action：动作

State：状态

policy：策略函数

$\pi(s,a)$

$\pi(a|s) = P(A=a|S=s)$ 即在当前状态S下选取动作a的概率

随机的原因：如果在S下a是确定的，那么便是熟练工，很容易让对手赢

Reward：奖励

old state $\xrightarrow{\text{action}}$ new action.

$p(S'=s'|S=s, A=a)$：在当前状态S下采取行动a产生新状态S'的概率

①动作是随机的 ②状态转移是随机的

状态 s_1，动作 a_1，奖励 r_1 …… 称为轨迹

Return：从t时刻开始到结束时的奖励之和.

$U_t = R_t + R_{t+1} + \cdots$

discounted return：折扣率回报：

$U_t = R_t + \gamma R_{t+1} + \gamma^2 R_{t+2} \cdots$

动作价值函数（策略π的）

$Q_\pi(s_t, a_t) = E[U_t | S_t = s_t, A_t = a_t]$

因为 U_t 依赖于动作 $A_t, A_{t+1}, A_{t+2} \cdots$ 和状态 $S_t, S_{t+1} \cdots$

最优动作价值函数：

$Q^*(s_t, a_t) = \max_\pi Q_\pi(s_t, a_t)$

状态价值函数：

$V_\pi(s_t) = E_A[Q_\pi(s_t, A)] = \sum_a \pi(a|s_t) Q_\pi(s_t, a)$

$= \int \pi(a|s_t) Q_\pi(s_t, a)\, da.$

其中 $A \sim \pi(\cdot|s_t)$

↓

看当前的状态好不好

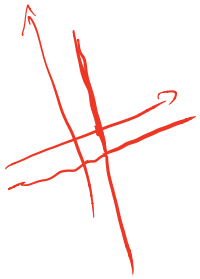
价值学习：

fea

feature

$g(w)$

$h(w)$ -

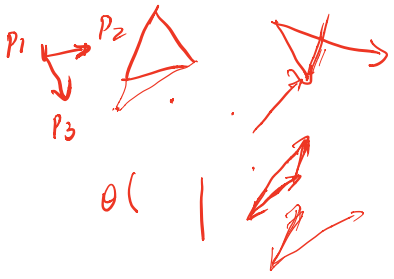


0.-3

1.-2

means -

a.b c.d



a.b

c-a, d-b

x.y

x-a, y-b

$(c-a)*(x-a)$

$+ \vec{(d-b)}(y-b)$

θ(