* Q- Learning ⇒

More general form of Q- Learning,

$$Q(a,s) \leftarrow Q(a,s) + \alpha\left(R(s) + \gamma \max\left[Q(a',s')\right] - Q(a,s)\right)$$

new Q-value

where,

$Q(a,s)$: Current Q-value

α : Learning rate

$R(s)$: Reward for taking an action in a state.

γ : Discount rate

$\max Q(a', s')$: Maximum expected future rewards by taking any action.

$Q(a,s)$: Current Q-value (specific)

For example : Consider a system with two states & two actions. $\alpha = 0.2$, $\gamma = 0.4$. Initially Q-table is empty. Calculate Q- value after performing following actions,

1. Current State : S_1, Reward : -10, Action (a_1) : $S_1 \rightarrow S_1$
2. Current State : S_1, Reward : -10, Action (a_2) : $S_1 \rightarrow S_2$
3. Current State : S_2, Reward : +10, Action (a_1) : $S_2 \rightarrow S_1$

⇒ Initial Q-table,

	S_1	S_2
a_1	0	0
a_2	0	0

$Q(a_1, s_1)$

1. $Q(a_1, s_1) \leftarrow Q(a_1, s_1) + \alpha\left(R(s) + \gamma \max_{a'(s'=s_1)}\left[Q(a', s')\right] - Q(a_1, s_1)\right)$

$Q(a_1, s_1) \leftarrow 0 + 0.2\left(-10 + 0.4 \max[0, 0] - 0\right)$

$Q(a_1, s_1) \leftarrow -2$

	S_1	S_2
a_1	-2	0
a_2	0	0

2. $Q(a_2, s_1) \leftarrow Q(a_2, s_1) + \alpha\left(R(a_2, s_1) + \gamma \max_{a'(s'=s_2)} [Q(a', s')] - Q(a_2, s_1)\right)$

$\leftarrow 0 + 0.2(-10 + 0.4 \max[0,0] - 0)$

$\leftarrow -2$

	s_1	s_2
a_1	-2	0
a_2	-2	0

3. $Q(a_1, s_2) \leftarrow Q(a_1, s_2) + \alpha\left(R(a_2, s_2) + \gamma \max_{a'(s'=s_1)} [-2, -2] - Q(a_1, s_2)\right)$

$\leftarrow 0 + 0.2(+10 + 0.4(-2) - 0)$

$\leftarrow 1.84$

CS: S2
R: +2
A1: S2->S1
Q(a1,S2) = 2.0192

	s_1	s_2
a_1	-2	1.84
a_2	-2	1.472

Current state: S2,
Reward: +5
A2: S2 -> S2
Q(a2,S2) = Q(a2,S2) +alpha (R(a2,s2) + disf * max [Q (a1,S2),Q(a2,S2)]-Q(a2,S2))

Q(a2,S2) = 0 + 0.2(+5 + 0.4 * mac[1.84,0] - 0)
=0.2(5+0.4(1.84)) = 1.472