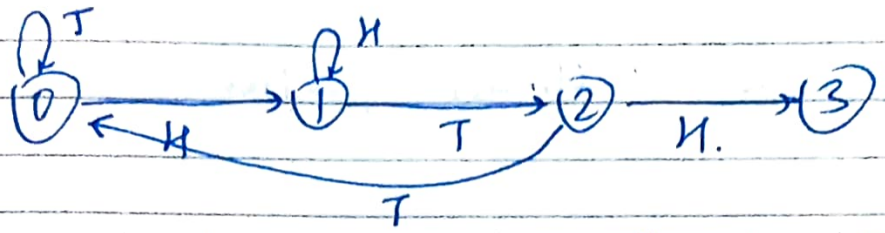
Date : / /
Page No. 1

Question 1

State # 0 :→ No part of pattern has been matched.
State 1 :- The first 'H' has been matched
State 2 :- The first 'H' followed by 'T' has been matched.
State 3 :- The pattern 'HTH' has been matched



Transition matrix

$$P = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{cccc} 0 & 1 & 2 & 3 \end{array} \\ \begin{bmatrix} 1/2 & 1/2 & 0 & 0 \\ 0 & 1/2 & 1/2 & 0 \\ 1/2 & 0 & 0 & 1/2 \\ 0 & 0 & 0 & 1 \end{bmatrix} \end{array}$$

Absorbing state is State #3

~~States~~ We can write the states of MRP as $\langle S, P, R, \gamma \rangle$ as follows

$S = \{ \emptyset, 'H', 'HT', 'HTH' \}$

$$P = \begin{array}{c} \\ \emptyset \\ H \\ HT \\ HTH \end{array} \begin{array}{c} \begin{array}{cccc} \emptyset & H & HT & HTH \end{array} \\ \begin{bmatrix} 1/2 & 1/2 & 0 & 0 \\ 0 & 1/2 & 1/2 & 0 \\ 1/2 & 0 & 0 & 1/2 \\ 0 & 0 & 0 & 1 \end{bmatrix} \end{array}$$

I set a reward for all states = -1 except absorbing state 'HTH'.

i.e

$R(s) = -1$ for $s \in \{\emptyset, 'H', 'HT'\}$ for every extra toss.

$R(s) = 0$ for $s = 'HTH'$.

The discount factor I take is $\gamma = 1$.

Now as per the Bellman equation

$$V(s) = E(r_{t+1} | s_t = s) + \gamma \sum_{s' \in S} P_{ss'} V(s')$$

applying it to all states.

$$V(\emptyset) = -1 + \tfrac{1}{2} V('H') + \tfrac{1}{2} V(\emptyset)$$

$$V('H') = -1 + \tfrac{1}{2} V('H') + \tfrac{1}{2} V(\emptyset)$$

$$V('HT') = -1 + \tfrac{1}{2} V(\emptyset) + \tfrac{1}{2} V('HTH')$$

$$V('HTH') = 0. \quad \{\text{as the 'HTH' is the absorbing state}\}$$

Solving the above system of equations.

$$V('HT') = -1 + \tfrac{1}{2} V(\emptyset) \rightarrow (1)$$

$$V('H') = -1 + \tfrac{1}{2} V('H') + \tfrac{1}{2} V('HT')$$

$$\Rightarrow V('H') = -2 + V('HT') \rightarrow (2)$$

$V(0) = -2 + V(1)$

$= -2 + (-2 + V(2))$

$\Rightarrow V(0) = -4 + V(2)$

using ② equation ①

$V(0) = -4 + (-1 + \frac{1}{2} V(0))$

$\Rightarrow V(0) = -5 + \frac{1}{2} V(0)$

$\Rightarrow V(0) = -10$

Now as we had taken Reward $R(s) = -1$ for every turns with -ve value being acting as a penalization factor.

Also $V(0)$: Signifies the expected number of turns required to reach goal state i.e value from state 0.

Hence No of Turns required to reach goal State = 10 {as -1 is awarded for every extra turns}. Therefore dividing -10 by -1 i.e $R(s)$.

Question 2

2(a)

(i) State Space :- defines the number of machines that are currently working.

$S : \{0, 1, 2, 3, \ldots N\}$

where N = total machines present in the production facility.

Any state has m machines working where $0 \leq m \leq N$. as per the problem statement.

(ii) Action Space :- Actions available at each state m are

→ Repair :- call the Repairman to repair all machines at the cost of $-\frac{N}{2}$ $.

→ No Repair :- Do nothing & continue with current number of working machines.

Therefore $A = \{$ Repair, No Repair $\}$.

(iii) Transition Probabilities $P(n/m, a)$

n = no of working machines of next day

m = no of working machines of current day

a = under the action a. i.e Repair/No repair

Date: / /
Page No. 5

~~P(A = No repair) =~~

P(n/m , a = "No repair") =

	n=0	n=1	n=2	n=3	n=4	n=N-1	n=N
m=0	1/2	1/2	0	0	0	0	0
m=1	1/3	1/3	1/3	0	0	0	0
m=2	1/4	1/4	1/4	1/4	0	0	0
m=3							
m=N	1/N+1	1/N+1	1/N+1				1/N+1

where m = currently correctly working machines
n = ~~no~~ no of machines that would start up correctly next day.

After repairing state always transitions to m = N (all machines working)

P(n/m, "Repair") =

~~P(n=N / m, Repair) = 1~~

i.e

	n=0	n=1	n=2	n=3	n=4	n=N-1	n=N
m=0	0	0	0	0	0	0	1
m=1	0	0	0	0	0	1/2	1/2
m=2	0	0	0	0	1/3	1/3	1/3
m=3							
m=N	1/N+1	1/N+1	1/N+1				1/N+1

(iv) Rewards.

$R(m, a = \text{"No Repair"}) = m$ { since each working machine generates a revenue of 1 $. Hence Reward for state when m machines are working is m.

$R(m, a = \text{"Repair"}) = \frac{-N}{2}$ { cost - lumpsum to repair all incorrectly working machines.

Note:- it is independent of m.

2. (b) I would use undiscounted setting, since the problem statement focusses on maximizing total long term profit, without any indication that future rewards should be less prioritized.

Hence Total Rewards

$$\sum_{t=0}^{\infty} \gamma^t R_t$$

$$= \sum_{t=0}^{\infty} R_t \quad \text{for } \gamma = 1 \text{ i.e undiscounted setting.}$$

(c) Given MDP $\langle S, A, P, R, \gamma \rangle$.

$N = 5$

S: State space S i.e $m \in \{0, 1, 2, 3, 4, 5\}$

P

m	0	1	2	3	4	5
0	0	0	0	0	0	0
1	1/2	1/2	0	0	0	0
2	1/3	1/3	1/3	0	0	0
3	1/4	1/4	1/4	1/4	0	0
4	1/5	1/5	1/5	1/5	1/5	0
5	1/6	1/6	1/6	1/6	1/6	1/6

A = "No Repair" = Action.

$\gamma = 1$ {as undiscounted setting assumed}

$R(m, a = \text{"No Repair"}) = m$.

i.e $R(0) = 0$
$R(1) = 1$
$R(2) = 2$
$R(3) = 3$
$R(4) = 4$
$R(5) = 5$.

using the Bellman ~~equatio~~ evaluation equation

$$V^{\pi} = R^{\pi} + \gamma P^{\pi} V^{\pi}.$$

$V(0) = 0 \quad \rightarrow (1)$ { as no machines are working }

$$V(1) = 1 + \frac{1}{2} V(0) + \frac{1}{2} V(1) \rightarrow (2)$$

$$V(2) = 2 + \frac{1}{3} \{ V(0) + V(1) + V(2) \} \rightarrow (3)$$

$$V(3) = 3 + \frac{1}{4} \{ V(0) + V(1) + V(2) + V(3) \} \rightarrow (4)$$

$$V(4) = 4 + \frac{1}{5} \{ V(0) + V(1) + V(2) + V(3) + V(4) \}$$

$$V(5) = 5 + \frac{1}{6} \{ V(0) + V(1) + V(2) + V(3) + V(4) + V(5) \}$$

Solving (1) & (2) we get

$$V(1) = 1 + \frac{1}{2}(0) + \frac{1}{2} V(1)$$

$$\Rightarrow V(1) = 2 \times 1 = 2 \quad \rightarrow (7)$$

Solving ~~7=4~~ (3) using values of $V(0)$ & $V(1)$

$$\Rightarrow V(2) = 2 + \frac{1}{3} \{ 0 + 2 + V(2) \}$$

$$\Rightarrow \frac{2}{3} V(2) = \frac{6+2}{3}$$

$$\Rightarrow V(2) = 8/2 = 4.$$

Date : / /
Page No. 9

Solving (4) using values of $V(0), V(1), V(2)$

$$V(3) = 3 + \frac{1}{4}\{0 + 2 + 4 + V(3)\}$$

$$\Rightarrow \frac{3}{4}V(3) = \frac{12 + 2 + 4}{4}$$

$$\Rightarrow V(3) = 18/3 = 6.$$

Solving (5) using values of $V(0), V(1), V(2), V(3)$

$$V(4) = 4 + \frac{1}{5}\{0 + 2 + 4 + 6 + V(4)\}$$

$$\frac{4}{5}V(4) = \frac{20 + 2 + 4 + 6}{5}$$

$$\Rightarrow V(4) = 32/4 = 8$$

Solving (5) using values of $V(0), V(1), V(2), V(3)$ & $V(4)$

$$V(5) = 5 + \frac{1}{6}\{0 + 2 + 4 + 6 + 8 + V(5)\}$$

$$\Rightarrow \frac{5}{6}V(5) = \frac{30 + 2 + 4 + 6 + 8}{6}$$

$$V(5) = \frac{50}{5} = 10$$

Hence
$V(0) = 0$
$V(1) = 2$
$V(2) = 4$
$V(3) = 6$
$V(4) = 8$
$V(5) = 10.$

Hence value of the policy for number of working machines $m = 5$ $V(5) = 10$.

2(d) Step 1 :- Initial Policy & it's Evaluation

Let's assume the initial policy π_1 to be a "No repair" policy for all the states

i.e { for states $m=1, m=2, m=3, m=4, m=5$ all has No repair, ~~No repair~~ as the action

i.e $\{0, 1, 2, 3, 4, 5\}$ = State Space

Initial { $m=0$ No repair, $m=1$ No repair, $m=2$ No repair, $m=3$ No repair, $m=4$ No repair, $m=5$ No repair }

and as from previous answer

$V^{\pi}(0) = 0$
$V^{\pi}(1) = 2$
$V^{\pi}(2) = 4$
$V^{\pi}(3) = 6$
$V^{\pi}(4) = 8$
$V^{\pi}(5) = 10.$ for the No repair policy

These above values represent the expected long term reward when starting with m working machines under "No repair" policy.

Step 2 Under the "Repair" action. ~~the~~ & using the Bellman evaluation equation

$$V^{\pi} = R^{\pi} + \gamma P^{\pi} V^{\pi}.$$

Transition matrix under Action = "Repair"

	m=0	m=1	m=2	m=3	m=4	m=5
m=0	0	0	0	0	0	1
m=1	0	0	0	0	1/2	1/2
m=2	0	0	0	1/3	1/3	1/3
m=3	0	0	1/4	1/4	1/4	1/4
m=4	0	1/5	1/5	1/5	1/5	1/5
m=5	1/6	1/6	1/6	1/6	1/6	1/6

Initial policy

Π (No Repair, No Repair, No Repair, No Repair, No Repair, No Repair)

Step 2 Improve the policy for one iteration.

~~$V(s) = \max_a$~~ $V^{\Pi} + R^{\Pi} + \gamma P^{\Pi} V^{\Pi}$.

for Action a = "Repair".

$R(m, \text{"Repair"}) = m - N/2$ where m = no of currently correctly working machines

Discount factor $\gamma = 1$.

Action a: Repair.

V(0) = 0 +

using:

$$V^{\wedge}(s) = R(s, \text{"Repair"}) + \gamma \left[P_{s0} V(s0) + P_{s1} V(s1) + P_{s2} V(s2) + P_{s3} V(s3) + P_{s4} V(s4) + P_{s5} V(s5) \right].$$

$$V(5) = 5 - \frac{5}{2} + 1 * \left[\frac{1}{6} * 0 + \frac{1}{6} * 2 + \frac{1}{6} * 4 + \frac{1}{6} * 6 + \frac{1}{6} * 8 + \frac{1}{6} * 10 \right].$$

$$\rightarrow V(5) = 2.5 + 30/6$$

$$= 7.5.$$

Similarly for

$$V(4) = 4 - \frac{5}{2} + 1 * \left[\frac{1}{5} * 0 + \frac{1}{5} * 2 + \frac{1}{5} * 4 + \frac{1}{5} * 6 + \frac{1}{5} * 8 + \frac{1}{5} * 10 \right]$$

$$= 4 - 2.5 + 30/5 = 7.5.$$

for calculating for V(3)

$$V(3) = 3 - \frac{5}{2} + 1 * \left[\frac{1}{4} \times 0 + \frac{1}{4} \times 0 + \frac{1}{4} * 4 + \frac{1}{4} * 6 + \frac{1}{4} \times 8 + \frac{1}{4} \times 10 \right]$$

$$= 3 - 2.5 + 7 = 7.5.$$

$$V(2) = 2 - \frac{5}{2} + 1 * \left[\frac{1}{3} \times 0 + \frac{1}{3} \times 0 + \frac{1}{3} \times 0 + \frac{1}{3} \times 6 + \frac{1}{3} \times 8 + \frac{1}{3} \times 10 \right]$$

$$= 2 - 2.5 + 8 = 7.5.$$

$$V(1) = 1 - \frac{5}{2} + 1 * \left[\frac{1}{2} \times 0 + \frac{1}{2} \times 0 \frac{1}{2} \times 0 + \frac{1}{2} \times 0 + \frac{1}{2} \times 8 + \frac{1}{2} \times 10 \right]$$

$$= 1 - 2.5 + 9 = 7.5.$$

$$V(0) = 0 - \frac{5}{2} + 1 * \left[0 \times 0 + 0 * 0 + 0 * 0 + 0 * 0 + 0 * 0 + 1 * 10 \right]$$

$$= 10 - 5/2 = 7.5.$$

Since

	Repair V(S, "Repair")	No Repair V(S, "No Repair")	Better Action
S0	7.5	0	Repair
S1	7.5	2	Repair
S2	7.5	4	Repair
S3	7.5	6	Repair.
S4	7.5	8	No Repair
S5	7.5	10.	No Repair.

So the initial policy

π ⎡"No Repair", "No Repair", "No Repair", "No Repair",
"No Repair", "No Repair", "No Repair"⎤

get's changed to

π ⎡"Repair", "Repair", "Repair", "Repair", "No Repair",
"No Repair"⎤

after one iteration

Question 3(a)

(a). At the terminal time step N player must take the reward based on the value of the dice as no further rolling is possible.

Therefore at $n=N$, the value function $V^N(s)$ for each state s is just the reward corresponding to the dice value s.

$$V^N(s) = R(s) = 3s^2+5.$$

$$V^n(s=1) = 3(1)^2+5 = 8$$
$$V^N(s=2) = 3(2)^2+5 = 17.$$
$$V^N(s=3) = 3(3^2)+5 = 32$$
$$V^N(s=4) = 3(4^2)+5 = 53.$$

(b) To compute $Q^{N-1}(s,a)$ for $a =$ "Quit" or $a =$ "continue", we need to evaluate the expected value of taking each action (Quit/continue) at the second to last time step $N-1$ given state s.

$$Q^{N-1}(s, \text{"Quit"}) = R(s) = 3s^2+5.$$

$$Q^{N-1}(s, \text{"continue"}) =$$

$$E[V^N(s')] = \frac{1}{4}\sum_{s=1}^{4} V^{n+1}(s)$$

$$\Rightarrow E[V^N(s')] = \frac{1}{4}\left(V^N(1)+V^N(2)+V^N(3)+V^N(4)\right)$$

$$= \frac{1}{4}(8+17+32+53) = \frac{1}{4}\times 110 = 27.5.$$

Therefore for "continue" action

$$Q^{N-1}(s, \text{"continue"}) = 27.5$$

for $s = 1$

$$Q^{N-1}(1, \text{"Quit"}) = 3(1^2) + 5 = 8$$

$$Q^{N-1}(1, \text{"continue"}) = 27.5$$

for $s = 2$

$$Q^{N-1}(2, \text{"Quit"}) = 3(2^2) + 5 = 17$$

$$Q^{N-1}(2, \text{"continue"}) = 27.5$$

for $s = 3$

$$Q^{N-1}(3, \text{"Quit"}) = 3(3^2) + 5 = 32$$

$$Q^{N-1}(3, \text{"continue"}) = 27.5.$$

for $s = 4$

$$Q^{N-1}(4, \text{"Quit"}) = 3(4^2) + 5 = 53$$

$$Q^{N-1}(4, \text{"continue"}) = 27.5.$$

s	$Q^{N-1}(s, \text{"Quit"})$	$Q^{N-1}(s, \text{"continue"})$
1	8	27.5
2	17	27.5
3	32	27.5
4	53	27.5.

(c) as $V^{N-1}(s) = \max\left(Q^{N-1}(s, \text{"Quit"}), Q^{N-1}(s, \text{"continue"})\right)$

Therefore

for $s=1$ $V^{N-1}(1) = \max(8, 27.5) = 27.5$

$V^{N-1}(2) = \max(17, 27.5) = 27.5$

$V^{N-1}(3) = \max(32, 27.5) = 32$

$V^{N-1}(4) = \max(53, 27.5) = 53$

(d) Applying Bellman equation for $V^{N-1}(s)$

$$V^{N-1}(s) = \max\left(Q^{N-1}(s, \text{"Quit"}), Q^{N-1}(s, \text{"continue"})\right)$$

$= \max(\ldots)$

$Q^{N-1}(s, \text{"Quit"}) = R(s) = 3s^2 + 5$.

$Q^{N-1}(s, \text{"continue"}) = E\left[V_{\pi}^{N}(s')\right] = \frac{1}{4}\sum_{s=1}^{4} V^{N}(s)$

Hence

$$V^{N-1}(s) = \max\left(3s^2+5, \frac{1}{4}\left(V^N(1) + V^N(2) + V^N(3) + V^N(4)\right)\right)$$

(c)

When the player choses to "continue" at time n-1 they roll the dice again.
Since the dice is fair the future state s' will be uniformly distributed over the possible outcomes.
Therefore the expected value of continuing at time n-1 is the average of the action value of continuing at time n.

$$Q_*^{N-1}(s, \text{"continue"}) = \frac{1}{4}\left(Q^N(1, \text{"continue"}) + Q^N(2, \text{"continue"}) + Q^N(3, \text{"continue"}) + Q^N(4, \text{"continue"})\right)$$

Thus $Q^{N-1}(s, \text{"continue"})$ depends on average value of $Q^N(s', \text{"continue"})$.

(d)

$$Q^N(s, \text{"quit"}) = 3s^2 + 5$$

$$Q^N(s, \text{"continue"}) = \frac{1}{4}\sum_{s'=1}^{4} V^{n+1}(s')$$

optimal policy

$$\pi^*(s, n) = \begin{cases} \text{"quit"} \text{ if } Q^n(s, \text{"quit"}) \geq Q^N(s, \text{"continue"}) \\ \text{"continue"} \text{ if } Q^N(s, \text{"quit"}) < Q^N(s, \text{"continue"}) \end{cases}$$

i.e

$$= \begin{cases} \text{"quit"} \text{ if } 3s^2 + 5 \geq \frac{1}{4}\sum_{s'=1}^{4} V^{N+1}(s') \\ \text{"continue"} \text{ if } \frac{1}{4}\sum_{s'=1}^{4} V^{N+1}(s') > 3s^2 + 5. \end{cases}$$

(g) The optimal policy in the dice game is Non Stationary as the decision to "Quit" or "continue" depends not only on the current state s, but also on the time step n.

As the game progresses & the number of remaining rolls decreases the player's optimal choice can change making the policy time dependent.