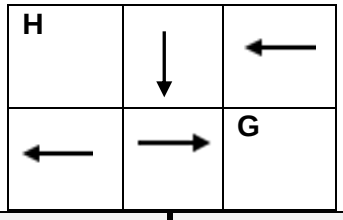
2. **Reward**:

- Any State to Hole $= -30$
- Any Sate to Goal $= +30$
- all remaining transitions $= +2$

H	S_0	S_1
S_2	S_3	**G**

A: $\{Left = 0, Right = 1\}, \gamma = 0.9\ ,\ \alpha = 0.1$

$b(.|s)$:



State	Action	Q-Value	State	Action	Q-Value
S_0	Left	0.3	S_2	Left	0.4
S_0	Right	0.9	S_2	Right	0.6
S_0	Up	0.5	S_2	Up	0.3
S_0	Down	0.6	S_2	Down	0.4
S_1	Left	0.6	S_3	Left	0.2
S_1	Right	0.2	S_3	Right	0.9
S_1	Up	0.7	S_3	Up	0.7
S_1	Down	0.8	S_3	Down	0.1

Sarsa: Current State: S_1, Current State: S_0

Q-L Current State: S_1, Current State: S_2

2)

$$Q(s,a) = Q(s,a) + \alpha\left[R(s,a,s') + \gamma Q_{bm}(s',a') - Q(s,a)\right]$$

Q2ai) (1) S_1 : (SARSA)

$s = S_1$, $s' = S_0$.

$a = \text{left}$, $a' = \text{down}$.

$$Q(S_1, \text{Left}) = Q(S_1, \text{Left}) + 0.1\left[2 + 0.9 * Q(S_0, \text{down}) - Q(S_1, \text{Left})\right].$$

$$Q(S_1, \text{Left}) = 0.6 + 0.1\left[2 + 0.9 * 0.6 - 0.6\right].$$

$$\boxed{Q(S_1, \text{Left}) = 0.794}$$

Q2aii) (2) S_0 (SARSA).

$$Q(S_0, \text{down}) = Q(S_0, \text{down}) + 0.1\left[2 + 0.9 * Q(S_3, \text{Right}) - Q(S_0, \text{down})\right].$$

$$Q(s_0, down) = 0.6 + 0.1[2 + 0.9 * 0.9 - 0.6]$$

$$Q(s_0, down) = 0.821$$

Q2bi) 2b) Q-Learning) S1 :

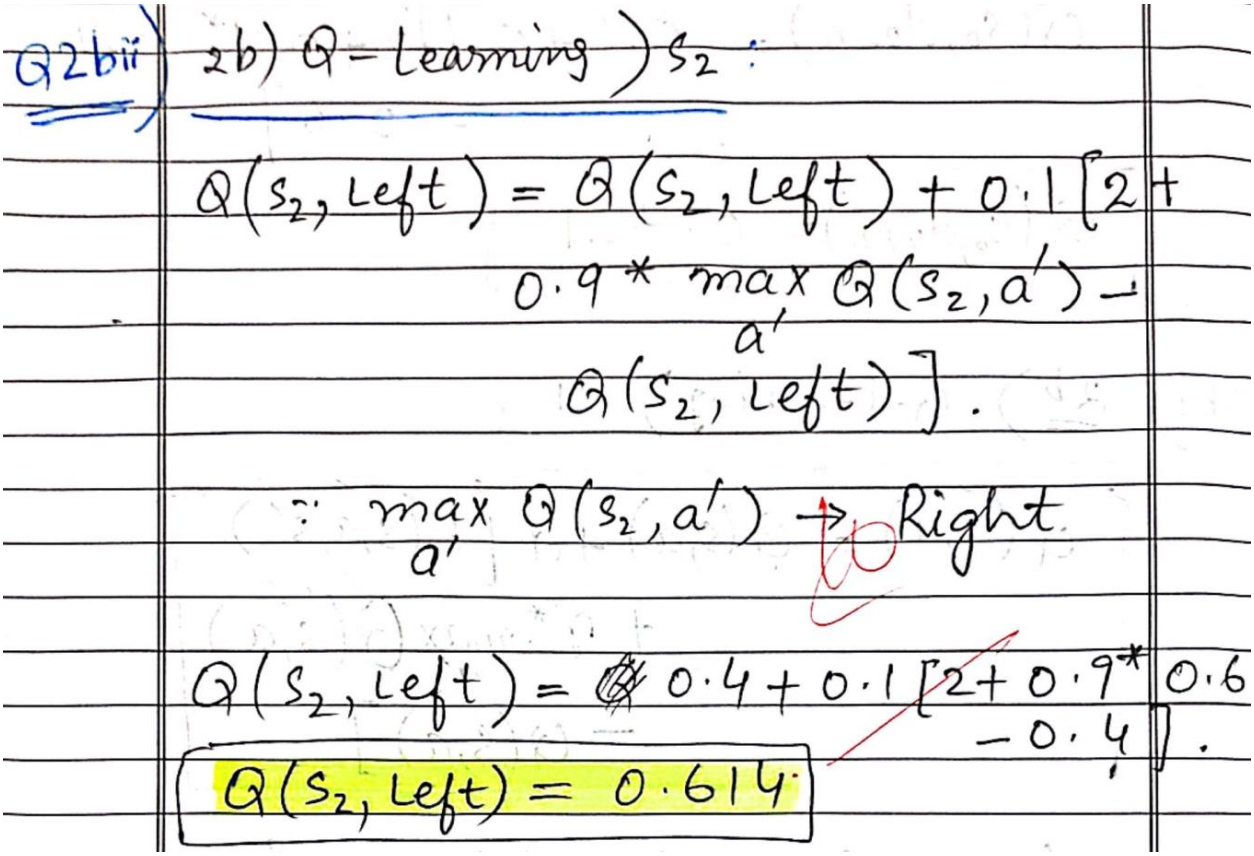
$$Q(s,a) = Q(s,a) + \alpha [R(s,a,s') + \gamma * \max_{a'} Q(s',a') - Q(s,a)]$$

$$Q(s_1, Left) = Q(s_1, Left) + 0.1[2 + 0.9 * Q(s_0, Right) - Q(s_1, Left)]$$

$\therefore \max_{a'} Q(s_0, a') \Rightarrow Right$

$$Q(s_1, Left) = 0.6 + 0.1[2 + 0.9 * 0.9 - 0.6]$$

$$Q(s_1, Left) = 0.821$$



3.

- Any State to Hole = -30
- Any Sate to Goal = $+30$
- all remaining transitions = $+2$

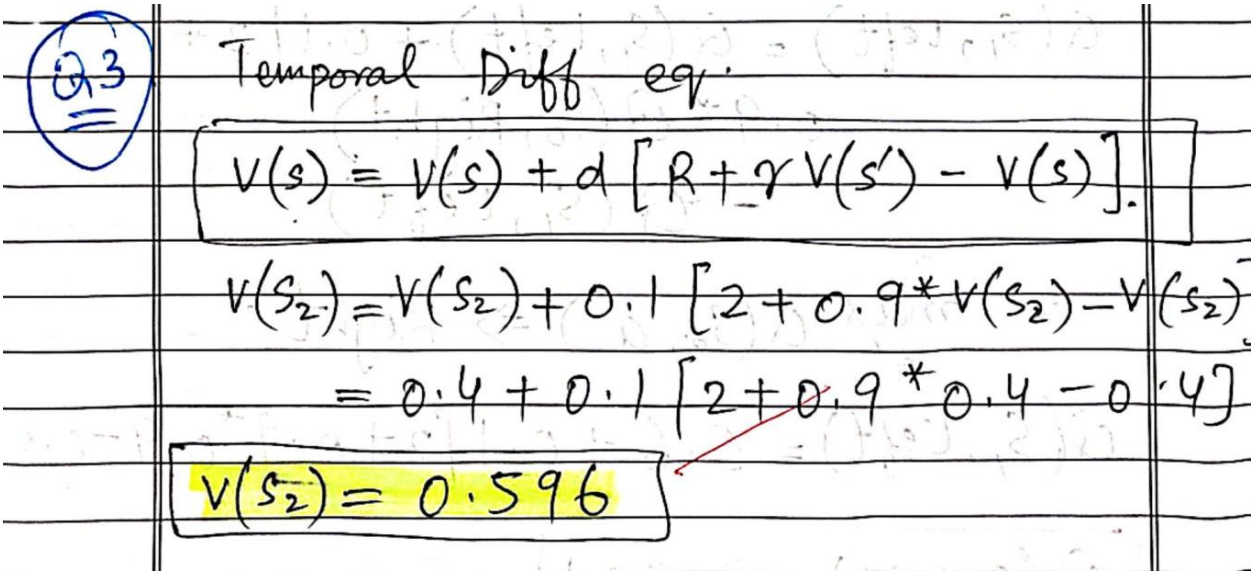
H	S_0	S_1
S_2	S_3	**G**

A: $\{Left = 0, Right = 1, Up = 2, Down = 3\}, \gamma = 0.9\ ,\ \alpha = 0.1$

H	↓	←
←	→	**G**

State	Value				
S_0	0.9	S_2	0.4	H	0
S_1	0.6	S_3	0.5	G	0

TD learning – Evaluation. Current State = S_2, Current State = S_3



State	Value
S_0	0.9
S_1	0.6
S_2	0.596
S_3	0.5
H	0
G	0

$$V(S_3) = V(S_3) + \alpha [R(S_3, \text{Right}, G) + \gamma * V(G) - V(S_3)]$$

$$10 = 0.5 + 0.1 [30 + 0.9 * 0 - 0.5]$$

$$V(S_3) = 3.45$$

State	Value
S_0	0.9
S_1	0.6
S_2	0.596
S_3	3.45
H	0
G	0

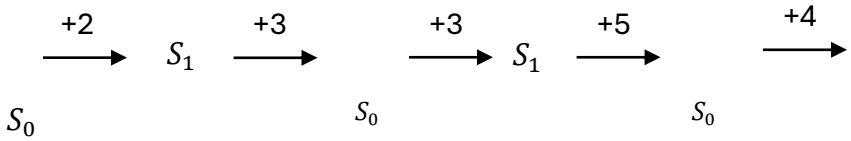
4.

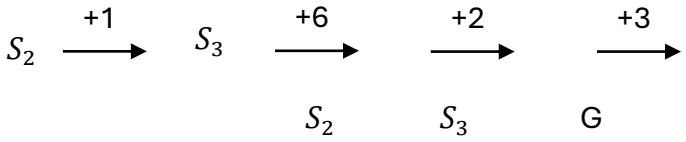
H	S_0	S_1
S_2	S_3	**G**

$\gamma = 0.9$,

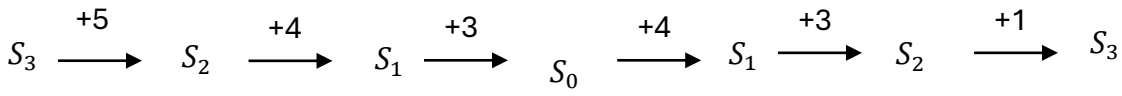
$\pi(. \mid s) = 0.5 \quad \forall s$

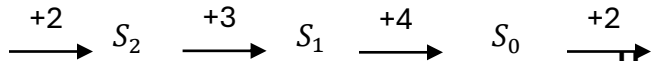
Trajectory – I:





Trajectory – II:





Compute following using Every Visit MC (EVMC):

1. $V(S_1)$
2. $V(S_3)$

Q4) Monte-carlo.

$V(s) = V(s) + \alpha [G - V(s')]$.

① $V(S_1)$:

Trajectory 1:

$S_0 \to S_1 \to S_0 \to S_1 \to S_0 \to S_2 \to S_3 \to S_2 \to S_3 \to G$.

$3+3+5+4+1+6+2+3 = 27$

$5+4+1+6+2+3 = 21$

Trajectory - II:

$S_3 \to S_2 \to S_1 \to S_0 \to S_1 \to S_2 \to S_3 \to S_2 \to S_1 \to S_0 \to H$.

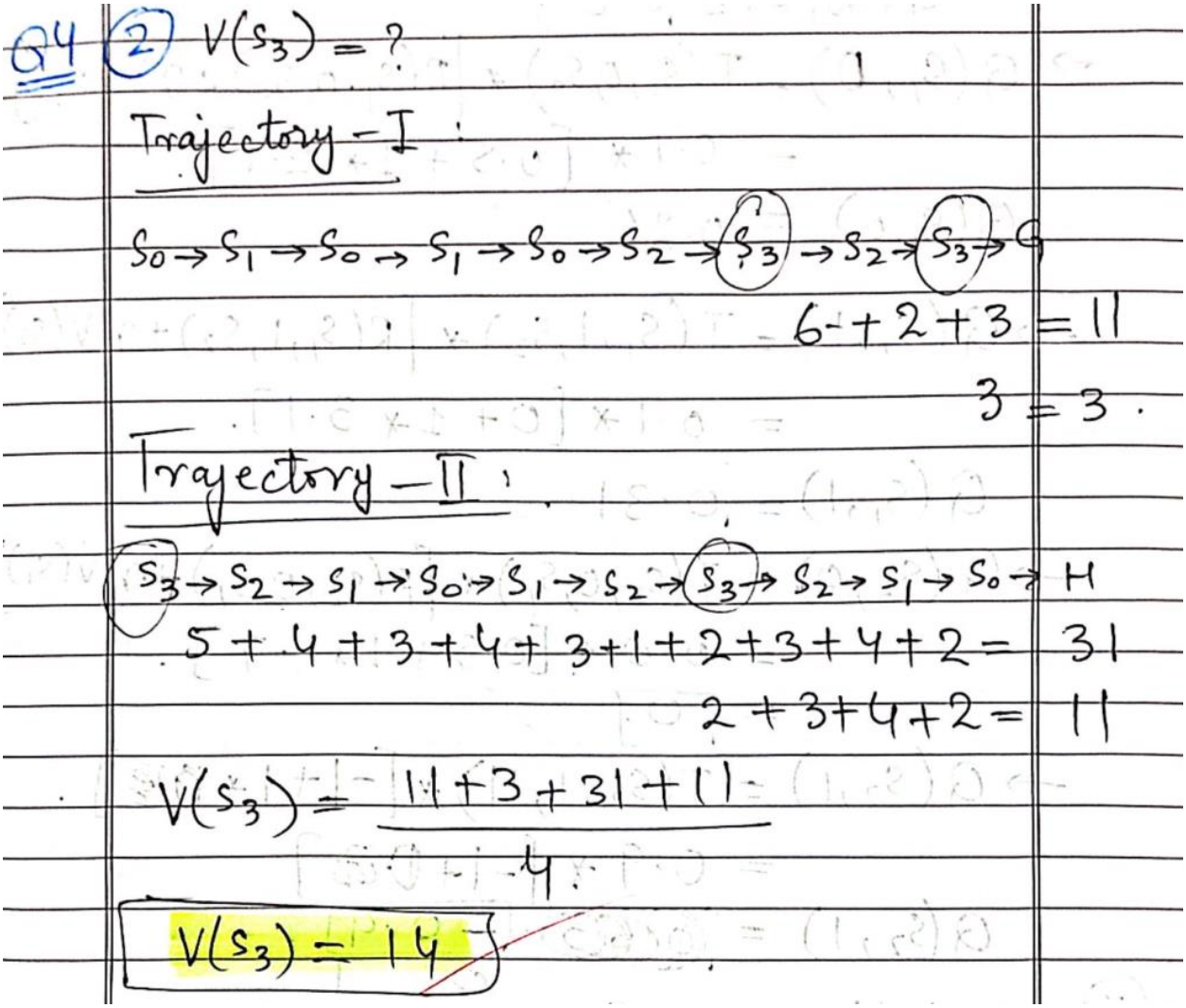
$3+4+3+1+2+3+4+2 = 22$

$3+1+2+3+4+2 = 15$

$4+2 = 6$.

$V(S_1) = \frac{27+21+22+15+6}{5}$

$V(S_1) = 18.2$



5

Value Iteration

G	S_0	S_1	S_2	**H**

A: $\{Left = 0, Right = 1\}, \gamma = 1$,

s	a	s'	T	R
S_0	1	S_1	0.1	0.5
S_0	0	G	0.8	+1
S_1	0	S_0	0.1	0.5
S_1	1	S_2	0.1	0
S_2	0	S_1	0.0	0.5
S_2	1	H	0.9	-1

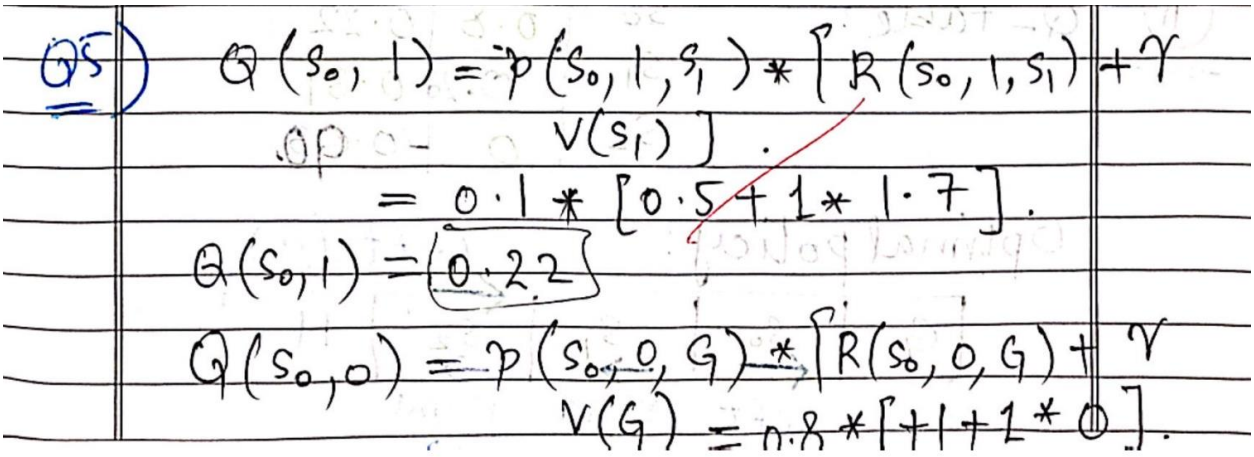
After running step 2, following optimal values (V^*) have been computed:

Optimal value function from Step 2 (not shown in in the question):

State	Value
S_0	2.1
S_1	1.7
S_2	3.1
H	0
G	0

Compute following using Value Iteration

I. Compute Q values from the Optimal value.
II. Update Q-table.
III. Compute Optimal Policy



$Q(s_0, 0) = \boxed{0.8}$

$\rightarrow Q(s_1, 0) = T(s_1, 0, s_0) * [R(s_1, 0, s_0) + \gamma V(s_0)]$

$= 0.1 * [0.5 + 1 * 2.1]$.

$Q(s_1, 0) = \boxed{0.26}$

$\rightarrow Q(s_1, 1) = T(s_1, 1, s_2) * [R(s_1, 1, s_2) + \gamma V(s_2)]$

$= 0.1 * [0 + 1 * 3.1]$.

$Q(s_1, 1) = \boxed{0.31}$

$\rightarrow Q(s_2, 0) = T(s_2, 0, s_1) * [R(s_2, 0, s_1) + \gamma V(s_1)]$

$= 0.0 * [0.5 + 1 * 1.7]$.

$= \boxed{0}$

$\rightarrow Q(s_2, 1) = T(s_2, 1, H) * [-1 + 1 * 0]$.

$= 0.9 * [-1 + 0]$

$Q(s_2, 1) = \boxed{-0.9}$

(ii) Q-table:

	0	1
s_0	0.8	0.22
s_1	0.26	0.31
s_2	0	-0.90

Optimal policy:

****