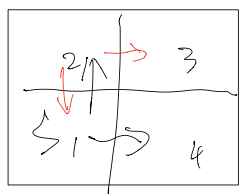


Quality

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha (r_t + \gamma \max_{a'} Q(s_{t+1}, a') - Q(s_t, a_t))$$

$Q(s_t, a_t)$: State Action
 α : learning rate
 r_t : reward
 γ : discount
 $\max_{a'} Q(s_{t+1}, a')$: self max
 $Q(s_t, a_t)$: current and best difference
 $r_t = r(s_t, a_t)$: reward value by state & action

0-1 不学习 固定情景



$$s_1 a \neq s_2 a$$

$$s_1(s, \uparrow)$$

$$Q_1 = 0 + \alpha (R + (\gamma \cdot \max Q') - Q_1)$$

$$Q_1 = \alpha R + \alpha \gamma \max + s_1$$

Q table

States

Action time Q 值

	↑	↓	→	←	
1	0.7	0.9	0.8	0.6	0
2	0.5	0.6	0.7	0.4	0
3	0.4	0.5	0.6	0.3	0
4	0.3	0.4	0.5	0.2	0
5	0.2	0.3	0.4	0.1	0
6	0.1	0.2	0.3	0.0	0