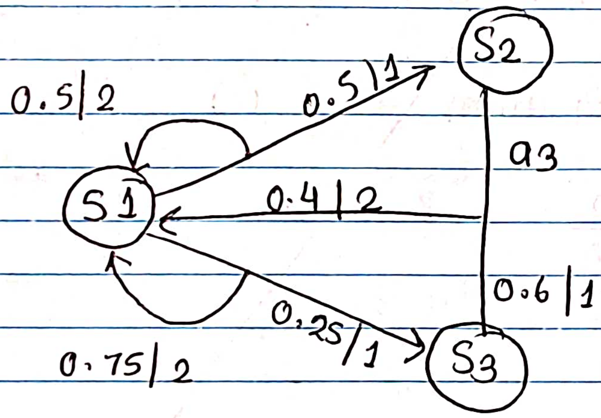
PAURAVI UDAY NAGARKAR
W1650209

homework No: 4.

Problem 1] An MDP state transition is given below. The agent wants to go from S1 or S2 to goal state S3. Suppose agent follows a fixed policy where it takes action a2 in state S1 and takes action a3 in state S2. for this fixed policy. Calculate expected cost to go from S1 to goal denoted as $V^{\pi}(S_2)$ & calculate expected cost to go from S2 to goal, denoted & associated $V^{\pi}(S_2)$.



→ given $T(S_1, a_2, S_1) = 0.5$
$C(S_1, a_2, S_1) = 2$

$$V^{\pi}(S_1) = \sum_{s'} T(s, a, s') \left[R(s,a,s') + \gamma V_k^{\pi}(s') \right]$$

$$V^{\pi}(S_1) = 0.5(2 + 1 * V^{\pi}(S_1) + 0.5(1 + V^{\pi}(S_2)]$$

$$V^{\pi}(S_1) = 1 + \frac{V^{\pi}(S_1)}{2} + \frac{1}{2} + \frac{V^{\pi}(S_2)}{2}$$

$$V^{\pi}(S_1) = T(S_1, a_2, S_2)\,[c(S_1, a_2, S_2) + \gamma \cdot V^{\pi}(S_2)] + T(S_1, a_2, S_1)\,[c(S_1, a_2, S_1) + \gamma \cdot V^{\pi}(S_1)]$$

$$= 0.5[1 + V^{\pi}(S_2)] + 0.5[2 + V^{\pi}(S_1)$$

$$= 0.5 + 0.5V^{\pi}(S_2) + 1 + 0.5V^{\pi}(S_1)$$

$$= 1.5 + 0.5V^{\pi}(S_1) + 0.5V^{\pi}(S_1)$$

$$0.5V^{\pi}(S_1) = 1.5 + 0.5V^{\pi}(S_2) \quad \ldots \quad (1)$$

$$0.5V^{\pi}(S_1) = 1.5 + 0.5\,\boxed{1.4 + 0.4V^{\pi}(S_1)}$$

we got this Equation By

$$V^{\pi}(S_2) = T(S_2, a_3, S_3).\,c(S_2, a_3, S_3) + 0] + T(S_2, a_3, a_1)\,c(S_2, a_3, S_1) + \gamma V^{\pi}(S_1))$$

$$= 0.6[1] + 0.8 + 0.4 * V^{\pi}(S_1)$$

$$V^{\pi}(S_2) = 1.4 + 0.4 * V^{\pi}(S_1) \quad \ldots \quad (2)$$

By Equation one & Equation two
We get.

$$0.5V^{\pi}(S_1) = 1.5 + 0.7 + 0.2V^{\pi}(S_1)$$

$$0.3V^{\pi}(S_1) = 2.3$$

$$V^{\pi}(S_1) = 7.33$$

$$V^{\pi}(S_2) = 1.4 + 0.4(V^{\pi} S_1))$$
$$= 1.4 + 2.93$$
$$= 4.33$$

hence Expected Cost to go from $S_1 = 7.33$ & $S_2 = 4.33$

Problem 2 Answer the following Question and show your work for Question b & c.

a] In Value iteration let K be iteration index. write a formula to update $Q_K(S, a)$ from $R(S, a, s')$ $T(S, a, S')$, $V_{K-1}(S')$, γ and write formula to Compute $V_K(S)$ from $Q_K(S, a)$

$Q_K(S, a) =$

$V_K(S) =$

a] → $$Q_K(S, a) = \sum_{S'} T(S, a, S')[R(S, a, S') + \gamma V_{K-1}(S')]$$

$$V_K(S) = \max_a Q_K(S, a)$$

B] → $$Q_1(A, 1) = 0[0 + 1*0] + 1(0 + 1*0)$$
$$= 0$$

$$Q_1(A, 2) = 1[1 + 1*0] + 0(0 + 0*0)$$
$$= 1$$

$$Q_1(A, 3) = 0.5[0 + 1*0] + 0.5[0 + 1*0]$$
$$= 0$$

$V_1(A) = \text{Max}(Q_1(A,1), Q_1(A,2), Q_1(A,3))$
$V_1(A) = \text{Max}(0, 1, 0)$
$V_1(A) = 1.$

$Q_1(B,1) = 0.5[10 + 1\times 0] + 0.5[0 + 1\times 0]$
$= 5 + 0$
$= 5$

$Q_1(B,2) = 1[0 + 1 * 0] + 0[0 + 1 * 0]$
$= 0$

$Q_1(B,3) = 0.5[2 + 1 * 0] + 0.5[4 + 1 * 0]$
$= 1 + 2$
$= 3$

$V_1(B) = \text{Max}(Q_1(B,1), Q_1(B,2), Q_1(B,3)]$

$V_1(B) = \text{Max}[5, 0, 3]$

$V_1(B) = 5$

$Q_2(A,1) = 0[0 + 1\times 1] + 1\times[0 + 1\times 5]$
$= 0 + 5$
$= 5$

$Q_2(A,2) = 1\times[1 + 1\times 1] + 0(0 + 1 * 5)$
$= 2$

$Q_2(A,3) = 0.5(0 + 1\times 1) + 0.5[0 + 1 * 5]$
$= 0.5 + 2.5$
$= 3$

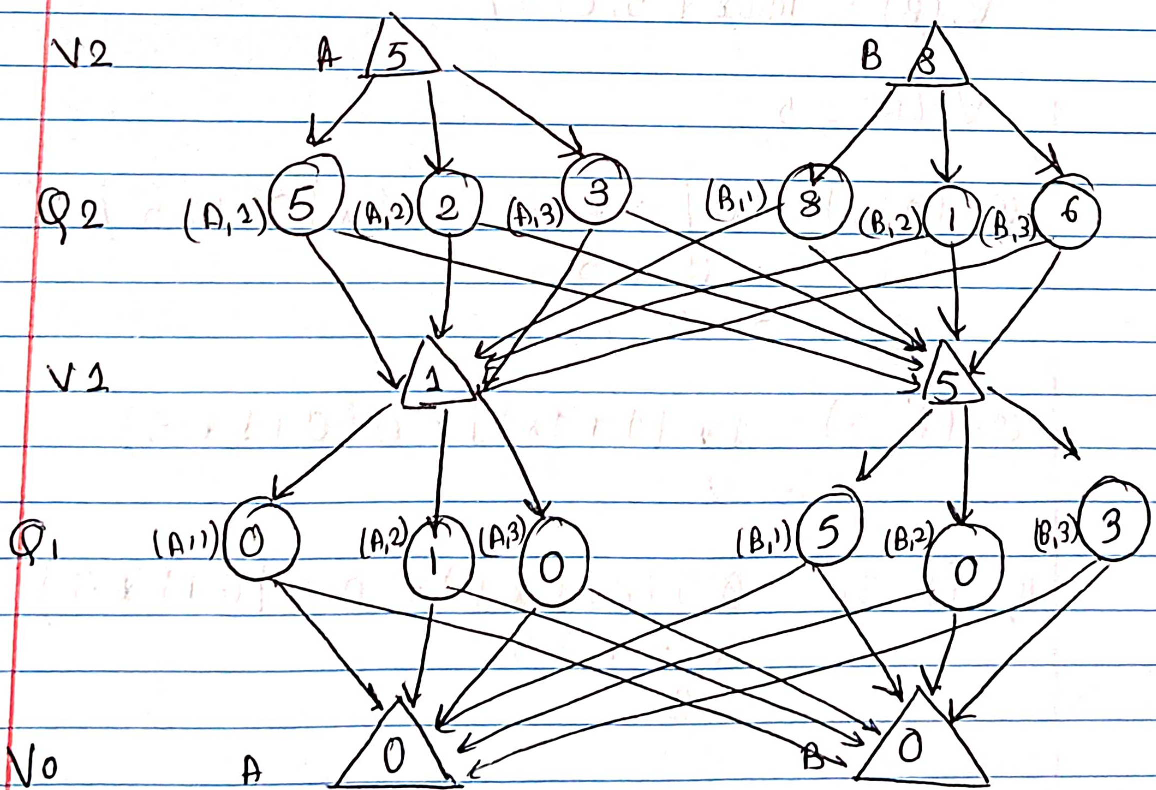
$Q_2(B,1) = 0.5[10+1] + 0.5[0+5]$
$= 5.5 + 2.5$
$= 8$

$Q_2(B,2) = 1[0+1] + 0[0+5]$
$= 1$

$Q_2(B,3) = 0.5[2+1] + 0.5[4+5]$
$= 1.5 + 4.5$
$= 6$

$V_2(A) = \text{Max}(5, 2, 3)$
$= 5$

$V_2(B) = \text{Max}(8, 1, 6)$
$= 8$



C. let $\pi_k(S)$ be optemal action in State S in k-th iteration fill the tables

→

S	$\pi_1(S)$
A	2
B	1

S	$\pi_2(S)$
A	1
B	1

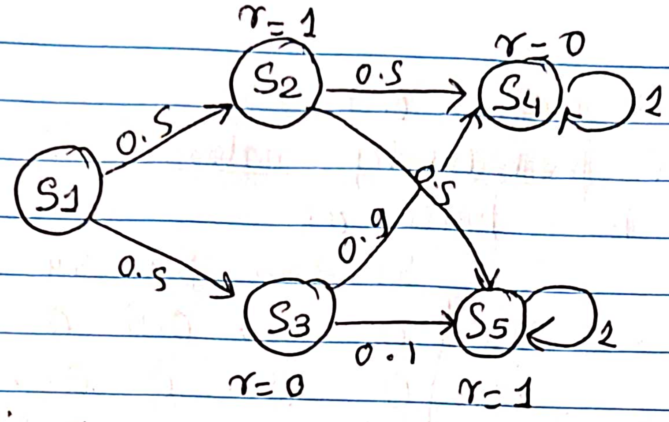
I) optemal action taken from A will be 2, as it gives maximum $V_2(A)$ value, Similarly for B, the optemal action will be 1.

II optemal action from A will be 1 as it will give maximum $V_1(A)$ value, Similarly for B optimal action will be 1

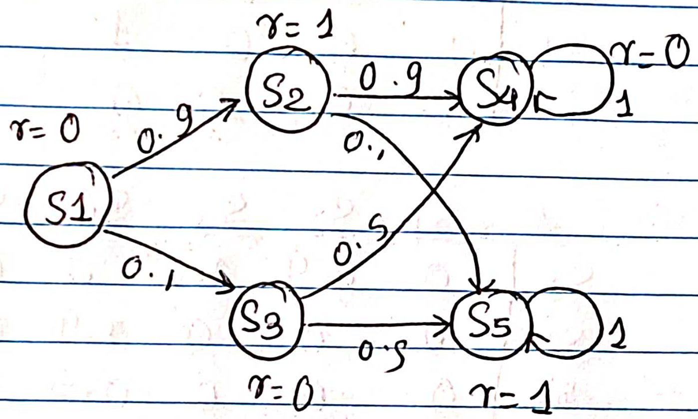
a) $Q_k(S,a) = \sum_{a} T(S,a,S')[R[S,a,S') + \gamma V_{k-1}(S')]$

$$V_k(S) = \max_{a} \sum T(S,a,S')(R(S,a,S') + \gamma V_{k-1}(S')]$$

Problem 3



Action 1



Calculate optimal values of states in MDP by Value iteration, assuming $\gamma = 0.9$ & initial values of states are zeros let $\epsilon = 10^{-4}$ for stoping criterion. The reward $r = R(S)$ is defined as a function of current state S, regardless of action a and next state S'

describe how the value iteration works, with proper math expression, then give the final Result

→ Ans To calculate the value of iteration.
The transition probabilities are mentioned on the arrows.

for Action 1 (Index 0)
transition probability matrix
can be built as

$$T[:,:,0] = \begin{array}{c|ccccc} & S_1 & S_2 & S_3 & S_4 & S_5 \\ \hline S_1 & 0 & 0.5 & 0.5 & 0 & 0 \\ S_2 & 0 & 0 & 0 & 0.5 & 0.5 \\ S_3 & 0 & 0 & 0 & 0.9 & 0.1 \\ S_4 & 0 & 0 & 0 & 1 & 0 \\ S_5 & 0 & 0 & 0 & 0 & 1 \end{array}$$

$s \quad s' \quad a$

$$T[:,:,1] = \begin{array}{c|ccccc} & S_1 & S_2 & S_3 & S_4 & S_5 \\ \hline S_1 & 0 & 0.9 & 0.1 & 0 & 0 \\ S_2 & 0 & 0 & 0 & 0.9 & 0.1 \\ S_3 & 0 & 0 & 0 & 0.5 & 0.5 \\ S_4 & 0 & 0 & 0 & 1 & 0 \\ S_5 & 0 & 0 & 0 & 0 & 1 \end{array}$$

$s \quad s' \quad a_1$

Rewards for both action are same as it does not depend on action or next state i.e.

$$R = \begin{array}{c|c} S_1 & 0 \\ S_2 & 1 \\ S_3 & 0 \\ S_4 & 0 \\ S_5 & 1 \end{array}$$

To calculate the optimal values of the states in MDP, we use Value Iteration method
Discounting factor $\gamma = 0.9$
$\varepsilon = 10^{-4}$

getting initial values $V_0(s)$ for all states S_1, S_2, S_3, S_4, S_5 as zero.

Running Bellman Equation for updating the Values at each iteration till values converge

$$V_{k+1}(S) \longleftarrow \max_a \sum_{S'} T(S,a,S')\left[R(S,a,S') + \gamma V_k(S')\right]$$

Here $R(S,a,S') = R(S)$

$$\therefore V_{k+1}(S) = R(S) + \gamma \max_a \sum_{S'} T(S,a,S') \cdot V_k(S')$$

where k = number of iteration.

① $V_0(S) = 0$ for S_1, S_2, S_3, S_4, S_5.

② $V_1(S_1) = R(S_1) + 0.9 \max[(0.5 * V_0(S_2) + (0.5 * V_0(S_3)), 0.9 * V_0(S_2) + 0.1 * V_0(S_3))]$

$= 0 + 0.9 * \max([0.5 * 0 + 0.5 * 0], [0.9 * 0 + 0.1 * 0])$

$\boxed{V_1(S_1) = 0}$

③ $V_1(S_2) = 1 + 0.9 \max[(0.5*0 + 0.5*0), (0.9*0 + 0.1*0)]$

$\boxed{V_1(S_2) = 1.}$

④ $V_1(S_3) = 0 + 0.9 \max[(0.5*0 + 0.5*0), (0.9*0 + 0.1*0)]$

$V_1(S_3) = 0$

Running Bellman Equation for updating the Values at each iteration till values converge

$$V_{k+1}(s) \longleftarrow \max_a \sum_{s'} T(s,a,s')\left[R(s,a,s')+\gamma V_k(s')\right]$$

Here $R(s,a,s') = R(s)$

$$\therefore V_{k+1}(s) = R(s) + \gamma \max_a \sum_{s'} T(s,a,s') . V_k(s')$$

where k = number of iteration.

(1) $V_0(s) = 0$ for S_1, S_2, S_3, S_4, S_5.

(2) $V_1(S_1) = R(S_1) + 0.9 \max[(0.5 * V_0(S_2) + (0.5 * V_0(S_3)),$
$0.9 * V_0(S_2) + 0.1 * V_0(S_3))]$

$= 0 + 0.9 * \max([0.5 * 0 + 0.5 * 0],$
$[0.9 * 0 + 0.1 * 0])$

$\boxed{V_1(S_1) = 0}$

(3) $V_1(S_2) = 1 + 0.9 \max[(0.5 * 0 + 0.5 * 0),$
$(0.9 * 0 + 0.1 * 0)]$

$\boxed{V_1(S_2) = 1.}$

(4) $V_1(S_3) = 0 + 0.9 \max[(0.5 * 0 + 0.5 * 0),$
$(0.9 * 0 + 0.1 * 0)]$

$V_1(S_3) = 0$

$V_1(S_4) = 0 + 0.9 * \text{Max}[(0.5*0) + (0.5*0),$
$(0.9*0) + (0.1*0)]$

$\boxed{V_1(S_4) = 0}$

$V_1(S_5) = 1 + 0.9 \text{ Max}[(0.5*0) + (0.5*0),$
$(0.9*0) + (0.1*0)]$

$\boxed{V_1(S_5) = 1}$

$V = \begin{bmatrix} 0 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$, we further calculate we calculate the same function

$V_2(S_1) = 0.81$
$V_2(S_2) = 1.45$
$V_3(S_3) = 0.45$
$V_2(S_4) = 0$
$V_2(S_5) = 1.9$

The above iteration goes on till the stopping criteria is met

$$\underset{S}{\text{Max}} \left| V_{k+1}(S) - V_k(S) \right| < \frac{\varepsilon(1-\gamma)}{\gamma}$$

↑

maximum value of absolute difference between two iteration values.

Rounding off values V(s) at each iteration, we get final optimal value at 52nd iteration

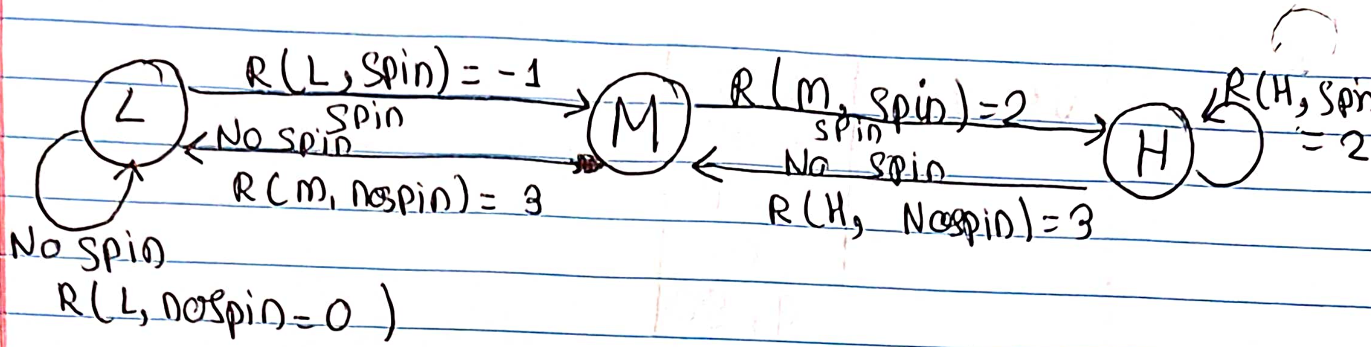
$$V S_2^* = \begin{bmatrix} 4.86 \\ 5.48 \\ 4.48 \\ 0 \\ 9.96 \end{bmatrix}$$

Hence the iteration goes for S2 to end the iteration.

Problem 4 Consider the following Problem. Consider a rover that operates on a slope and uses Solar panels to recharge. It can be in one of three states, high, medium, low of slope. If it spins it wheels, it climbs the slope in each time step or stay high.

a] Draw State transition graph for this MDP. use arrows to represent State transitions, specify the action and reward on each arrow

→ It has three States
low
medium
High.
If Spins wheel It climb Slope in each time step
If Non spin Slides down the Slope.
Reward / slope gain 3 unit of energy per step



L = low M = medum H - high

→ Spinning uses '1 unit' per time step.

$\therefore$ Rewards - $R(S, a)$

$R(L, \text{Spin}) = -1$ $R(L, \text{Nospin}) = 0$

$R(M, \text{Spin}) = 3 - 1 = 2$ $R(M, \text{Nospin}) = 3$

$R(H, \text{Spin}) = 3 - 1 = 2$ $R(H, \text{nospin}) = 3$

b] Solve the MDP using value value iteration with discount factor of 0.8 Start with 0 values. let $\epsilon = 10^{-4}$ for stopping criterion : what are state values after two iteration

→ Solving with Bellman Equation

$$V_{K+1}(S) = \max_a \sum_{S'} T(S, a, S') \left[R(S, a, S') + \gamma \cdot V_K(S') \right]$$

where $K \to$ number of iteration.

Starting with initial values of state L, M, H.

$V_0(S) = 0$ for L, M, H, $\gamma = 0.8$

$T(S, a, S') = 1$ for all low (L)

$V_{k+1}(L, \text{Spin}) = -1 + 0.8 * V_k(M)$

$V_{k+1}(L, \text{Nospin}) = 0 + 0.8 * V_k(L)$

$V_{k+1}(L) = \text{Max}[V_{k+1}(L, \text{Spin}), V_{k+1}(L, \text{nospin})]$

$\left.\begin{array}{l} V(L, \text{spin}) = -1 + 0.8*0 \\ V_1(L, \text{nospin}) = 0 + 0.8*0 \end{array}\right\}$ $V_1(L) = \text{Max}(-1, 0)$

$\boxed{V_1(L) = 0}$

We well formulate $V_1(M)$ for High & Medium as well

$V_1(M, \text{Spin}) = 2 + 0.8*0 = 2$
$V_1(M, \text{nospin}) = 3 + 0.8*0 = 3$
$V_1(M) = \text{Max}(2, 3)$
$\boxed{V_1(M) = 3}$

$V_1(H, \text{spin}) = 2 + 0.8*0 = 2$
$V_1(H, \text{nospin}) = 3 + 0.8*0 = 3$
$\therefore V_1(H) = \text{Max}(2, 3)$
$\boxed{V_1(H) = 3}$

$V_1(S) = \begin{bmatrix} 0 \\ 3 \\ 3 \end{bmatrix}$

Similarly formulating for second iteration

$V_2(L, S) = -1 + 0.8*3 = 1.4$
$V_2(L, \text{Nospin}) = 0 + 0.8*0 = 0$
$V_2(L) = \max(1.4, 0)$
$\boxed{V_2(L) = (1.4)}$

$V_2(M, \text{spin}) = 2 + 0.8 * 3 = 4.4$

$V_2(M, \text{Nospin}) = 3 + 0.8 * 0 = 3$

$V_2(M) = \max(4.4, 3)$

$\boxed{V_2(M) = \cancel{\text{PAO}}\ 4.4}$

$V_2(H, \text{spin}) = 2 + 0.8 * 3 = 4.4.$

$V_2(H, \text{Nospin}) = 3 + 0.8 * 3 = 5.4.$

$V_2(H) = \max(4.4, 5.4)$

$\boxed{V_2(H) = 5.4}$

Running the Value we get

$$V_2 = \begin{bmatrix} 1.4 \\ 4.4 \\ 5.4 \end{bmatrix}$$

Running values iteration till threshold.

$$\max_S \left| V_{k+1}(S) - V_k(S) \right| < \varepsilon \frac{(1-\gamma)}{\gamma}$$

Where $\varepsilon = 10^{-4}$, $\gamma = 0.8$.

→ Rounding off the Value at each iteration we get final optimal solution at 54^{th} iteration as

$$V54 = \begin{bmatrix} 8.78 \\ 12.22 \\ 12.78 \end{bmatrix}$$

c] Determine the optimal policy by using optimal values obtained in Question b. What method do you use to obtain optimal policy? Write out math expression of optimal policy.

→ To determine the optimal policy by using optimal policy, we use policy extraction (
The formula for the same is

$$\pi^*(S) = \arg\max_{a} \sum_{S'} T(S,a,S') * [(R(S,a,S') + \gamma V^*(S')]$$

$$\pi^*(S) = \arg\max_{a} Q(S,a)$$

Using the optimal values :-

$Q^*(S,a)$: $Q^*(L, \text{Spin}) = -1 + 0.8 * (12.22)$
$= 8.78$

$Q^*(S,a)$: $Q^*(L, \text{Nospin}) = 0 + 0.8 * (8.78)$
$= 7.02$

$Q^*(S,a) = Q^*(L) = \text{Max}(8.78, 70.2)$
$= 8.78$
i.e Max value for spin

$Q^*(M, \text{Spin}) = 2 + 0.8 * (12.78) = 12.22$
$(M, \text{Nospin}) = 3 + 0.8 * (8.78) = 10.02$
$Q^*(M) = \text{Max}(12.22, 10.02)$
$= 12.22.$
i.e Max value for spin

$$Q^*(H, Spin) = 2 + 0.8 * (12.78) \quad 12.22$$
$$Q^*(H, Nospin) = 2 + 0.8 * (12.22) = 12.78$$
$$Q^*(H) = Max(12.22, 12.78)$$
$$Q^*(H) = 12.78$$

Maximum Value we get for Nospin.

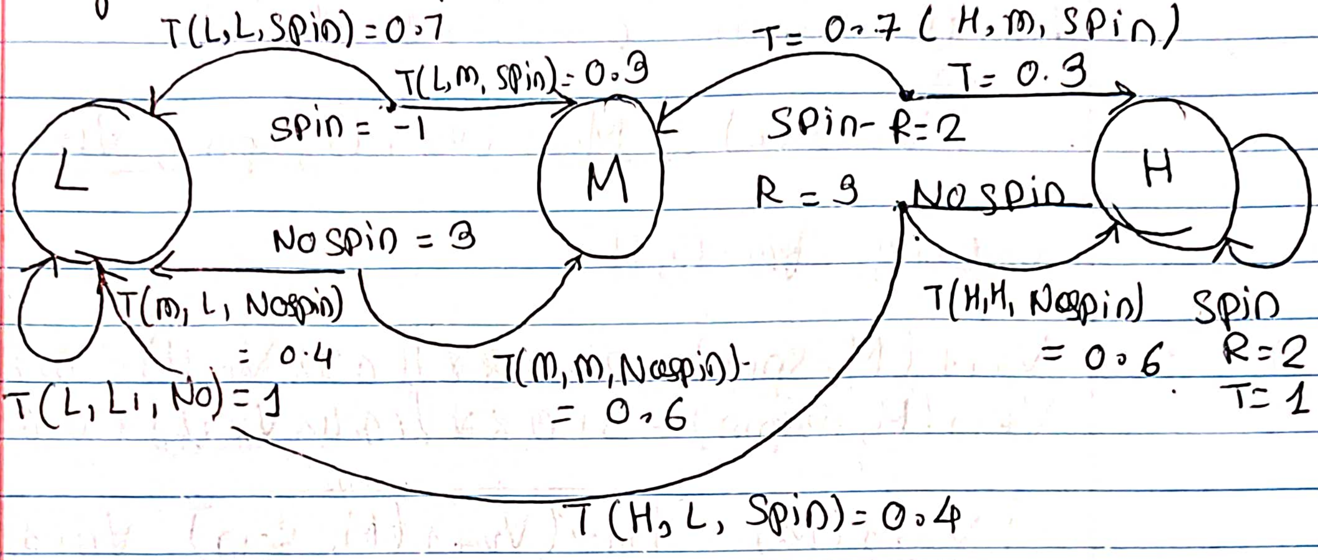
Based on above Q^* Values.
The optimal policies for each State is max over each action at particular State

$\pi^*(L) =$ Spin
$\pi^*(M) =$ Spin
$\pi^*(H) =$ Nospin.

Problem 5 MDP- Stochastic case
The Robot Problem: if it spins its wheels, it climbs the slope in each time step. all with Probability 0.3 and it stays where it is with Probability 0.7 if it does not spin its wheel it slides down the low with Probability 0.4 & stays where it is with Probability 0.6. If it spins its wheels at State high then it stays at high with Probability 1.

a] Draw transition graph for this MDP. use arrows to represent State transition, specify the action and reward on each row

→ graphical Representation of Situation



b] Solve MDP using Value iteration with discount factor of 0.8. $\epsilon = 10^{-4}$. What are states Value after two iterations.

→ Solving MDP using Value iteration with bellman's updated equation as

$$V_{k+1}(S) \leftarrow \max_{a} \sum_{S'} T(S, a, S')[R(S, a, S') + \gamma V_k(S')]$$

Starting with 0 as initial Values of State L, M, H

$V_0(S) = 0$ for L, M, H.

$\gamma = 0.8$

$T(S, S', a)$ = as indicated on diagram above (transition Probability from state S to S' with action a)

$V_{k+1}(L, \text{Spin}) = -1 + 0.8 * 0.3 * V_k(M) + 0.8 * 0.7 \times V_k(L)$
$V_{k+1}(L, \text{Nospin}) = 0 + 0.8 * 1 * V_k(L)$

$$\boxed{V_{k+1}(L) = \text{Max}\left(V_{k+1}(L, \text{Spin}), V_{(k+1)}(L, \text{nospin})\right)}$$

Similarly for M, H.

$V_{k+1}(M, \text{Spin}) = 2 + 0.8 * [(0.3 * V_k(H) + 0.7 * V_k(M)]$
$V_{k+1}(M, \text{Nospin}) = 3 + 0.8 * [(0.4 \times V_k(L) + 0.6 * V_k(M)]$

$$\boxed{V_{k+1}(M) = \text{Max}\left(V_{k+1}(M, \text{Spin}), V_{k+1}(M, \text{Nospin})\right)}$$

$V_{k+1}(H, \text{spin}) = 2 + 0.8 * 1 * V_k(H)$

$V_{k+1}(H, \text{Nospin}) = 3 + 0.8 * [0.4 * V_k(L) + 0.6 \times V_k(H)]$

$$\boxed{V_{k+1}(H) = \text{Max}\left(V_{k+1}(H, \text{Spin}), V_{k+1}(H, \text{Nospin})\right)}$$

Plugging the values as above: 1st iteration values are

$$V_1(S) = \begin{array}{c} L \\ M \\ H \end{array} \begin{bmatrix} \max(-1, 0) \\ \max(2, 3) \\ \text{Max}(2, 3) \end{bmatrix} = \begin{bmatrix} 0 \\ 3 \\ 3 \end{bmatrix}$$

for $k = 1$

$$V_2(S) = \begin{array}{c} L \\ M \\ H \end{array} \begin{bmatrix} \text{Max}(-0.28, 0) \\ \text{Max}(4.40, 4.44) \\ \text{Max}(4.40, 4.44) \end{bmatrix} = \begin{bmatrix} 0 \\ 4.44 \\ 4.44 \end{bmatrix}$$

Running the Value iterations till

$$\max_{S} \left| V_{k+1}(S) - V_k(S) \right| < \varepsilon \frac{(1-\gamma)}{\gamma}$$

$\varepsilon = 10^{-4}$, $\gamma = 0.8$

Rounding off to two decimal Places

$$V^*_{k+1} = \begin{bmatrix} 3.18 \\ 9.98 \\ 9.98 \end{bmatrix}$$

c] optimal policy by using optimal Values obtained in Question (b). What method do you use to obtain optimal policy? write math expression.

→ optimal policy $\pi^*(S) = \arg\max_{a} Q^*(S,a)$

$Q^*(L, \text{Spin}) = 0.3(-1 + 0.8 * 10) + 0.7(-1 + 0.8 \times 3.18)$
$= 3.18$

$Q^*(L, \text{Nospin}) = 1(0 + 0.8 * 3.18)$
$= 2.54$

$Q^*(L) = \max(3.18, 2.54)$
$= 3.18 \rightarrow$ spin.

$Q^*(m, \text{spin}) = 0.3(2 + 0.8 * 10) + 0.7(2 + 0.8 * 10)$
$= 10$

$Q^*(m, \text{Nospin}) = 0.4(3 + 0.8 * 3.18) + 0.6(3 + 0.8 * 10)$
$= 8.52$

$Q^*(M) = \text{Max}(10, 8.52)$

$Q^*(M) = 10$

optimal policy in spin state.

$Q^*(H, \text{Spin}) = 1(2 + 0.8 * 10) = 10$

$Q^*(H, \text{Nospin}) = 0.4(3 + 0.8 * 3.18) + 0.6(3 + 0.8 * 10)$

$= 8.82$

optimal policy is in spin State.

$Q^*(H) = \text{Max}(10, 8.82)$

$Q^*(H) = 10$

Thus we achieve optimal policy as spin in all the states

$\pi^*(L) \rightarrow$ Spin

$\pi^*(M) \rightarrow$ Spin

$\pi^*(H) \rightarrow$ Spin