Quality

$Q(s_t, a_t)$ ← $Q(s_t, a_t) + \alpha \times (r_t + [\gamma \times \max Q(s_{t+1}, a_{t+1})] - Q(s_t, a_t))$

state, action

learning rate — 0-1 — 不变 —— 固定情景

reward

discount — 降权/加权

bellman

$\max Q = Q(s_{t+1}, a_{t+1})$

rand $\{a_1, a_2, a_3$

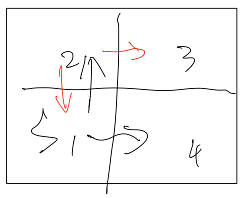
余数多少最大值

当前和最好差多少

$r_t = r(s_t, a_t)$

reward 值 by state & action

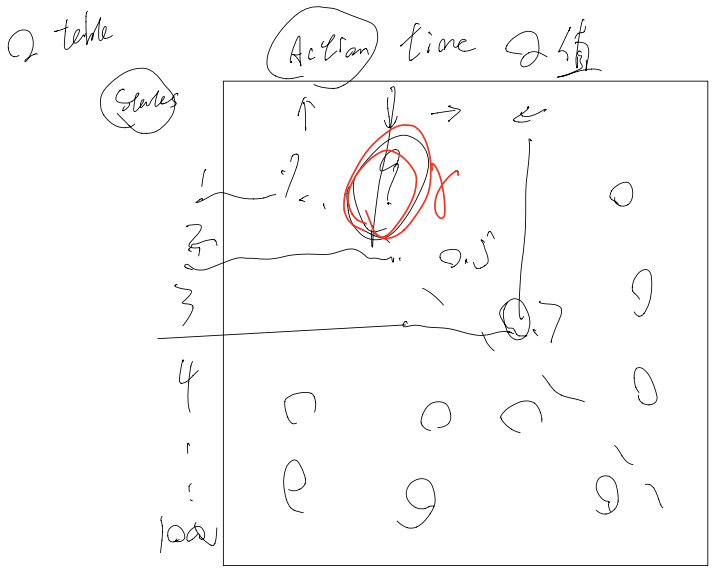
$s_1 a \neq s_2 a$



$S_1(s, \uparrow)$

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

$0 + \alpha R + \alpha \gamma \max + S_1$


Q table
Actions
time
States