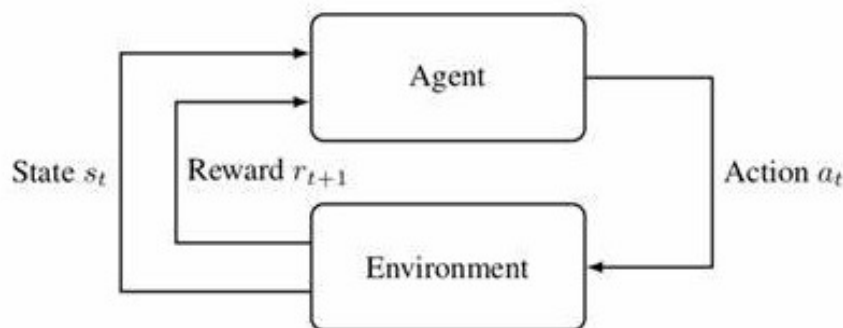


马尔科夫决策过程

Markov Decision Process

马尔可夫决策过程

- 强化学习任务通常使用马尔可夫决策过程（Markov Decision Process，简称MDP）来描述：
- 机器处在一个环境中,每个状态为机器对当前环境的感知;机器只能通过动作来影响环境,当机器执行一个动作后，会使得环境按某种概率转移到另一个状态;环境会根据潜在的奖赏函数反馈给机器一个奖赏。
- 综合而言，强化学习主要包含四个要素：状态、动作、转移概率以及奖赏函数。



马尔科夫性

只要知道现在，将来和过去条件独立

➤ 定义：如果在t时刻的状态 S_t 满足等式： $\mathbb{P}[S_{t+1}|S_t] = \mathbb{P}[S_{t+1}|S_1, \dots, S_t]$

那么这个状态被称为马尔科夫状态，或者说该状态满足马尔科夫性。

➤ 马尔科夫性的要点：

- 状态 S_t 包含了所有历史相关信息
- 或者说历史的所有状态的相关信息都在当前状态 S_t 上体现出来
- 一旦 S_t 知道了，那么 S_1, \dots, S_{t-1} 都可以被抛弃
- 数学上可以认为：状态是将来的充分统计量

示例？

示例

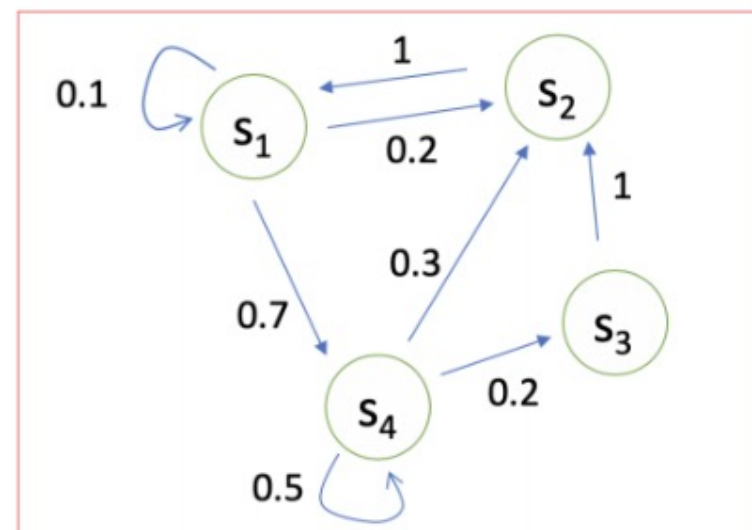
- 下棋时，只关心当前局面
- 打俄罗斯方块时，只关心当前屏幕
-

马尔科夫过程

➤ 马尔科夫过程又叫做马尔科夫链 (Markov Chain), 它是一个无记忆的随机过程, 可以用一个元组 $\langle S, P \rangle$ 表示, 其中:

- S 是有限数量的状态集 $S = \{s_1, s_2, s_3, \dots, s_t\}$
- P 是状态转移概率矩阵 $p(S_{t+1} = s' | s_t = s)$ 其中 s' 表示下一时刻的状态, s 表示当前状态

$$P = \begin{pmatrix} P(s_1|s_1) & P(s_2|s_1) & \dots & P(s_N|s_1) \\ P(s_1|s_2) & P(s_2|s_2) & \dots & P(s_N|s_2) \\ \vdots & \vdots & \ddots & \vdots \\ P(s_1|s_N) & P(s_2|s_N) & \dots & P(s_N|s_N) \end{pmatrix}$$



马尔科夫奖励过程

➤ 马尔科夫奖励过程是在马尔科夫过程基础上增加了奖励函数 R 和衰减系数 γ ，用 $\langle S, R, P, \gamma \rangle$ 表示

○ R ：表示 S 状态下某一时刻的状态 S_t 在下一个时刻（ $t + 1$ ）能获得的奖励的期望

$$R_s = E[R_{t+1}|S_t = s]$$

○ G_t ：收获 G_t 为在一个马尔科夫奖励链上从 t 时刻开始往后所有的奖励的有衰减的收益总和

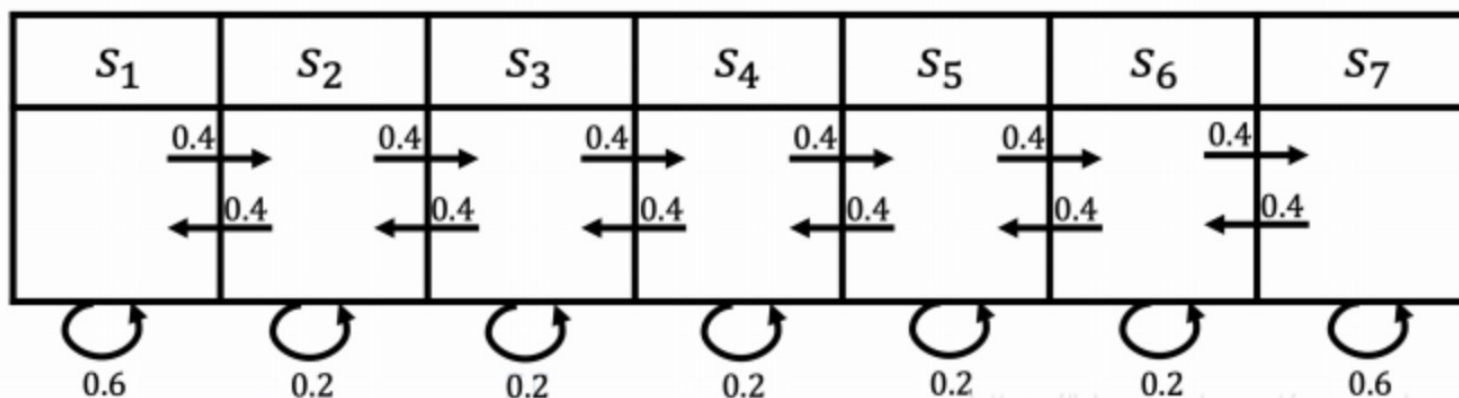
$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots + \gamma^{T-t-1} R_T$$

○ γ ：折扣因子（discount factor $\in [0, 1]$ ）

示例

例子：对于如下状态，我们设定进入 S_1 状态奖励为 5，进入 S_7 状态奖励为 10，其余状态奖励为 0。则 R 可以如下表示： $R = [5, 0, 0, 0, 0, 0, 10]$ ，折扣因子 γ 为 0.5。则对于下面两个马尔可夫过程获得的奖励为：

- $S_4, S_5, S_6, S_7 : 0 + 0.5 \times 0 + 0.5 \times 0 + 0.125 \times 10 = 1.25$
- $S_4, S_3, S_2, S_1 : 0 + 0.5 \times 0 + 0.25 \times 0 + 0.125 \times 5 = 0.625$



Bellman Equation 贝尔曼方程

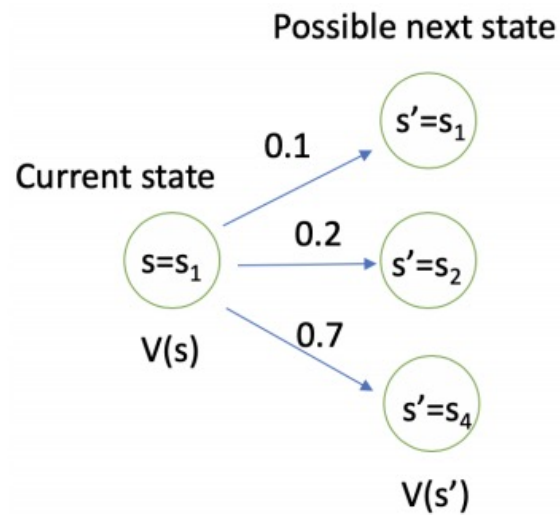
$$\begin{aligned}v(s) &= E[G_t | S_t = s] \\&= E[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots | S_t = s] \\&= E[R_{t+1} + \gamma(R_{t+2} + \gamma R_{t+3} + \cdots) | S_t = s] \\&= E[R_{t+1} + \gamma v(S_{t+1}) | S_t = s] \\&= \underbrace{E[R_{t+1} | S_t = s]}_{\text{当前的奖励}} + \underbrace{\gamma E[v(S_{t+1}) | S_t = s]}_{\text{下一时刻状态的价值期望}}\end{aligned}$$

$$V(s) = \underbrace{R(s)}_{\text{Immediate reward}} + \underbrace{\gamma \sum_{s' \in S} P(s'|s) V(s')}_{\text{Discounted sum of future reward}}$$

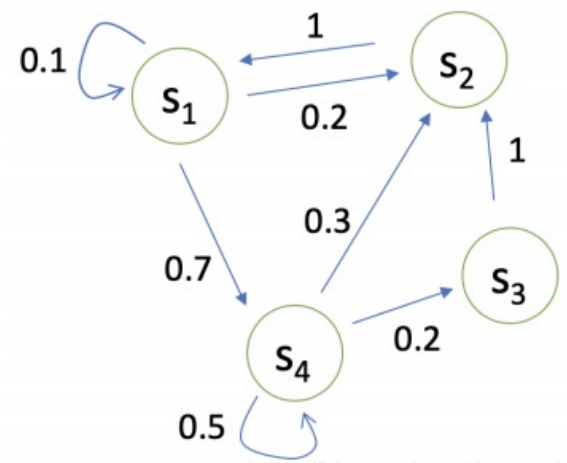
价值函数 $v(s)$ 有两部分组成：一个是当前获得的奖励的期望，即 $R(s)$ ；
另一个是下一时刻的状态期望，即下一时刻可能的状态能获得奖励期望与对应状态转移概率乘积的和

$$V(s) = \underbrace{R(s)}_{\text{Immediate reward}} + \underbrace{\gamma \sum_{s' \in S} P(s'|s)V(s')}_{\text{Discounted sum of future reward}}$$

➤ 状态 s_1 的价值函数为：



Markov Transition matrix



$$V(s_1) = R(s_1) + \gamma(0.1 * V(s_1) + 0.2 * V(s_2) + 0.7 * V(s_4))$$

马尔科夫决策过程 (Markov Decision Process)

➤ 马尔科夫决策过程在马尔科夫奖励过程的基础上加了 Decision 过程，即多了一个动作集合，表示为 $\langle S, A, P, R, \gamma \rangle$ ，其中：

- P 和 R 都与具体的行为 a 对应，而不像马尔科夫奖励过程那样仅对应于某个状态。

- A 表示有限的行为集合

- S 表示有限的状态集合

- P^a is dynamics / transition model for each action

$$P(s_{t+1} = s' | s_t = s, a_t = a) = P[S_{t+1} = s' | S_t = s, A_t = a]$$

- R 是奖励函数

$$R(s_t = s, a_t = a) = E[R_t | s_t = s, a_t = a]$$

马尔科夫决策过程-策略 (policy)

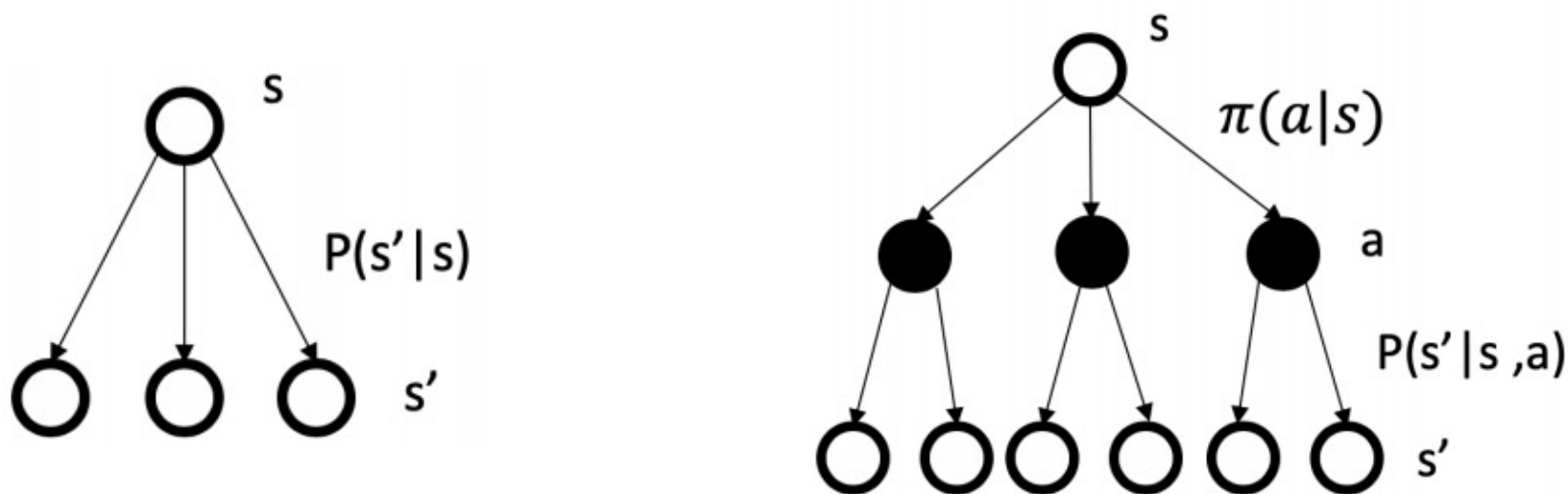
➤ 用 π 表示策略的集合，元素 $\pi(a|s)$ 表示某一状态 s 采取可能的行为 a 的概率

$$\pi(a|s) = P(A_t = a | S_t = s)$$

- Policy完整定义个体行为方式，即包括了个体在各状态下的所有行为和概率
- 同时某一确定的Policy是静态的，与时间无关
- Policy仅和当前的状态有关，与历史信息无关，但是个体可以随着时间更新策略

状态转移概率: $P^\pi(s'|s) = \sum_{a \in A} \pi(a|s) P(s'|s, a)$

奖励函数: $R^\pi(s) = \sum_{a \in A} \pi(a|s) R(s, a)$



策略可以理解[为行动指南](#)，更加规范地描述个体的行为。

既然有了行动指南，就要判断行动指南的价值，故而引入[基于策略的价值函数](#)。

根据策略，当前状态 s 可以到多个其他状态； \longrightarrow **状态价值**

当前状态 s 可以经过某个动作 a 到达某个状态； \longrightarrow **行为价值**

基于策略的状态价值函数 (state value function)

➤ $V(s)$ 表示从状态 s 开始，遵循当前策略时所获得的收获的期望

$$v_{\pi}(s) = E_{\pi}[G_t | S_t = s]$$

- G_t 参照马科夫奖励过程。

➤ 如果状态 s 是一个好的状态，如何选择动作到达这个状态，这时就需要判断动作的好坏，衡量行为价值。


- 行为价值函数

基于策略的行为价值函数 (action value function)

➤ $q_{\pi}(s, a)$: 当前状态 s 执行某一具体行为 a 所能得到的收获的期望

$$q_{\pi}(s, a) = E_{\pi}[G_t | S_t = s, A_t = a]$$

Bellman 公式推导

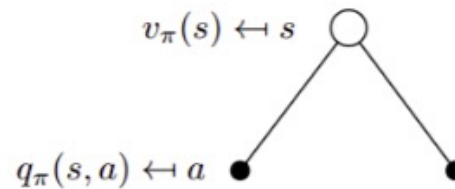

$$q_{\pi}(s, a) = R(s, a) + \gamma \sum_{s' \in S} P(s' | s, a) \cdot V_{\pi}(s')$$

在某一个状态下采取某一个行为的价值，可以分为两部分：
其一是离开这个状态的价值，
其二是所有进入新的状态的价值于其转移概率乘积的和。

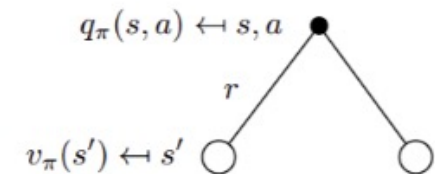
➤ 由状态价值函数和行为价值函数的定义

- 策略用来描述各个不同状态下执行各个不同行为的概率
- 状态价值是遵循当前策略时所获得的收获的期望，即状态 s 的价值体现为在该状态下遵循某一策略而采取所有可能行为的价值按行为发生概率的乘积求和。

$$v_{\pi}(s) = \sum_{a \in A} \pi(a|s) \cdot q_{\pi}(s, a)$$



$$v_{\pi}(s) = \sum_{a \in A} \pi(a|s) q_{\pi}(s, a)$$



$$q_{\pi}(s, a) = \mathcal{R}_s^a + \gamma \sum_{s' \in S} \mathcal{P}_{ss'}^a v_{\pi}(s')$$

https://blog.csdn.net/mobvember_chopin

$$q_{\pi}(s, a) = \mathcal{R}(s, a) + \gamma \sum_{s' \in S} P_{(s'|s,a)} \cdot \sum_{a' \in A} \pi(a'|s') \cdot q_{\pi}(s', a')$$

$$v_{\pi}(s) = \sum_{a \in A} \pi(a|s) \left(\mathcal{R}(s, a) + \gamma \sum_{s' \in S} P_{(s'|s,a)} \cdot v_{\pi}(s') \right)$$

最优价值函数

- 解决强化学习问题即要寻找一个最优的策略让个体在与环境交互过程中获得始终比其它策略都要多的收获，这个最优策略为 π_* 。
- 一般来说，比较难去找到一个全局最优策略，但是可以通过比较若干不同策略的优劣来确定一个较好的策略，也就是局部最优解。
- 一般是通过对应的价值函数来比较策略的优劣，也就是说，寻找较优策略可以通过寻找较优的价值函数来完成。

➤ **最优状态价值函数**是所有策略下产生的众多状态价值函数中的最大者，即： $V_*(s) = \max_{\pi} V_{\pi}(s)$

➤ **最优动作价值函数**是所有策略下产生的众多动作状态价值函数中的最大者，即： $q_*(s, a) = \max_{\pi} q_{\pi}(s, a)$

○ 最大化最优行为价值函数

$$\pi^*(a|s) = \begin{cases} 1, & \text{if } a = \operatorname{argmax} q_*(s, a) \\ 0, & \text{otherwise} \end{cases}$$

➤ 当到达最优的时候，一个状态的价值就等于在当前状态下最大的那个动作价值

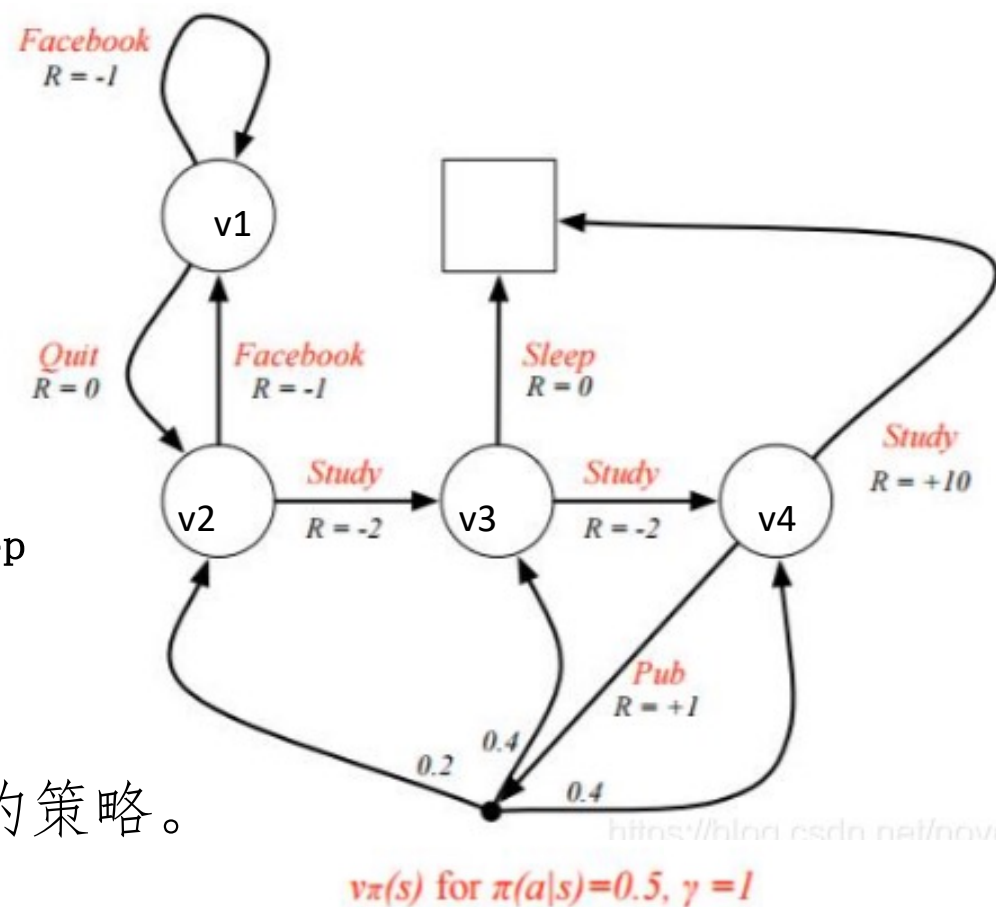
$$q_*(s, a) = R(s, a) + \gamma \sum_{s' \in S} P_{s's}^a \cdot V_*(s')$$

$$v_*(s) = \max_a (R(s, a) + \gamma \sum_{s' \in S} P_{s's}^a \cdot v_*(s'))$$

$$q_*(s, a) = R(s, a) + \gamma \sum_{s' \in S} P_{s's}^a \cdot \max_{a'} q_*(s', a')$$

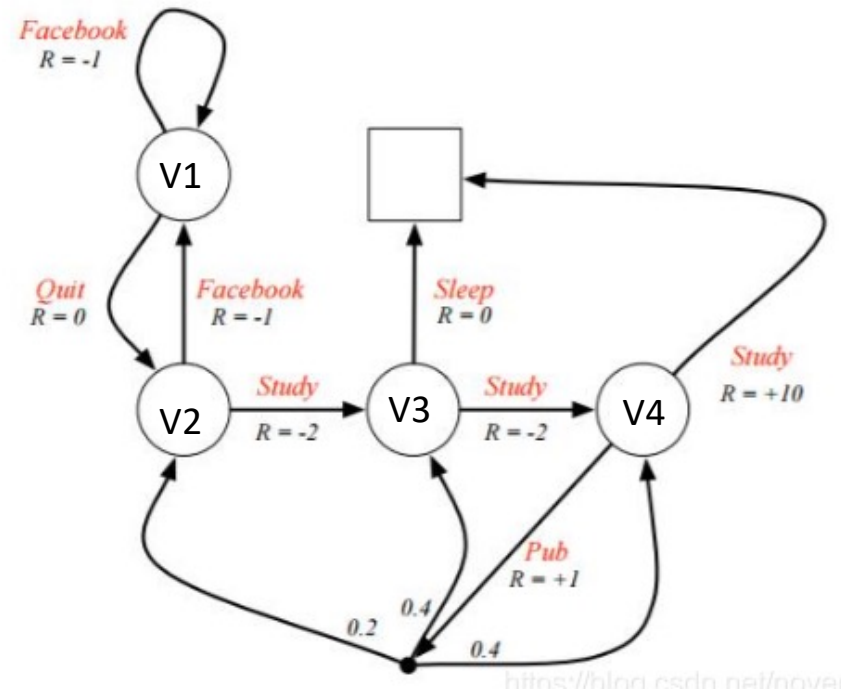
示例

- 左下圆圈位置是起点
- 方框那个位置是终点
- 动作有 study, pub, facebook, quit, sleep
- 动作的即时奖励R已经标出
- 目标是找到最优的动作价值函数或者状态价值函数，进而找出最优的策略。



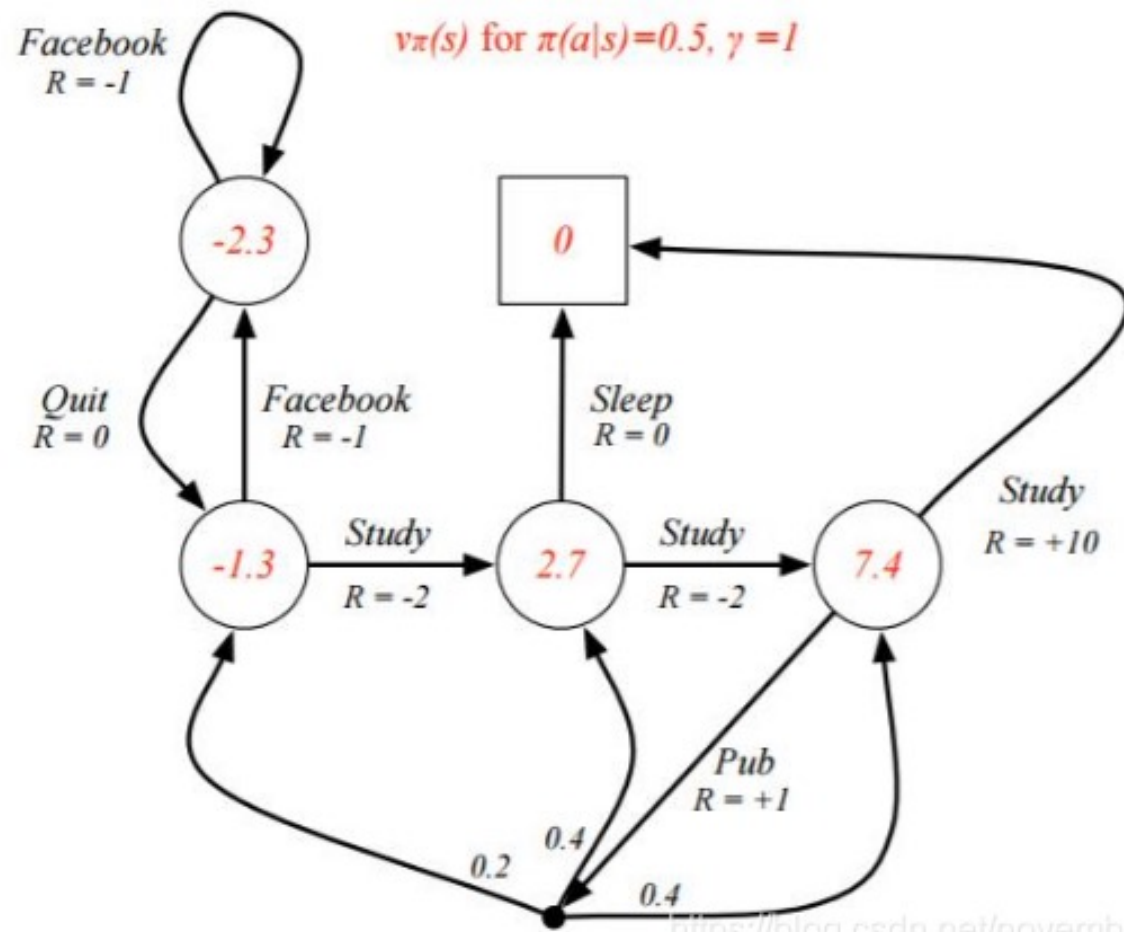
$$v_{\pi}(s) = \sum_{a \in A} \pi(a|s) \left(R(s, a) + \gamma \sum_{s' \in S} P(s'|s, a) \cdot v_{\pi}(s') \right)$$

- 对于 v_1 : $v_1 = 0.5 * (-1 + v_1) + 0.5 * (0 + v_2)$
- 对于 v_2 : $v_2 = 0.5 * (-1 + v_1) + 0.5 * (-2 + v_3)$
- 对于 v_3 : $v_3 = 0.5 * (0 + 0) + 0.5 * (-2 + v_4)$
- 对于 v_4 : $v_4 = 0.5 * (10 + 0) + 0.5 * (1 + 0.2 * v_2 + 0.4 * v_3 + 0.4 * v_4)$



$v_{\pi}(s)$ for $\pi(a|s)=0.5, \gamma=1$

$$V_1 = -2.3; \quad v_2 = -1.3; \quad v_3 = 2.7; \quad v_4 = 7.4$$



求最优的动作价值函数 $V_*(s)$ $q_*(s, a) = R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a \max_{a'} q_*(s', a')$

$$q_*(s_4, \text{study}) = 10$$

$$q_*(s_4, \text{pub}) = 1 + 0.2 * \max_{a'} q_*(s_2, a') + 0.4 * \max_{a'} q_*(s_3, a') + 0.4 * \max_{a'} q_*(s_4, a')$$

$$q_*(s_3, \text{sleep}) = 0$$

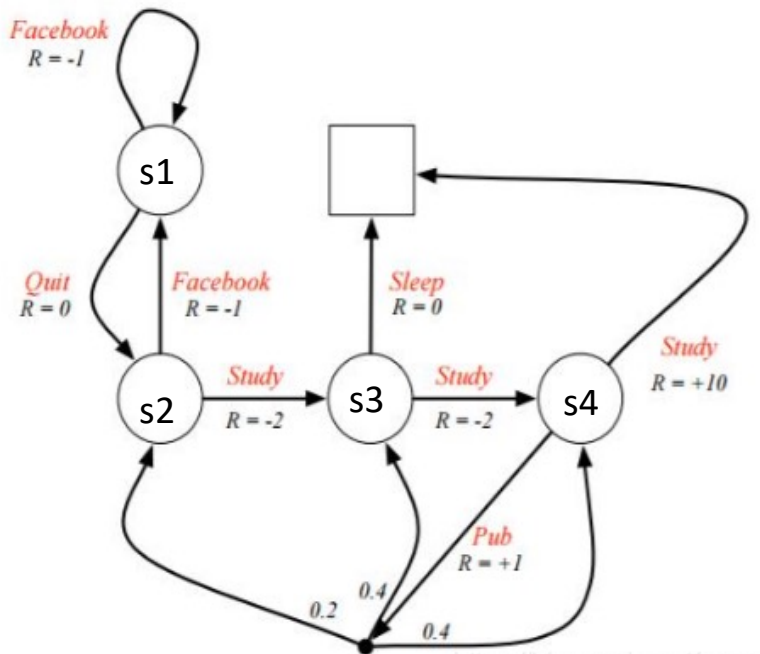
$$q_*(s_3, \text{study}) = -2 + \max_{a'} q_*(s_4, a')$$

$$q_*(s_2, \text{study}) = -2 + \max_{a'} q_*(s_3, a')$$

$$q_*(s_2, \text{facebook}) = -1 + \max_{a'} q_*(s_1, a')$$

$$q_*(s_1, \text{facebook}) = -1 + \max_{a'} q_*(s_1, a')$$

$$q_*(s_1, \text{quit}) = 0 + \max_{a'} q_*(s_2, a')$$



$v_\pi(s)$ for $\pi(a|s)=0.5$, $\gamma=1$

