_t(s'|s, a) & = & \mathbb{P}^\pi(X_{t+1} = s' | X_t = s, Y_t = a) \\ & = & \mathbb{P}^\pi(X_t + Y_t W_t - Z_t = s' | X_t = s, Y_t = a) \\ & = & \mathbb{P}(s + a W_t - Z_t = s') \end{array}$$

  To get a deterministic value for this expression, we consider the following cases involving s', s

  i). $s = 0$.

The queue is empty so no-one can leave the queue (ie $\mathbb{P}(Z_t = 0) = 1$).

$$\begin{aligned} p_t(s'|0,a) &= \mathbb{P}(aW_t = s') \\ &= \begin{cases} 1 & \text{if } a=0, s'=0 \\ \mathbb{P}(W_t=0) & \text{if } a=1, s'=0 \\ \mathbb{P}(W_t=1) & \text{if } a=1, s'=1 \\ 0 & \text{otherwise} \end{cases} \\ &= \begin{cases} 1 & \text{if } a=0, s'=0 \\ 1-p & \text{if } a=1, s'=0 \\ p & \text{if } a=1, s'=1 \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

ii). $s = n$.
The queue is full so no-one can join the queue (ie $\mathbb{P}(W_t = 0) = 1$).

$$\begin{aligned} p_t(s'|n,a) &= \mathbb{P}(n - Z_t = s') \\ &= \begin{cases} \mathbb{P}(Z_t=0) & \text{if } s'=n \\ \mathbb{P}(Z_t=1) & \text{if } s'=n-1 \\ 0 & \text{if } s'<n-1 \end{cases} \\ &= \begin{cases} 1-q & \text{if } s'=n \\ q & \text{if } s'=n-1 \\ 0 & \text{if } s'<n-1 \end{cases} \end{aligned}$$

It is noteworthy that these results are independent of the action taken a.

iii). $s' = s,\ s \notin \{0, n\}$.
The queue has not changed size. Thus, either: no movements have occurred; or, one customer joined the queue and another customer left the queue. The second scenario is only possible if the agent allows customers to join the queue (ie $a = 1$).

$$\begin{aligned} p_t(s|s,a) &= \mathbb{P}(s + aW_t - Z_t = s) \\ &= \mathbb{P}(aW_t - Z_t = 0) \\ &= \begin{cases} \mathbb{P}(Z_t=0) & \text{if } a=0 \\ \mathbb{P}(W_t=Z_t) & \text{if } a=1 \end{cases} \\ &= \begin{cases} 1-q & \text{if } a=0 \\ pq+(1-p)(1-q) & \text{if } a=1 \end{cases} \end{aligned}$$

iv). $s' = s + 1,\ s \notin \{0, n\}$.
The queue has become longer.

$$\begin{aligned} p_t(s+1|s,a) &= \mathbb{P}(s + aW_t - Z_t = s+1) \\ &= \begin{cases} \mathbb{P}(-Z_t=1) & \text{if } a=0 \\ \mathbb{P}(W_t-Z_t=1) & \text{if } a=1 \end{cases} \\ &= \begin{cases} 0 & \text{if } a=0 \\ \mathbb{P}(W_t=1)\mathbb{P}(Z_t=0) & \text{if } a=1 \end{cases} \\ &= \begin{cases} 0 & \text{if } a=0 \\ p(1-q) & \text{if } a=1 \end{cases} \end{aligned}$$

v). $s' = s-1,\ s \notin \{0,n\}$.
The queue has become shorter.

$$\begin{array}{rcl}p_t(s-1|s,a) &=& \mathbb{P}(s+aW_t-Z_t=s-1)\\ &=& \mathbb{P}(aW_t-Z_t=-1)\\ &=& \begin{cases}\mathbb{P}(Z_t=1) & \text{if } a=0\\ \mathbb{P}(W_t-Z_t=-1) & \text{if } a=1\end{cases}\\ &=& \begin{cases}q & \text{if } a=0\\ \mathbb{P}(W_t=0,Z_t=1) & \text{if } a=1\end{cases}\\ &=& \begin{cases}q & \text{if } a=0\\ (1-p)q & \text{if } a=1\end{cases}\end{array}$$

vi). *All other cases.*
All other cases require either the number of people in the queue to become negative or to change by more than 1 in a single time-step, both of these are impossible so have 0 probability.

$$p_t(s'|0,a)=0$$

- *Immediate Costs*

  For costs in time-period t we always have to pay cX_t for the length of the queue. Additionally, if a customer is rejected C is payed as well. Mathematically, a customer is rejected if $W_t=1$ and $Y_t=0$. We can summarise the total cost G_t incurred in time-period t as

  $$G_t := g_t(W_t,X_t,Y_t) = cX_t + \mathbb{1}\{W_t=1,Y_t=0\}\cdot C$$

  This assumes that $Y_t=0$ if the queue is already full.

- *Equivalent Objective*

  The objective of this problem is to minimise expected total cost

  $$\mathbb{E}^\pi\left[\sum_{t=0}^{N-1}G_t\right] = \mathbb{E}^\pi\left[\sum_{t=0}^{N-1}g_t(W_t,X_t,Y_t)\right]$$

  This expectations depends upon the number of customers wishing to join the queue $Y_0,\dots,Y_{N-1}$ which is independent of a policy π so does not fit within the framework a *Markov Decision Problem.* Thus I shall transform the expect total cost.

  $$\begin{array}{rcl}\mathbb{E}^\pi\left[\sum_{t=0}^{N-1}G_t\right] &=& \sum_{t=0}^{N-1}\mathbb{E}^\pi[G_t]\\ &=& \sum_{t=0}^{N-1}\mathbb{E}^\pi\left[\mathbb{E}^\pi[G_t|X_t,Y_t]\right] \text{ by Tower Property}\\ &=& \sum_{t=0}^{N-1}\mathbb{E}^\pi\left[\mathbb{E}^\pi[g_t(W_t,X_t,Y_t)|X_t,Y_t]\right]\end{array}$$

  Define reward functions $r_t(s,a)$ and $r_N(s)$

  $$\begin{array}{rcl}r_N(s) &=& 0\\ r_t(s,a) &=& -\mathbb{E}^\pi[g_t(W_t,X_t,Y_t)|X_t=s,Y_t=a]\\ &=& -\mathbb{E}^\pi[g_t(W_t,s,a)|X_t=s,Y_t=a]\\ &=& -\mathbb{E}^\pi[g_t(W_t,s,a)]\\ &=& -p\cdot g_t(1,s,a)-(1-p)\cdot g_t(0,s,a)\\ &=& -p(cs+\mathbb{1}\{a=0\}C)-(1-p)cs\\ &=& -cs-\mathbb{1}\{a=0\}\cdot pC\end{array}$$

Using these reward functions the total expected reward can be rephrased

$$-\mathbb{E}^\pi\left[r_N(X_n)+\sum_{t=0}^{N-1}r_t(X_t,Y_t)\right] \tag{1}$$

The equivalent objective is to find a policy $\pi\in HR(T)$ which maximises (1).

Question 3) (b)
Solve the formulated Markov decision problem for the following case

- $N=2,n=3$.
- $p=\dfrac{1}{2},q=\dfrac{1}{4}$.
- $C=2,c=1$.

Answer 3) (b)
From the markov decision problem formulated in `3) (a)` and the given conditions we can state the following properties of the system

$$\begin{array}{rcl}T&=&\{0,1\}\\ S&=&\{0,1,2,3\}\\ A&=&\{0,1\}\\ A(s)&=&\begin{cases}\{0,1\}&\text{if }s\in\{0,1,2\}\\ \{0\}&\text{if }s=3\end{cases}\end{array}$$

The transition probabilities $p_t(s'|s,a)$ are defined in the tables below, separated by what action a is taken.

$$p_t(s'|s,0)=\begin{array}{c|cccc}\text{s}\backslash\text{s'}&0&1&2&3\\\hline 0&1&0&0&0\\ 1&\text{q}&\text{1-q}&0&0\\ 2&0&\text{q}&\text{1-q}&0\\ 3&0&0&\text{q}&\text{1-q}\end{array}=\begin{array}{c|cccc}\text{s}\backslash\text{s'}&0&1&2&3\\\hline 0&1&0&0&0\\ 1&1/4&3/4&0&0\\ 2&0&1/4&3/4&0\\ 3&0&0&1/4&3/4\end{array}$$

$$p_t(s'|s,1)=\begin{array}{c|cccc}\text{s}\backslash\text{s'}&0&1&2&3\\\hline 0&\text{1-p}&\text{p}&0&0\\ 1&\text{q(1-p)}&\text{pq+(1-p)(1-q)}&\text{p(1-q)}&0\\ 2&0&\text{q(1-p)}&\text{pq+(1-p)(1-q)}&\text{p(1-q)}\\ 3&0&0&\text{q}&\text{1-q}\end{array}=\begin{array}{c|cccc}\text{s}\backslash\text{s'}&0&1&2&3\\\hline 0&1/2&1/2&0&0\\ 1&1/4&1/8&3/8&0\\ 2&0&1/4&1/8&3/8\\ 3&0&0&1/4&3/4\end{array}$$

The terminal cost value is $r_2(s)=0$. The cost function values $r_t(s,a)$ are given in the table below and are the same $\forall\ t\in\{0,1\}$.

$$r_t(s,a)=\begin{array}{c|cc}\text{s}\backslash\text{a}&0&1\\\hline 0&-pC&0\\ 1&-c-pC&-c\\ 2&-2c-pC&-2c\\ 3&-3c-pC&-3c\end{array}=\begin{array}{c|cc}\text{s}\backslash\text{a}&0&1\\\hline 0&-1&0\\ 1&-2&-1\\ 2&-3&-2\\ 3&-4&-3\end{array}$$

To find the optimal policy π^* we use the *dynamic programming algorithm* which is defined as

$$\begin{array}{rcl}w_t^*(s,a)&:=&r_t(s,a)+\sum_{s'\in S}u_{t+1}^*(s')p_t(s'|s,a)\\ u_t^*(s)&=&\max_{a\in A(s)}(w_t^*)\\ d_t^*(s)&=&\text{argmax}_{a\in A(s)}(w_t^*)\end{array}$$

where $u_2^*(s) := r_2(s) = 0 \ \forall \ s \in S$. Specifically, we need to determine $u_t^*(s),\ d_t^*(s)$ for all states s in each time-period $t \in \{1, 0\}$.

- *Time-Period* $t = 1$.

  In this time-period

$$\begin{array}{rcl} w_1^*(s,a) & = & r_1(s,a) + \sum_{s' \in \{0,1,2,3\}} u_2^*(s) p_t(s'|s,a) \\ & = & r_1(s,a) \text{ since } u_2^*(s) = 0 \ \forall \ s \in S \end{array}$$

  This gives the following table of values for $w_1^*(s,a)$

$w_1^*(s,a) =$

s\a	0	1
0	-1	0
1	-2	-1
2	-3	-2
3	-4	-3

  This table shows that taking action $a = 1$ produces the greatest expected reward in all states s. This is summarised in the following table for $u_1^*(s), d_1^*(s)$

s	$u_1^*(s)$	$d_1^*(s)$
0	0	1
1	-1	1
2	-2	1
3	-3	1

- *Time-Period* $t = 0$.

  In this time-period

$$\begin{array}{rcl} w_0^*(s,a) & = & r_0(s,a) + \sum_{s' \in \{0,1,2,3\}} u_1^*(s) p_t(s'|s,a) \\ & = & r_0(s,a) - p_t(1|s,a) - 2 \cdot p_t(2|s,a) - 3 \cdot p_t(3|s,a) \end{array}$$

  This gives the following table of values for $w_0^*(s,a)$

$w_1^*(s,a) =$

s\a	0	1
0	-1+0+0+0	0-1/2+0+0
1	-2-3/4+0+0	-1-1/8-6/8+0+0
2	-3-1/4-6/4+0	-2-1/4-2/8-9/8
3	-4+0-2/4-9/4	-3+0-2/4-9/4

$=$

s\a	0	1
0	-1	-1/2
1	-11/4	-15/8
2	-19/4	-29/8
3	-27/4	-23/4

  Again, this table shows that taking action $a = 1$ produces the greatest expected reward in all states s. This is summarised in the following table for $u_1^*(s), d_1^*(s)$

s	$u_0^*(s)$	$d_0^*(s)$
0	-1/2	1
1	-15/8	1
2	-29/8	1
3	-23/4	1

The optimal policy is

$$\pi^* = (d_0^*(s), d_1^*(s)) = (1, 1)\ \forall\ s$$

The optimal value function is

$$u_0^*(s) = \begin{cases} -1/2 & \text{if } s = 0 \\ -15/8 & \text{if } s = 1 \\ -29/8 & \text{if } s = 2 \\ -23/4 & \text{if } s = 3 \end{cases}$$

Question 4)
Consider the following machine maintenance problem.
A machine is operated over time-periods $0, \dots, N-1$. The machine can be in one of states $1, \dots, M$. For $s \in \{1, \dots, M-1\}$, we assume that the condition of the machine is better in state s than in state $s+1$ (ie 1 is best condition, M is worst condition).

At the start of each period, the state of the machine is known and one of the following actions is taken

- The machine is allowed to operate in the current state for one more period.
- The machine is repaired: If the machine is in state $s \in \{2, \dots, M\}$ at the beginning of the time-period, then its state can immediately be restored to (a better) state in $s' \in \{1, \dots, s-1\}$ at cost $c_r(s, s')$.

The cost of operating the machine in state $s \in \{1, \dots, M\}$ during a time-period is $c_o(s)$. (If the machine is in state s after the action taken at the beginning of a time-period, cost $c_0(s)$ is incurred during that period). We assume worst states cost more

$$c_o(1) < \dots < c_0(M)$$

During each time-period, the state of the machine can become worse or it may stay unchanged. The machine changes its state randomly:

- If $s \in \{1, \dots, M-1\}$ is the state after the action taken at the beginning of the time-period, the machine will be in state $s' \in \{s, \dots, M\}$ at the end of this period with probability $p(s'|s)$ where $p(s'|s) \geq 0$ and $\sum_{s'=s}^{M} p(s'|s) = 1$.
- If the machine is in state M after the action taken at the beginning of the time-period, the machine will remain in state M at the end of this period with probability one.

The problem is to decide, at the beginning of each time-period, what action to take so as to expect total cost of operating the machine over period $0, \dots, N-1$ is minimal. The decision should be based on the state of the machine.

Question 4) (a)
Formulate the described machine maintenance problem as a finite horizon Markov decision problem. More precisely, state the state-space S, action-space A, time-horizon T, transition probabilities $p_t(s'|s,a)$ and immediate rewards $r_t(s,a)$.

Answer 4) (a)
Let X_t be the system state at the start of time-period X_t and Y_t be the action the agent takes at the start of time-period Y_t.

- *Decision Epochs* - Start of each time-period.
- *Time-Horizon* - $T = \{0, \dots, N-1\}$.
- *Agent Actions.*

  At the start of each turn the agent can either repair the machine to a specified state $s \in \{1, \dots, M-1\}$ or not repair. This can be encoded into Y_t as

$$Y_t = \begin{cases} 0 & \text{if machine is \underline{not} repaired} \\ s & \text{if machine is repaired to state } s \end{cases}$$

- *Action-Space.*

  Given the specification of Y_t the action-space is $A := \{0, \dots, M-1\}$.

- *System-States.*

  Knowing the state of the machine at the beginning of a time-period is sufficient for the agent to determine which action to take. This knowledge can be encoded into X_t by setting X_t to be the state of the machine at the start of period t.

- *State-Space.*

  By the definition of X_t, the state-space is equivalent to the set of states the machine can take $S = \{1, \dots, M\}$.

- *Admissible Action-Space.*

  The agent can always choose not to repair the machine, but if they choose to repair the machine they can only repair it to a better state. This gives admissible action-spaces

$$A(s) = \{0, \dots, s-1\},\ s \in S$$

- *Immediate Costs.*

  In each time-period there is always a cost for running the machine $c_o(s)$, where s is the state after the agent has taken their action. There is an additional cost $c_r(s, s')$ if the agent chooses to repair the machine from state s to state s' (ie if $Y_t = s' \neq 0$).

  Let Z_t be the state of the machine after the agent takes their action. If the agent does not repair (ie $Y_t = 0$), then the state of the machine after the agent's action is the same as before (ie $Z_t = X_t$). However, if the agent chooses to repair the machine (ie $Y_t = s' \in A(X_t) \setminus \{0\}$) then the machine's state after the action will be whatever state the agent chose to repair the machine to (ie $Z_t = Y_t$). This gives the following definition for Z_t

$$Z_t = \begin{cases} X_t & \text{if } Y_t = 0 \\ Y_t & \text{if } Y_t \in A(X_t) \setminus \{0\} \end{cases}$$

  We can define the total cost in time-period t as

$$\begin{array}{rcl} G_t & := & c_o(Z_t) + \mathbb{1}\{Y_t \in A(X_t) \setminus \{0\}\} c_r(X_t, Z_t) \\ & = & \begin{cases} c_o(X_t) & \text{if } Y_t = 0 \\ c_o(Y_t) + c_r(X_t, Y_t) & \text{if } Y_t \in A(s) \setminus \{0\} \end{cases} \end{array}$$

  For clarity of dependency, define the function

$$g_t(X_t, Y_t) := \begin{cases} c_o(X_t) & \text{if } Y_t = 0 \\ c_o(Y_t) + c_r(X_t, Y_t) & \text{if } Y_t \in A(s) \setminus \{0\} \end{cases}$$

- *Transition Probabilities.*

$$\begin{array}{rcl} p_t(s'|s, a) & := & \mathbb{P}^\pi(X_{t+1} = s' | X_t = s, Y_t = a) \\ & = & \begin{cases} p(s'|s) & \text{if } a = 0 \\ p(s'|a) & \text{if } a \in A(s) \setminus \{0\} \end{cases} \end{array}$$

  where $p(\cdot|\cdot)$ are the probabilities of the machine degrading as defined in the question.

- *Equivalent Objective* The objective of this problem is to minimises the total expected cost

$$\begin{aligned}
\mathbb{E}^\pi\left[\sum_{t=0}^{N-1} G_t\right] &= \mathbb{E}^\pi\left[\sum_{t=0}^{N-1} g_t(X_t, Y_t)\right] \\
&= \sum_{t=0}^{N-1} \mathbb{E}^\pi[g_t(X_t, Y_t)] \\
&= \sum_{t=0}^{N-1} \mathbb{E}^\pi\left[\mathbb{E}^\pi[g_t(X_t, Y_t)]\middle|X_t = s, Y_t = a\right] \text{ by Tower Property.}
\end{aligned}$$

Minimisation does not fit within the Markov decision problem framework, so I shall make this a maximisation problem. Consider the following reward functions

$$\begin{aligned}
r_N(s) &:= 0 \\
r_t(s,a) &:= -\mathbb{E}^\pi[g_t(X_t, Y_t)|X_t = s, Y_t = a] \\
&= -\mathbb{E}^\pi[g_t(s,a)] \\
&= \begin{cases} -c_o(s) & \text{if } a = 0 \\ -c_o(s) - c_r(s,a) & \text{if } a \in \{1, \dots, s-1\} \end{cases}
\end{aligned}$$

Using these definitions, our objective can be restated as wishing to maximise the following

$$-\mathbb{E}^\pi\left[r_N(s) + \sum_{t=0}^{N-1} r_t(s,a)\right]$$

Question 4) (b)
Find the optimal policy for the Markov decision problem formulated it `4) (a)` under the following conditions

- $M = 3, N = 2.$
- $p(1|1) = \frac{1}{2},\ p(2|1) = \frac{1}{4},\ p(3|1) = \frac{1}{4},\ p(2|2) = \frac{1}{4},\ p(3|2) = \frac{3}{4}.$
- $c_r(2,1) = 1,\ c_r(3,2) = 2,\ c_r(3,1) = 4.$
- $c_o(1) = 1,\ c_o(2) = 2,\ c_o(3) = 5.$

Answer 4) (b)
The system described in the question has the follow probabilities for degradation and repair costs

$p(s'|s) =$

$s\backslash s'$	1	2	3
1	1/2	1/4	1/4
2	ND	1/4	3/4
3	ND	ND	1

$c_r(s,s') =$

$s'\backslash s$	1	2	3
1	ND	1	4
2	ND	ND	2
3	ND	ND	ND

where "ND" denotes "Not Defined", this is due to the actions required for these values to be needed are not admissible in these states.

From the markov decision problem formulated in `4) (a)` and the given conditions, we can state the following properties of the system

$$\begin{array}{rcl} T &=& \{0,1\} \\ S &=& \{1,2,3\} \\ A &=& \{0,1,2\} \\ A(s) &=& \{0,\dots,s-1\} \end{array}$$

The transition probabilities $p_t(s'|s,a)$ are defined in the tables below, separated by what action a is taken.

$$p_t(s'|s,0) = \begin{array}{c|ccc} s\backslash s' & 1 & 2 & 3 \\ \hline 1 & 1/2 & 1/4 & 1/4 \\ 2 & 0 & 1/4 & 3/4 \\ 3 & 0 & 0 & 1 \end{array}$$

$$p_t(s'|s,1) = \begin{array}{c|ccc} s\backslash s' & 1 & 2 & 3 \\ \hline 1 & \text{NA} & \text{NA} & \text{NA} \\ 2 & 1/2 & 1/4 & 1/4 \\ 3 & 1/2 & 1/4 & 1/4 \end{array}$$

$$p_t(s'|s,2) = \begin{array}{c|ccc} s\backslash s' & 1 & 2 & 3 \\ \hline 1 & \text{NA} & \text{NA} & \text{NA} \\ 2 & \text{NA} & \text{NA} & \text{NA} \\ 3 & 0 & 1/4 & 3/4 \end{array}$$

where "NA" denotes that action a is not an admissible action for that state s (ie $a \notin A(s)$).

The terminal cost value is $r_2(s) = 0$. The cost function values $r_t(s,a)$ are given in the table below and are the same for all $t \in \{0,1\}$.

$$r_t(s,a) = \begin{array}{c|ccc} s\backslash a & 0 & 1 & 2 \\ \hline 1 & c_o(1) & \text{NA} & \text{NA} \\ 2 & c_o(2) & c_o(2)+c_r(2,1) & \text{NA} \\ 3 & c_o(3) & c_o(3)+c_r(3,1) & c_o(3)+c_r(3,2) \end{array} = \begin{array}{c|ccc} s\backslash a & 0 & 1 & 2 \\ \hline 1 & -1 & \text{NA} & \text{NA} \\ 2 & -2 & -3 & \text{NA} \\ 3 & -5 & -9 & -7 \end{array}$$

To find the optimal policy π^* we use the *dynamic programming algorithm* which is defined as

$$\begin{array}{rcl} w_t^*(s,a) &:=& r_t(s,a) + \sum_{s' \in S} u_{t+1}^*(s') p_t(s'|s,a) \\ u_t^*(s) &=& \max_{a \in A(s)} (w_t^*) \\ d_t^*(s) &=& \text{argmax}_{a \in A(s)} (w_t^*) \end{array}$$

where $u_2^*(s) := r_2(s) = 0 \ \forall \ s \in S$. Specifically, we need to determine $u_t^*(s)$, $d_t^*(s)$ for all states s in each time-period $t \in \{1,0\}$.

- *Time-Period* $t = 1$.

  In this time-period

$$\begin{array}{rcl} w_1^*(s,a) &=& r_1(s,a) + \sum_{s' \in \{1,2,3\}} u_2^* p_t(s'|s,a) \\ &=& r_1(s,a) \text{ since } u_2^*(s) = 0 \ \forall \ s \in S \end{array}$$

  This gives the following table of values for $w_1^*(s,a)$

$$w_1^*(s,a)=\begin{array}{c|ccc} s\backslash a & 0 & 1 & 2 \\ \hline 1 & -1 & \text{NA} & \text{NA} \\ 2 & -2 & -3 & \text{NA} \\ 3 & -5 & -9 & -7 \end{array}$$

In all cases taking action $a=0$ (ie not repairing) yields the best results. This is summarised in the following table of results for $u_1^*(s), d_1^*(s)$.

s	$u_1^*(s)$	$d_1^*(s)$
1	-1	0
2	-2	0
3	-5	0

- *Time-Period* $t=0$.

  In this time-period

$$\begin{aligned} w_0^*(s,a) &= r_0(s,a)+\sum_{s'\in\{1,2,3\}} u_1^* p_t(s'|s,a) \\ &= r_0(s,a)-p_t(1|s,a)-2p_t(2|s,a)-5p_t(3|s,a) \end{aligned}$$

  This gives the following table of values for $w_0^*(s,a)$

$$w_0^*(s,a)=\begin{array}{c|ccc} s\backslash a & 0 & 1 & 2 \\ \hline 1 & \text{-1-1/2-2/4-5/4} & \text{NA} & \text{NA} \\ 2 & \text{-2+0-2/4-15/4} & \text{-3-1/2-2/4-5/4} & \text{NA} \\ 3 & \text{-5+0+0-5} & \text{-9-1/2-2/4-5/4} & \text{-7+0-2/4-15/4} \end{array} = \begin{array}{c|ccc} s\backslash a & 0 & 1 & 2 \\ \hline 1 & \text{-13/4} & \text{NA} & \text{NA} \\ 2 & \text{-25/4} & \text{-21/4} & \text{NA} \\ 3 & \text{-10} & \text{-45/4} & \text{-45/4} \end{array}$$

  This time the optimal action is different for different states. I summarise the optimal actions in the table below

s	$u_1^*(s)$	$d_1^*(s)$
1	-13/4	0
2	-21/4	1
3	-10	0

The optimal strategy π^* is

$$\pi^* := (d_1^*(s), d_2^*(s))$$

and the optimal value function is

$$u_0^*(s)=\begin{cases} -13/4 & \text{if } s=0 \\ -21/4 & \text{if } s=1 \\ -10 & \text{if } s=2 \end{cases}$$