Q-Learning

(a) Q-learning update rule for a sample (s, a, s', r):

$$Q(s,a) \leftarrow Q(s,a) + \alpha\left[r(s,a,s') + \gamma \max_{a'} Q(s',a') - Q(s,a)\right]$$

or

$$Q(s,a) \leftarrow (1-\alpha)\, Q(s,a) + \alpha\left[r(s,a,s') + \gamma \max_{a'} Q(s',a')\right]$$

(b) (i) Sample : $(s = S_1, a = A_1, s' = S_1, r = (-10))$

Initialisation :

Q	S_1	S_2
A_1	0	0
A_2	0	0

Update :

$$Q(S_1,A_1) \leftarrow Q(S_1,A_1) + \alpha\Big[r(S_1,A_1,S_1) + \gamma \cdot \big[\max[Q(S_1,A_1), Q(S_1,A_2)\big] - Q(S_1,A_1)$$

$$Q(S_1,A_1) \leftarrow 0 + 0.5 \times \left[-10 + (0.5 \times \max(0,0)) - 0\right]$$

$$Q(S_1,A_1) \leftarrow (-5)$$

Resulting Q-table

Q	S_1	S_2
A_1	-5	0
A_2	0	0

(ii) Sample : $(s = S_1, a = A_2, s' = S_2, r = (-10))$

$$Q(S_1,A_2) \leftarrow Q(S_1,A_2) + \alpha \times \left[r(S_1,A_2,S_2) + \gamma \times \max(Q(S_2,A_1), Q(S_2,A_2)) - Q(S_1,A_2))\right]$$

$$Q(S_1,A_2) \leftarrow 0 + 0.5 \times \left[(-10) + 0.5 \times \max(0,0) - 0\right]$$

$$Q(S_1,A_2) \leftarrow 0 + 0.5 \times (-10)$$

$$Q(S_1,A_2) \leftarrow (-5)$$

Resulting Q-table :

Q	S_1	S_2
A_1	-5	0
A_2	-5	0

(iii) Sample: $(s = S_2 ,\ a = A_1 ,\ s' = S_1 ,\ r = (+20))$

$$Q(S_2,A_1) \leftarrow Q(S_2,A_1) + \alpha \cdot \left[r(S_2,A_1,S_1) + \gamma \cdot \max(Q(S_1,A_1), Q(S_1,A_2)) - Q(S_2,A_1) \right]$$

$$Q(S_2,A_1) \leftarrow 0 + 0.5 \times [+20 + 0.5 \times \max(-5,-5) - 0]$$

$$Q(S_2,A_1) \leftarrow 0 + 0.5 \times [20 - 2.5]$$

$$Q(S_2,A_1) \leftarrow 8.75$$

Resulting Q-table:

Q	S_1	S_2
A_1	-5	8.75
A_2	-5	0

(iv) Sample: $(s = S_1 ,\ a = A_2 ,\ s' = S_2 ,\ r = (-10))$

$$Q(S_1,A_2) \leftarrow Q(S_1,A_2) + \alpha \times \left[r(S_1,A_2,S_2) + \gamma \times \max(Q(S_2,A_1), Q(S_2,A_2)) - Q(S_1,A_2) \right]$$

$$Q(S_1,A_2) \leftarrow -5 + 0.5 \times [(-10) + 0.5 \times \max(8.75, 0) - (-5)]$$

$$Q(S_1,A_2) \leftarrow -5 + 0.5 \times [(-10) + 0.5 \times 8.75 + 5]$$

$$Q(S_1,A_2) \leftarrow -5.3125$$

Resulting Q-table:

Q	S_1	S_2
A_1	-5	8.75
A_2	-5.3125	0

(c) Optimal policy:

$$\pi^*(s) = \underset{a}{\operatorname{argmax}}\ Q(s,a)$$

$$\pi^*(s_1) = \underset{(A_1,A_2)}{\operatorname{argmax}} \left[\underset{\underset{(-5)}{\Downarrow}}{Q(s_1,A_1)} , \underset{\underset{(-5.3125)}{\Downarrow}}{Q(s_1,A_2)} \right] = A_1$$

$$\pi^*(s_2) = \underset{(A_1,A_2)}{\operatorname{argmax}} \left[\underset{\underset{(8.75)}{\Downarrow}}{Q(s_2,A_1)} , \underset{\underset{(0)}{\Downarrow}}{Q(s_2,A_2)} \right] = A_1$$