强化学习

- MDP 马尔可夫决策过程
- 值函数
- 值迭代
- 策略迭代

监督 → 训练集

学习模型

非监督 → 未标记数据

水平监督，比 小

强化学习 奖励信号 → Alpha 下棋 → 对以分配到具体步

MDP $(S, A, \{P_{sa}\}, \gamma, R)$

S - 状态集

A - 动作集

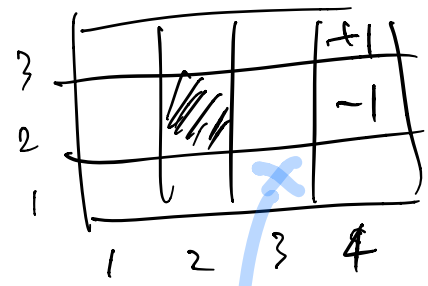
P_{sa} - 状态转移分布

$\sum_{s'} P_{sa}(s') = 1 \quad P_{sa}(s') \geq 0$

γ - discount factor $(0 \leq \gamma \leq 1)$

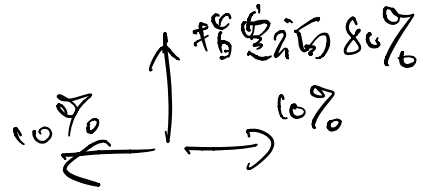
R 梁的强效 $R: S \to R$

例1



S — 11个

$A = \{\uparrow, \downarrow, \leftarrow, \rightarrow\}$



$P_{(3,1),\uparrow}((3,2)) = 0.8$

$P_{(3,1),\uparrow}((4,1)) = 0.1$

$P_{(3,1),\uparrow}((2,1)) = 0.1$

$P_{(3,1),\uparrow}(\ldots) = 0$

$R((4,3)) = +1$

$R((4,2)) = -1$

$R(S) = -0.02$ other

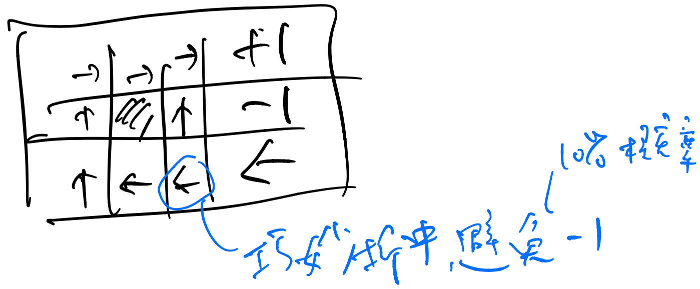
S_0 选 a_0 $S_1 \sim P_{S_0 a_0}$

选 a_1 $S_2 \sim P_{S_1 a_1}$

$R(S_0) + \gamma R(S_1) + \gamma^2 R(S_2) + \gamma^3 R(S_3)$ $0 \le \gamma < 1$

策略 $(a_0, a_1, \ldots)$ 让 $E(\downarrow)$ 最大

Policy $\pi : S \longrightarrow A$



定义 V^π, V^*, π^*

π^* 最佳 P^*

$\forall \pi$, 定义值函数 $V^\pi : S \to \mathbb{R}$

$V^\pi(s)$ = 预期执行 s 总收益

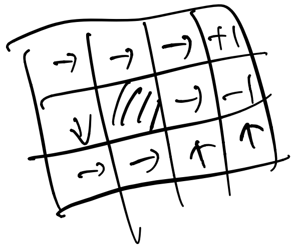
$$= E[R(s_0) + \gamma R(s_1) + \gamma^2 R(s_2) + \cdots \mid \pi, s_0 = s]$$

$$= E[R(s) + \gamma \underbrace{(R(s_1) + \gamma R(s_2) + \cdots)}_{V^\pi(s_1)} \mid \pi, s = s_0]$$

$$V^\pi(s) = R(s_0) + \gamma \sum_{s_1} P_{s_0 \pi(s)}(s_1) V^\pi(s_1)$$

贝尔曼方程

π 固定 $\rightarrow V^\pi$?

P: 

$$V^{\pi}((3,1)) = R((3,1)) + \gamma[0.8V^{\pi}((3,2)) + 0.1V^{\pi}((4,1)) + 0.1V^{\pi}((2,1))]$$

每点...

↓ 11个等式约束 → 解出 V^{π}

$$V^*(s) = \max_{\pi} V^{\pi}(s)$$

V^* 贝尔曼方程

$$V^*(s) = R(s) + \max_a \gamma \sum_{s'} P_{sa}(s') V^*(s')$$

$$\pi^*(s) = \arg\max_a \sum_{s'} P_{sa}(s') V^*(s')$$

↓ 是最优

∴ P^*

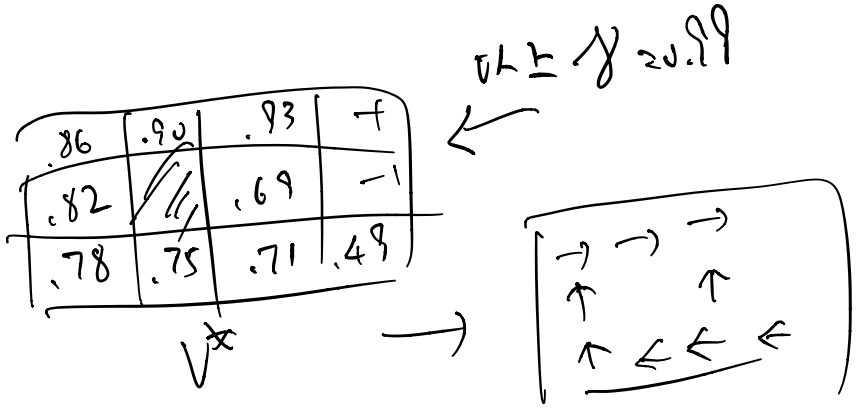
值迭代　用 Bellman 方程求最优

$$V(s) = 0$$

$\forall s$　$V(s) := R(s) + \max_a \gamma \sum_{s'} P_{sa}(s') V(s')$

$V(s) \xrightarrow{\text{收敛}} V^*(s)$　// $V := B(V)$

每次只更新1个，异步 (更快?)



Policy iteration

random π

Repeat: $V := V^{\pi}$ ($V = B(V)$) ← 计算成本 <1000

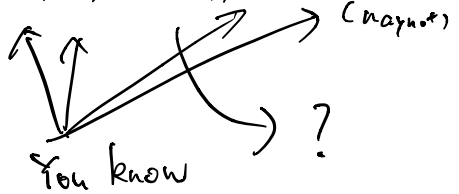
Let $\pi(s) := \arg\max_a \sum_{s_1} P_{sa}(s_1) V(s_1)$

$V \to V^*$ $\pi \to \pi^*$

收敛吗？（... 老师说收敛不知道 但可证收敛，证明不考也默，不写）

如果不知转移概率？

$(S, A, \{P_{sa}\}, \gamma, R)$



$$P_{sa}(s_i) = \frac{\text{到}s_i\text{ by } s \text{ and action } a}{s \text{ and action } a}$$

有限状态处理算法

Repeat {

在MDP中用π action得到经验

update P_{sa} 估计

解 Bellmans 得经验 V [估计]

$\pi := \arg\max_a \sum_{s_i} P_{sa}(s_i) V(s_i)$

}