1 (Larger) Machine replacement problem

Consider a production machine on a factory floor that can be in three different conditions: perfect, worn and broken. When operating the machine, it has a probability θ of degrading one stage (that is, going from perfect to worn, or from worn to broken). The factory owner can choose to replace the machine at a cost R. If it is broken, then he acquires a cost c for not being able to produce new products, and if it is worn, he acquires a cost $c/2$ for producing imperfect items (at each time-step). He wants to find an optimal policy for T time-steps that minimizes his expenses. Assume that the cost at the end of the horizon is zero, regardless of state.

a) Model the problem as an MDP, then answer the following question: What is the correct transition matrix? *Note:* The states are indexed as perfect (1), worn (2) and broken (3).

$$P(\text{replace}) = \begin{bmatrix} \_\_\_\_ & \_\_\_\_ & \_\_\_\_ \\ \_\_\_\_ & \_\_\_\_ & \_\_\_\_ \\ \_\_\_\_ & \_\_\_\_ & \_\_\_\_ \end{bmatrix} \text{ and } P(\text{continue}) = \begin{bmatrix} \_\_\_\_ & \_\_\_\_ & \_\_\_\_ \\ \_\_\_\_ & \_\_\_\_ & \_\_\_\_ \\ \_\_\_\_ & \_\_\_\_ & \_\_\_\_ \end{bmatrix}.$$

b) For $\theta = 0.5$, $R = 8$, $c = 6$, $T = 2$, solve the MDP *by hand*. That is, compute the optimal cost-to-go and the optimal policy. Then answer the following questions:

- $u_0^*(\text{Worn}) =$

  _____

- $a_0^*(\text{Broken}) =$

  _____

Solution:

a) We use the formalism introduced in Part 2 of the course, and in particular, in each state (even s_T), an action will be selected. Alternative representations can be proposed, e.g. having a terminal state where no action is selected. We can model the problem as the following finite time-horizon MDP:

- Time horizon: T
- State space: $\mathcal{S} = \{$ P (perfect), W (Worn), B (Broken) $\}$
- Actions: $\mathcal{A} = \{$ Continue (C), Replace (R) $\}$
- Rewards: the reward function is stationary is given by: $r(P, C) = 0$, $r(W, C) = -c/2$, $r(B, C) = -c$, $r(\cdot, R) = -R$
- Transitions: The transition probabilities are also stationary, with:

$$P(\cdot|\cdot, R) = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix} \text{ and } P(\cdot|\cdot, C) = \begin{bmatrix} 1-\theta & \theta & 0 \\ 0 & 1-\theta & \theta \\ 0 & 0 & 1 \end{bmatrix}.$$

b) To comply to the formalism given in the course, we compute u_1^*(Worn) (called u_0^*(Worn) in the text of the home-work). We use Bellman equation: for $s \in \mathcal{S}$,

$$u_{t-1}^*(s) = \max_{a=C \text{ or } R} \{r(s, a) + \sum_{j \in \mathcal{S}} P(j|s, a) u_t^*(j)\}.$$

We start by computing the terminal reward associated to the best action:
$u_2^*(P) = \max\{0, -R\} = 0$ (C is optimal)
$u_2^*(W) = \max\{-\frac{c}{2}, -R\} = -\frac{c}{2} = -3$ (C is optimal)
$u_2^*(B) = \max\{-c, -R\} = -c = -6$ (C is optimal)

We can now compute u_1^*:
$u_1^*(P) = \max\{\theta u_2^*(W) + (1-\theta) u_2^*(P), \theta u_2^*(W) + (1-\theta) u_2^*(P) - R\} = -1.5$ (C is optimal)
$u_1^*(W) = \max\{-\frac{c}{2} + \theta u_2^*(B) + (1-\theta) u_2^*(W), \theta u_2^*(W) + (1-\theta) u_2^*(P) - R\} = -7.5$ (C is optimal)
$u_1^*(B) = \max\{-c + u_2^*(B), \theta u_2^*(W) + (1-\theta) u_2^*(P) - R\} = -9.5$ (R is optimal)

We deduce that

$$u_1^*(W) = -7.5, \quad a_1^*(B) = R$$

2 Optimal Stopping

You observe a fair coin being tossed T times. You may stop observing at any time, and when you do you receive as a reward the proportion of heads observed. For example, if the first toss is head, you should stop immediately. Your problem is to identify a stopping rule maximizing the average reward.

a) Model the problem as an MDP. How many states will you use? _____
Justify your answer and write Bellman's equations.

b) Establish by induction one of the following statement. Which one is true? _____
Let $V_t(n)$ denote the maximal average reward if after t tosses, we got n heads.
(A) For all t and n, $V_t(n+1) \geq V_t(n)$
(B) For all t and n, $V_t(n+1) \leq V_t(n)$
(C) For all t and n, $V_t(n+1) = V_t(n)$

c)[*1] One of the following policies is optimal. Which one? Justify your choice._____
(A) After the second toss, stop only if the number of heads reaches $T/2$
(B) Never stop, except when the first toss is head
(C) After t tosses and n observed heads, stop if and only if $n > \frac{t}{2}$

d) The coin is biased, with an unknown bias. Propose an off-policy RL algorithm converging to the optimal policy. Your algorithm should work with one of the following behavior policies. Which one? _____
(A) After t tosses and n observed heads, stop if and only if $n \geq \lfloor \frac{t}{2} \rfloor + 1$
(B) Never stop, i.e., always select the same action.

[1] A difficult question – not qualifying to pass the HW.

Solution.

a)

The problem can be modelled as a finite time horizon MDP as follows.

- State space: $\mathcal{S} = \{\emptyset\} \cup \{(t,n) \in \mathbb{N}^2 : n \leq t, t \leq T\}$, initial state $(0,0)$ and in this state we are forced to continue. $\emptyset$ is the state reached after we decide to stop.
- Actions: Continue (C), Stop (S)
- Rewards: Terminal: $r_T((T,n),\cdot) = \frac{n}{T}$ (you are forced to stop at T). Non-terminal: For $t < T$, $r((t,n),S) = \frac{n}{t}$ and $r((t,n),C) = 0 = r(\emptyset,\cdot)$.
- Transitions: For $t < T$, $p((t+1,n)|(t,n),C) = \frac{1}{2} = p((t+1,n+1)|(t,n),C)$ (unbiased coin), $p(\emptyset|(t,n),S) = 1 = p(\emptyset|\emptyset,\cdot)$.
- Time-horizon and objective: Finite-horizon $T < \infty$, $\mathbb{E}\{\sum_{t=0}^{T} r(s_t,a_t)\}$.

The number of states used is hence $\sum_{t=0}^{T}(t+1) = \frac{(T+1)(T+2)}{2}$.

Bellman's equation: $V^\star(T,n) = \frac{n}{T}$, and for all $t < T$:

$$V^\star(t,n) = \max\{\frac{n}{t}, \frac{1}{2}(V^\star(t+1,n) + V^\star(t+1,n+1))\}$$

b) Answer: (A)
By backward induction on t, we prove that $V_t(n)$ is increasing in n, for all t.
For $t = T$, the result holds because $V_T(n) = n/T$ is increasing in n.
Assume that the result holds at time $t+1$. Then:

$$V_t(n) = \max\{\frac{n}{t}, \frac{1}{2}(V_{t+1}(n) + V_{t+1}(n+1))\}$$

and hence, V_t is the maximum of two increasing functions. Thus, V_t is increasing.

c) Answer: (C)

The solution consists in proving that (A), (B) present suboptimal policies. This can be done easily.

d) Answer: (B)

Unlike what we said during the course where we emphasise that in general it is important to explore all actions, to know their rewards, here the reward function depends on the state only. More precisely, in state (s,n) we know the reward we would get if we stopped (n/t). So knowing the state gives the information about the rewards on the "stop" action without actually exploring this action. Hence by always selecting the "continue" action is the right thing to do!