# 强化学习

## Reinforcement learning

## 第二节

## 马尔可夫决策过程

张世周

# Outlines

- 2.1 马尔可夫过程
- 2.2 马尔可夫收益过程
- 2.3 马尔可夫决策过程
- 2.4 关于MDPs的扩展

# 介绍马尔可夫决策过程

- 马尔可夫决策过程形式化地描述了强化学习的环境
- 这个环境是可以被完全观察的（当前状态完全描述了该过程）
- 几乎所有的强化学习问题都可以形式化为马尔可夫决策过程，例如：
  - 最优控制问题主要是连续MDPs
  - 部分可观测问题可以被转换成MDPs
  - 多臂赌博机是只有一个状态的MDP

# 2.1 马尔可夫过程

**（1）随机变量**

随机变量是指可以随机地取不同值的变量。

**（2）概率分布**

用来描述随机变量在每个可能取到的值处的可能性大小。

**（3）条件概率**

策略Π（a|s）是条件概率。

**（4）期望**

函数f(x)关于某分布P(x)的期望是指，当x由分布P(x)产生、f作用于x时，f(x)的平均值。

$$E_{x\sim P}[f(x)] = \sum_{x} P(x)f(x) \qquad E_{x\sim P}[f(x)] = \int p(x)f(x)dx$$

# 2.1 马尔可夫过程

## （5）随机过程

**定义2.1.1** 设$(\Omega, \mathscr{F}, P)$是一个概率空间，$T$是一个实的参数集，定义在$\Omega$和$T$上的二元函数$X(\omega,t)$，如果对于任意固定的$t \in T$，$X(\omega,t)$是$(\Omega, \mathscr{F}, P)$上的随机变量，则称$\{X(\omega,t), \omega \in \Omega, t \in T\}$为该概率空间上的随机过程，简记为$\{X(t), t \in T\}$.

https://www.countbayesie.com/blog/2015/8/30/picture-guide-to-probability-spaces

# 2.1 马尔可夫过程

## （5）随机过程

**定义2.1.2** 设$\{X(t), t \in T\}$是随机过程，则当$t$固定时，$X(t)$是一个随机变量，称之为$\{X(t), t \in T\}$在$t$时刻的状态. 随机变量$X(t)$($t$固定，$t \in T$)所有可能的取值构成的集合，称为随机过程的**状态空间**，记为$S$.

**定义2.1.3** 设$\{X(t), t \in T\}$是随机过程，则当$\omega \in \Omega$固定时，$X(t)$是定义在上$T$不具有随机性的普通函数，记为，$x(t)$,称为随机过程的一个**样本函数**。其图像成为随机过程的一条**样本曲线**（轨道或实现）。

# 2.1 马尔可夫过程

## （5）随机过程

**例2.2.2** 设有一质点在$x$轴上作随机游动，在$t=0$时质点处于$x$轴的原点$O$，在$t=1,2,\ldots$时质点可以在$x$轴上正向或反向移动一个单位，作正向移动一个单位的概率为$p$，作反向移动一个单位的概率为$q=1-p$，在$t=n$时，质点所处的位置为$X_n$，则$\{X_n, n=1,2,\ldots\}$为一随机过程，其参数集$T=\{0,1,2,\ldots\}$，状态空间$S=\{\ldots,-2,-1,0,1,2,\ldots\}$。

# 2.1 马尔可夫过程

## 马尔可夫性

未来只与现在的状态有关和过去无关

定义：

如果一个状态 $S_t$ 拥有马尔可夫性，则当且仅当满足以下公式

$$P[S_{t+1} \mid S_t] = P[S_{t+1} \mid S_1, \ldots, S_t]$$

- 状态从历史中获取所有相关信息
- 一旦知晓这个状态，历史就可以被扔掉
- 换句话说就是状态是未来的充分统计数字

## 状态转移矩阵

对于一个具有马尔可夫性的状态 $S$ 和他的后续状态 $S'$ 来说，状态转移概率定义为 $P_{ss'} = P[S_{t+1} = s' \mid S_t = s]$

状态转移矩阵 $P$ 定义了从所有状态 $S$ 到所有后续状态 $S'$ 的转移概率，

$$\mathcal{P} = \begin{matrix} \\ from \end{matrix} \begin{matrix} \overset{to}{} \\ \begin{bmatrix} \mathcal{P}_{11} & \dots & \mathcal{P}_{1n} \\ \vdots & & \\ \mathcal{P}_{n1} & \dots & \mathcal{P}_{nn} \end{bmatrix} \end{matrix}$$

每行的概率之和为1

## 马尔可夫链

马尔可夫链是一个无记忆的随机过程，即一系列的随机状态S1，S2，...

具有马尔可夫性质

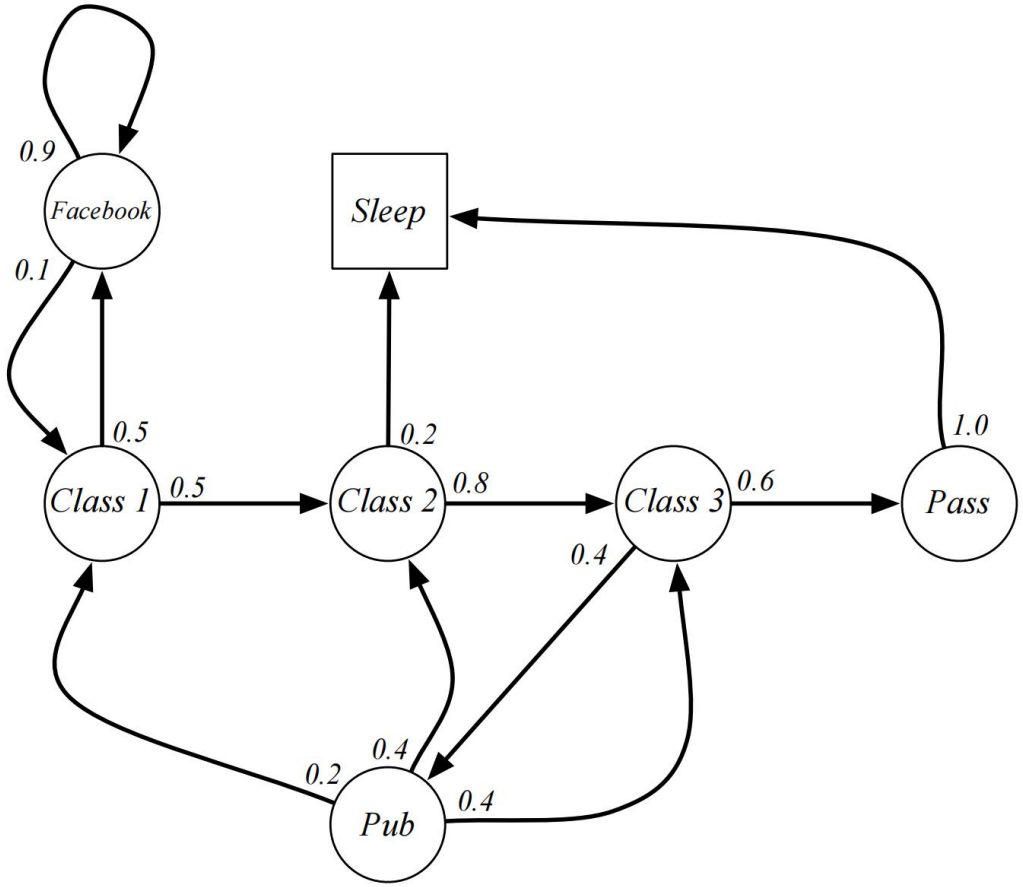
**定义**

一个马尔可夫链是一个包含 $< S, P >$的元组

- $S$ 是一个有限状态集
- $P$ 是一个状态转移矩阵,其中 $P_{ss'} = P[S_{t+1} = s' \mid S_t = s]$

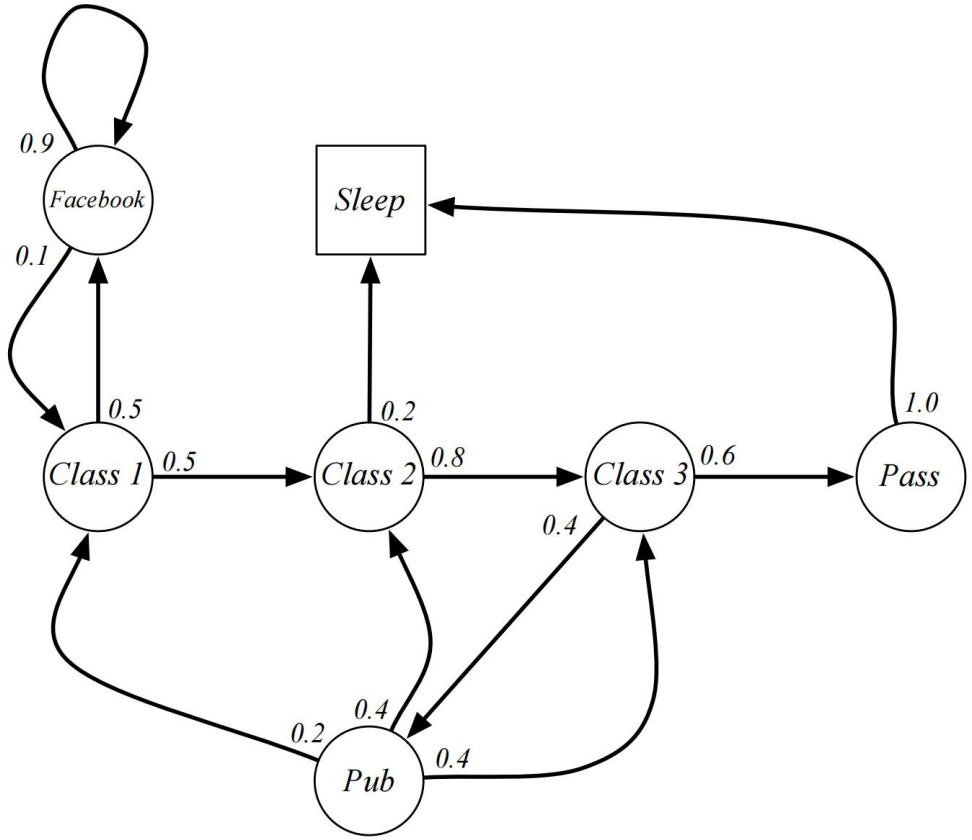
# 2.1 马尔可夫过程

## 例：学生马尔可夫链



# 2.1 马尔可夫过程

- 学生马尔可夫链采样



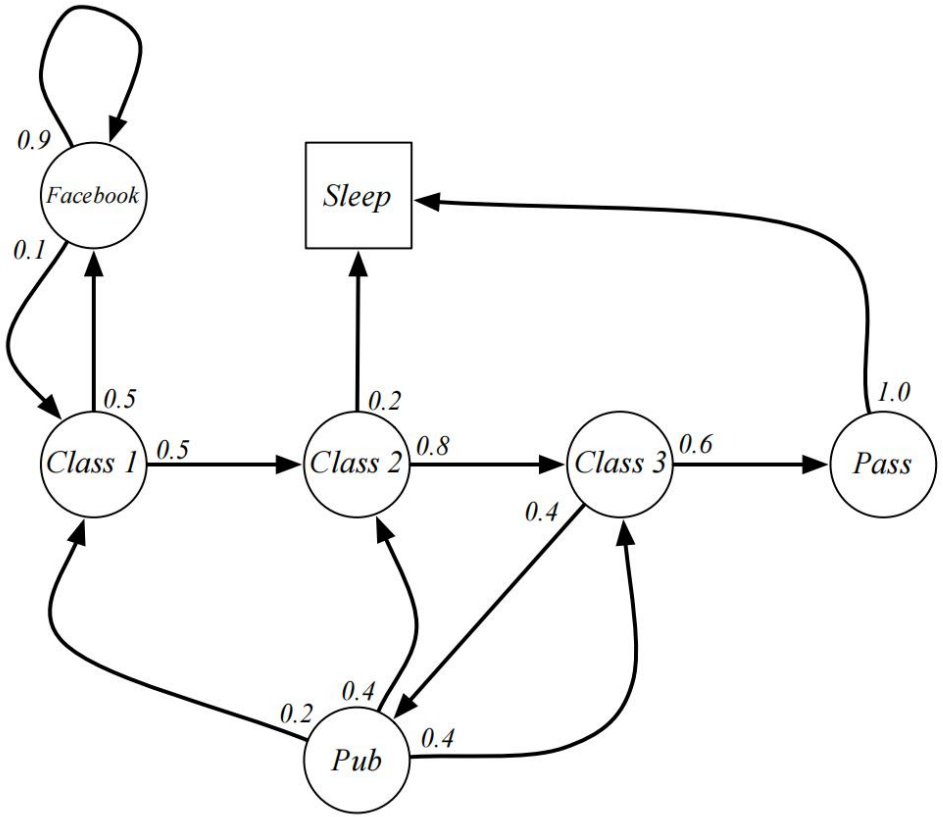

学生马尔可夫链样本(episode)从Class1开始令S1=C1

$$S_1, S_2, \ldots, S_T$$

- C1 C2 C3 Pass Sleep
- C1 FB FB C1 C2 Sleep
- C1 C2 C3 Pub C2 C3 Pass Sleep
- C1 FB FB C1 C2 C3 Pub C1 FB FB FB C1 C2 C3 Pub C2 Sleep

# 2.1 马尔可夫过程

## 学生马尔可夫链状态转移矩阵




$$\mathcal{P} = \begin{array}{c} \\ C1 \\ C2 \\ C3 \\ Pass \\ Pub \\ FB \\ Sleep \end{array} \begin{array}{c} \begin{array}{ccccccc} C1 & C2 & C3 & Pass & Pub & FB & Sleep \end{array} \\ \begin{bmatrix} & 0.5 & & & & 0.5 & \\ & & 0.8 & & & & 0.2 \\ & & & 0.6 & 0.4 & & \\ & & & & & & 1.0 \\ 0.2 & 0.4 & 0.4 & & & & \\ 0.1 & & & & & 0.9 & \\ & & & & & & 1 \end{bmatrix} \end{array}$$

# 马尔可夫收益过程（MRP）

马尔可夫收益过程是一个带有状态价值的马尔可夫链

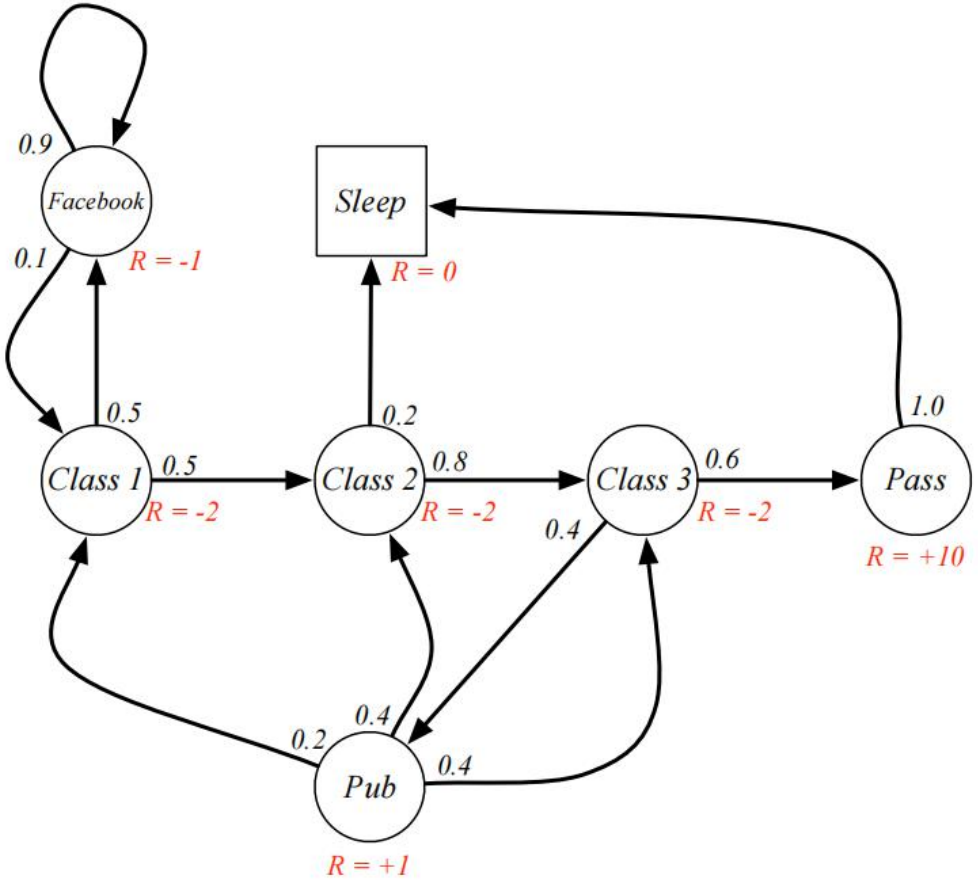
**定义**

一个马尔可夫收益过程是一个包含 $< S, P, R, \gamma >$ 的元组

- $S$ 是一个有限状态集
- $P$ 是一个状态转移矩阵, $P_{ss'} = P[S_{t+1} = s' \mid S_t = s]$
- $R$ 是一个收益函数, $R_S = E[R_{t+1} \mid S_t = s]$
- $\gamma$ 是一个折扣率 $\gamma \in [0,1]$

# 2.2马尔可夫收益过程

## 例：学生马尔可夫收益过程



# 2.2马尔可夫收益过程

## 回报

**定义**

回报是从t时刻开始的总折扣奖励，记为 $G_t$

$$G_t = R_{t+1} + \gamma R_{t+2} + \ldots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$

- 折扣 $\gamma \in [0,1]$ 是未来收益的当前价值
- K+1时刻的收益 $R$ 在当前的价值是 $\gamma^k R$
- 这种价值观是即时回报高于延迟回报。当 $\gamma$ 为0时，对未来的价值没有考虑；当 $\gamma$ 为1时，未来的价值等同于现在的价值

# 2.2马尔可夫收益过程

为什么会有折扣率？

- 从数学上来说，使用折扣率很方便
- 避免循环马尔可夫过程中的无限回报
- 未来具有不确定性
- 如果奖励是经济性的，那么即时奖励可能比延迟奖励获得更多的利息
- 动物/人类的行为表现出对即时奖励的偏好
- 有时可以使用无折扣率的马尔可夫奖励过程（即$\gamma=1$），例如，所有序列都是可以被完全预料到的

# 价值函数

## 状态s的长期价值由价值函数v(s)来表示

**定义**

**MRP的状态价值函数v(s)是状态s回报的期望**

$$v(s) = E[G_t \mid S_t = s]$$

# 2.2马尔可夫收益过程

## 例：学生MRP回报

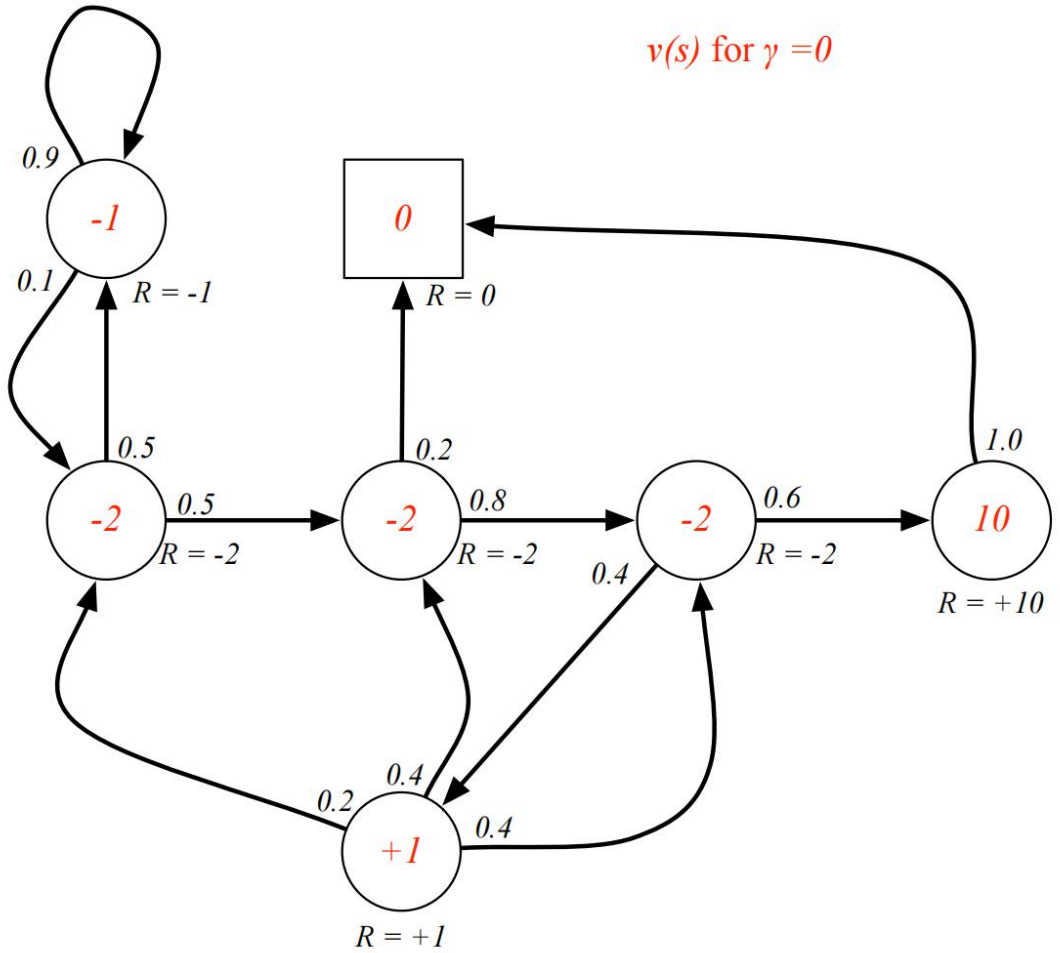
学生MRP的样本回报从Class1开始，令S1=C1且 $\gamma = \frac{1}{2}$

$$G_1 = R_2 + \gamma R_3 + \ldots + \gamma^{T-2} R_T$$

| | | | |
|---|---|---|---|
| C1 C2 C3 Pass Sleep | $v_1 = -2 - 2 * \frac{1}{2} - 2 * \frac{1}{4} + 10 * \frac{1}{8}$ | $=$ | $-2.25$ |
| C1 FB FB C1 C2 Sleep | $v_1 = -2 - 1 * \frac{1}{2} - 1 * \frac{1}{4} - 2 * \frac{1}{8} - 2 * \frac{1}{16}$ | $=$ | $-3.125$ |
| C1 C2 C3 Pub C2 C3 Pass Sleep | $v_1 = -2 - 2 * \frac{1}{2} - 2 * \frac{1}{4} + 1 * \frac{1}{8} - 2 * \frac{1}{16} \ldots$ | $=$ | $-3.41$ |
| C1 FB FB C1 C2 C3 Pub C1 ... FB FB FB C1 C2 C3 Pub C2 Sleep | $v_1 = -2 - 1 * \frac{1}{2} - 1 * \frac{1}{4} - 2 * \frac{1}{8} - 2 * \frac{1}{16} \ldots$ | $=$ | $-3.20$ |

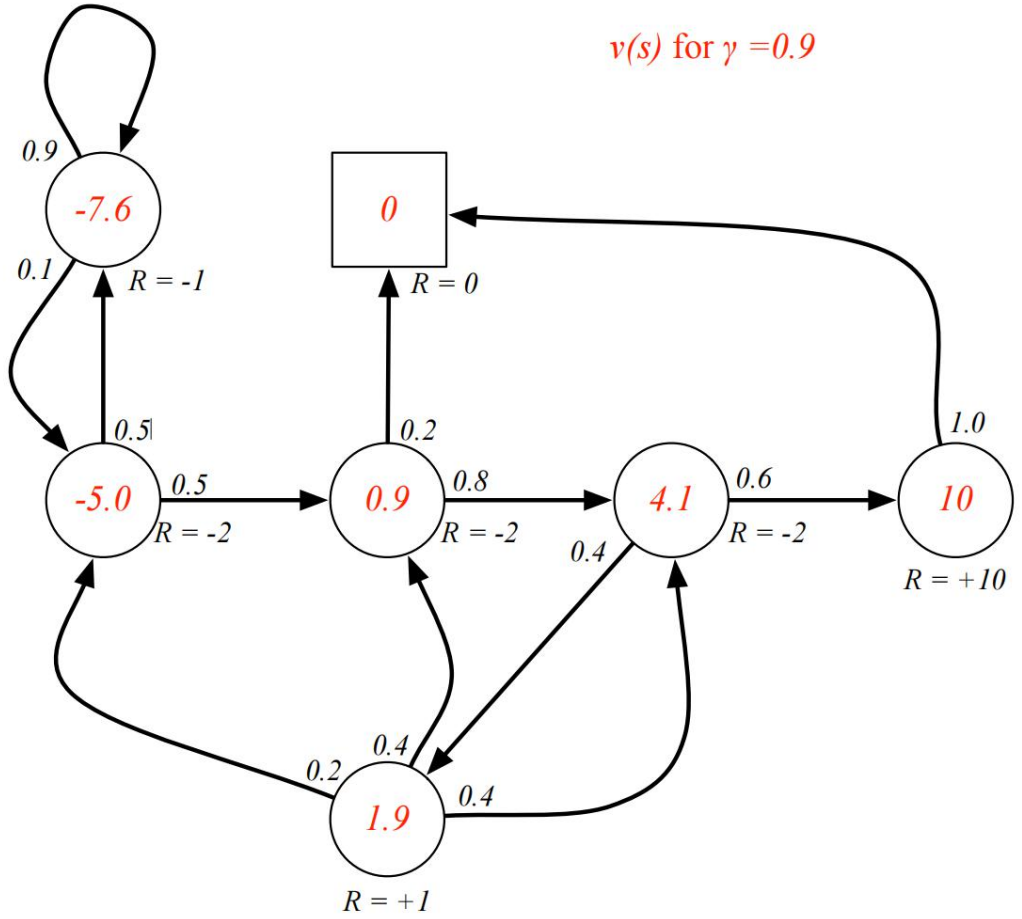
# 2.2马尔可夫收益过程

## 学生MRP状态价值函数（1）



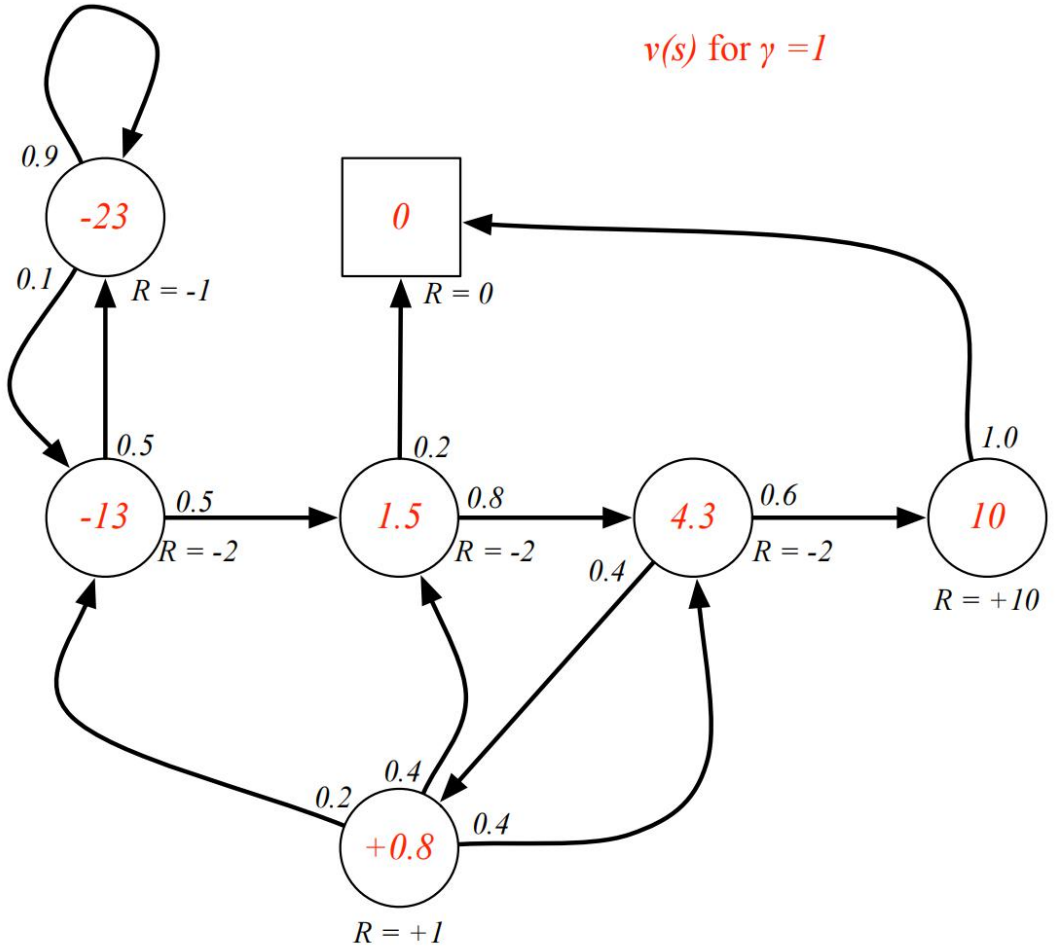
# 2.2马尔可夫收益过程

- 学生MRP状态价值函数（2）



# 2.2马尔可夫收益过程

## 学生MRP状态价值函数（3）



## MRPs的贝尔曼方程

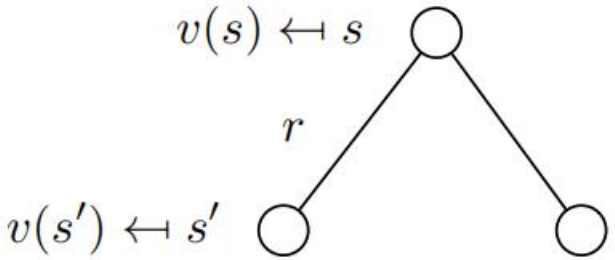
这个价值函数可以被分成两个部分：

- 立即收益 $R_{t+1}$
- 后继状态的折扣价值 $\gamma v(S_{t+1})$

$$
\begin{aligned}
v(s) &= E[G_t \mid S_t = s] \\
&= E[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots \mid S_t = s] \\
&= E[R_{t+1} + \gamma(R_{t+2} + \gamma R_{t+3} + \ldots) \mid S_t = s] \\
&= E[R_{t+1} + \gamma G_{t+1} \mid S_t = s] \\
&= E[R_{t+1} + \gamma v(S_{t+1}) \mid S_t = s]
\end{aligned}
$$

# MRPs的贝尔曼方程（2）

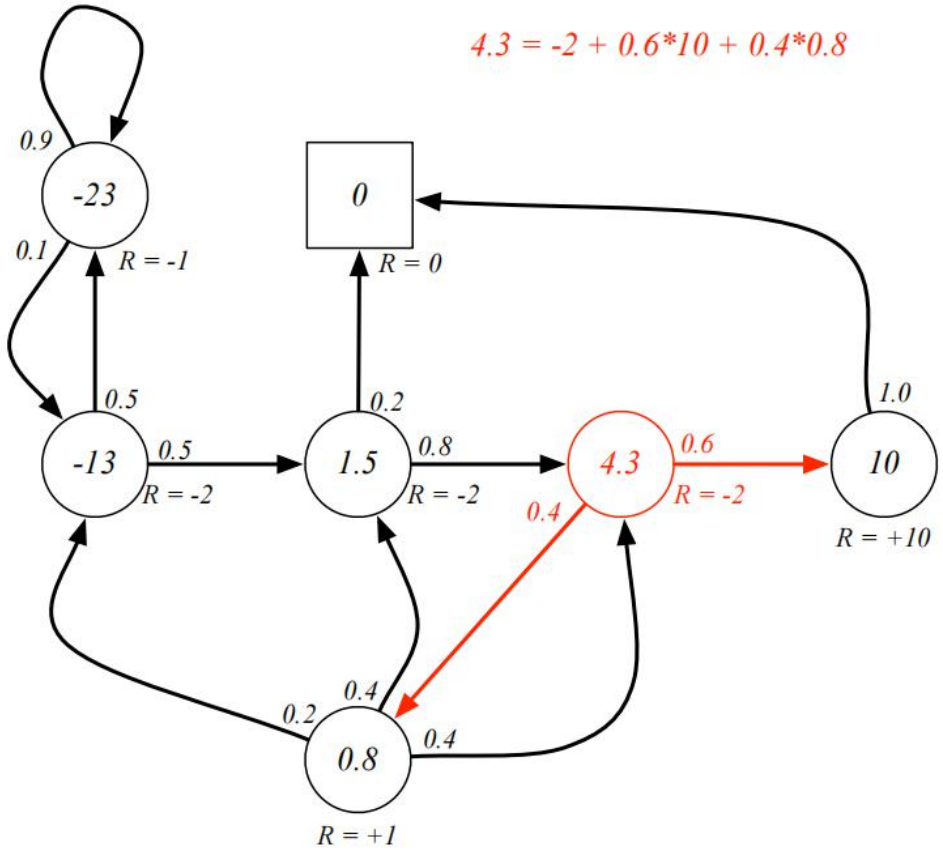
$$v(s) = E[R_{t+1} + \gamma v(s+1) \mid S_t = s]$$




$$v(s) = R_s + \gamma \sum_{s' \in S} P_{ss'} v(s')$$

# 2.2马尔可夫收益过程

## 例：学生MRP的贝尔曼方程



## 矩阵形式的贝尔曼方程

贝尔曼方程可以用矩阵简洁地表示

$$v = R + \lambda P v$$

其中v是一个列向量，每个元素对应着一个状态

$$\begin{bmatrix} v(1) \\ \vdots \\ v(n) \end{bmatrix} = \begin{bmatrix} \mathcal{R}_1 \\ \vdots \\ \mathcal{R}_n \end{bmatrix} + \gamma \begin{bmatrix} \mathcal{P}_{11} & \dots & \mathcal{P}_{1n} \\ \vdots & & \\ \mathcal{P}_{11} & \dots & \mathcal{P}_{nn} \end{bmatrix} \begin{bmatrix} v(1) \\ \vdots \\ v(n) \end{bmatrix}$$

# 2.2马尔可夫收益过程

## 解贝尔曼方程

- 贝尔曼方程是一个线性方程
- 它可以被直接解出来：

$$v = R + \lambda P v$$

$$(I - \gamma P)v = R$$

$$v = (I - \gamma P)^{-1} R$$

- n个状态的计算复杂度为 $O(n^3)$
- 只有小型的MRPs才可以使用直接解法

有许多间接的解贝尔曼方程的方法，比如：

- 动态规划
- 蒙特卡罗评估
- 时间差分学习

# 2.3马尔可夫决策过程

马尔可夫决策过程（MDP）是在马尔可夫收益过程的基础上添加了决策。它的环境中的所有状态都具有马尔可夫性。

**定义**

一个马尔可夫决策过程是一个包含 $< S, A, P, R, \gamma >$ 的元组

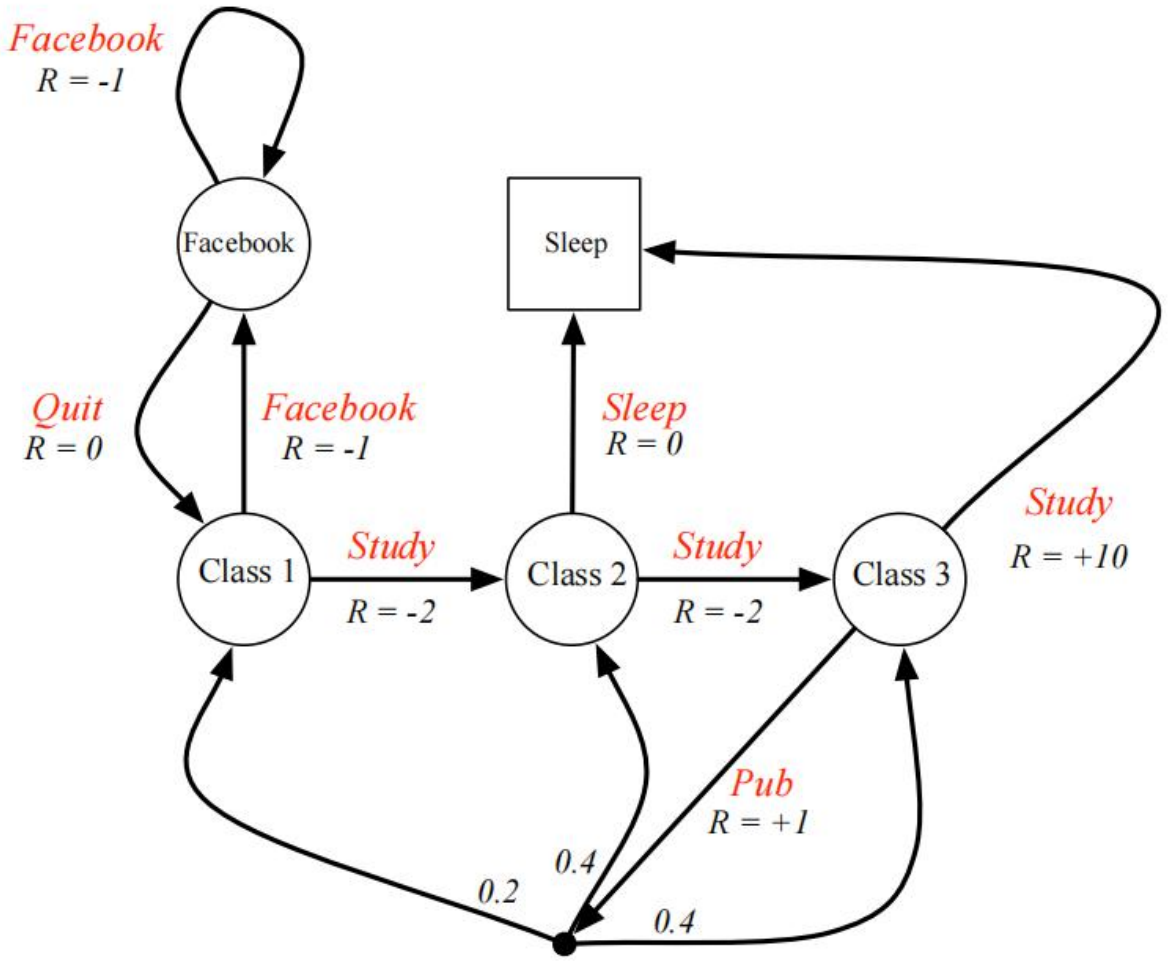
- $S$ 是一个有限状态集
- $A$ 是一个有限动作集
- $P$ 是一个状态转移可能性矩阵

$$P_{ss'}^{a} = P[S_{t+1} = s' \mid S_t = s, A_t = a]$$

- $R$ 是收益函数 $R_s^a = E[R_{t+1} \mid S_t = s, A_t = a]$
- $\gamma$ 是折扣率 $\gamma \in [0,1]$

# 2.3马尔可夫决策过程

- 例：学生MDP



## 策略（1）

**定义**

策略π是在给定状态条件下的动作集上的分布：

$$\pi(a \mid s) = P[A_t = a \mid S_t = s]$$

- 策略完全定义了智能体的行为
- MDP策略依赖当前状态而不是历史
- 换句话说策略是平稳的，独立于时间

$$A_t \sim \pi(\cdot \mid S_t), \forall t > 0$$

# 2.3马尔可夫决策过程

## 策略（2）

- 贪婪策略

$$\pi_*(a|s)=\begin{cases}1 & \text{if } a=\arg\max\limits_{a\in A} q_*(s,a)\\ 0 & \text{otherwise}\end{cases}$$

- ε-greedy策略

$$\pi(a|s)\leftarrow\begin{cases}1-\varepsilon+\dfrac{\varepsilon}{|A(s)|} & if\ a=\text{argmax}_a Q(s,a)\\ \dfrac{\varepsilon}{|A(s)|} & if\ a\neq\text{argmax}_a Q(s,a)\end{cases}$$

- 高斯策略

$$\pi_\theta=\mu_\theta+\varepsilon,\ \varepsilon\sim N(0,\sigma^2)$$

- 玻尔兹曼分布策略

$$\pi(a|s,\theta)=\frac{\exp(Q(s,a,\theta))}{\Sigma_b\exp(h(s,b,\theta))}$$

## 策略（3）

- 给定一个MDP，$M =< S, A, P, R, \gamma >$ 和一个策略 π
- 状态序列S1，S2...是一个马尔可夫过程 $< S, P^{\pi} >$
- 状态和收益序列S1, R2, S2, ...是一个马尔可夫收益过程 $< S, P^{\pi}, R^{\pi}, \gamma >$
- 其中

$$P_{ss'}^{\pi} = \sum_{a \in A} \pi(a \mid s) P_{ss'}^{a}$$

$$R_{s}^{\pi} = \sum_{a \in A} \pi(a \mid s) R_{s}^{a}$$

# 2.3马尔可夫决策过程

## 价值函数

**定义**

MDP的状态价值函数$v_\pi(s)$是从状态s开始，然后跟随策略 π 的期望收益

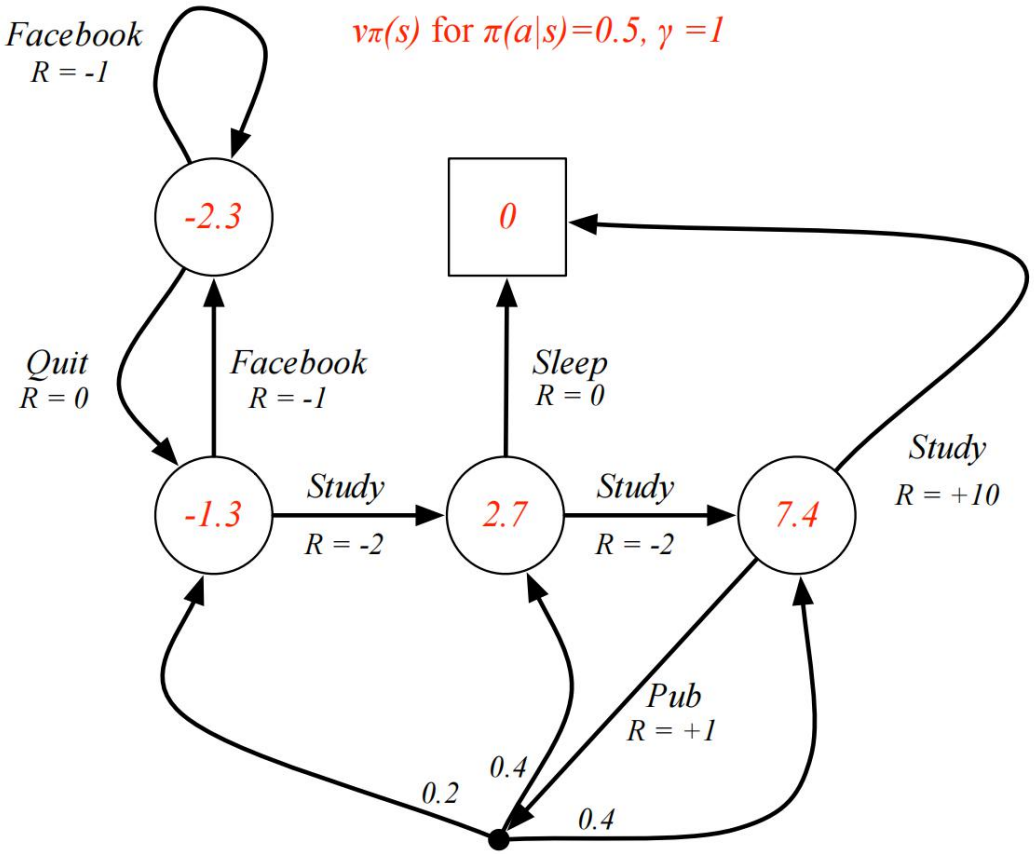
$$v_\pi(s) = E_\pi[G_t \mid S_t = s]$$

**定义**

动作价值函数$q_\pi(s,a)$是从状态s开始，采取动作a，然后遵循策略π的预期收益

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

# 例：学生MDP的状态价值函数



## 贝尔曼期望方程

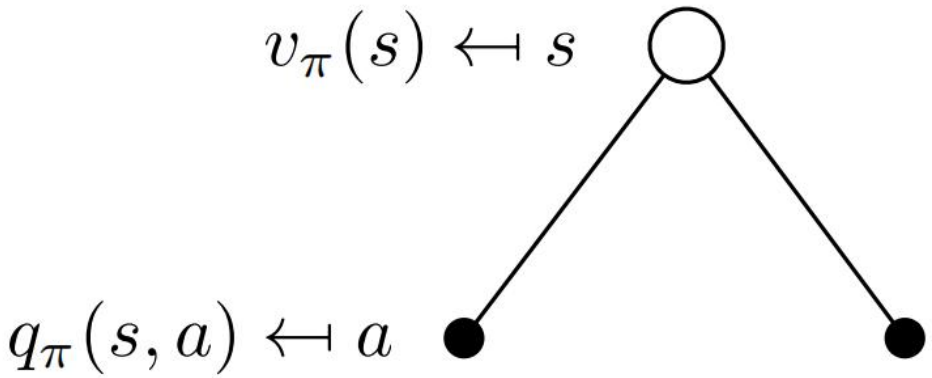
状态价值函数也可以被分解为立即收益加上折扣之后的后继状态的价值

$$v_\pi(s) = E_\pi[R_{t+1} + \gamma v(S_{t+1}) \mid S_t = s]$$

动作价值函数也可以被如此分解

$$q_\pi(s,a) = E_\pi[R_{t+1} + \gamma q_\pi(S_{t+1}, A_{t+1}) \mid S_t = s, A_t = a]$$

## 关于 $v_\pi$ 的贝尔曼期望方程




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

# 2.3马尔可夫决策过程

## 关于 $v_\pi$ 的贝尔曼期望方程

- 用 $t$ 时刻的动作价值函数表示 $t$ 时刻的状态价值函数：

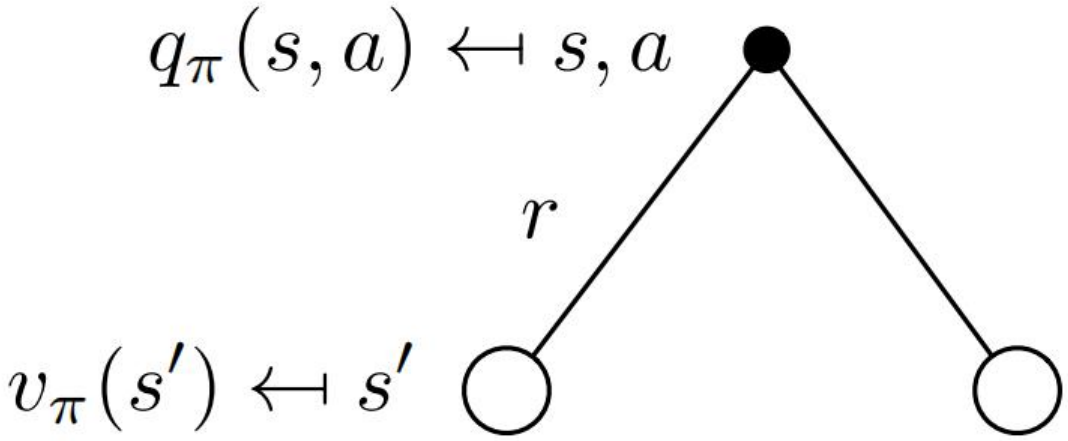
$$v_\pi(s)=\sum_a \pi(a|s)q_\pi(s,a),\quad s\in\mathcal{S}$$

（推导：对任一状态 $s\in\mathcal{S}$，有

$$\begin{aligned}
v_\pi(s)&=\mathrm{E}_\pi[G_t|S_t=s]\\
&=\sum_g g\Pr[G_t=g|S_t=s]\\
&=\sum_g g\sum_a \Pr[G_t=g,A_t=a|S_t=s]\\
&=\sum_g g\sum_a \Pr[A_t=a|S_t=s]\Pr[G_t=g|S_t=s,A_t=a]\\
&=\sum_a \Pr[A_t=a|S_t=s]\sum_g g\Pr[G_t=g|S_t=s,A_t=a]\\
&=\sum_a \Pr[A_t=a|S_t=s]\mathrm{E}_\pi[G_t|S_t=s,A_t=a]\\
&=\sum_a \pi(a|s)q_\pi(s,a)
\end{aligned}$$

# 2.3马尔可夫决策过程

## 关于 $q_\pi$ 的贝尔曼期望方程




$$q_\pi(s,a) = R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a v_\pi(s')$$

# 2.3马尔可夫决策过程

## 关于 $q_\pi$ 的贝尔曼期望方程

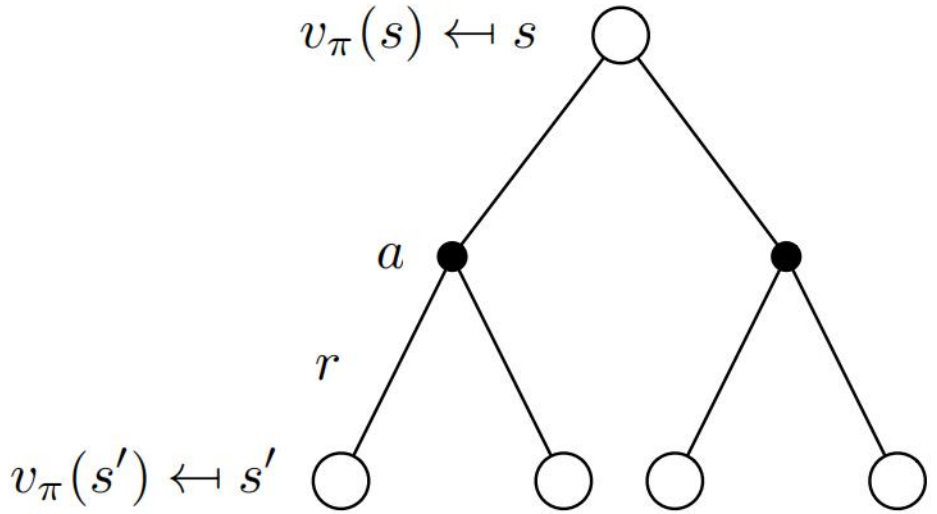
$$
\begin{aligned}
q_\pi(s,a) &= \mathrm{E}_\pi[G_t \mid S_t=s, A_t=a] \\
&= \mathrm{E}_\pi[R_{t+1}+\gamma G_{t+1} \mid S_t=s, A_t=a] \\
&= \mathrm{E}_\pi[R_{t+1} \mid S_t=s, A_t=a] + \gamma \mathrm{E}_\pi[G_{t+1} \mid S_t=s, A_t=a] \\
&= \sum_{s',r} p(s',r \mid s,a)\left[r+\gamma v_\pi(s')\right]
\end{aligned}
$$

(推导：对任意的状态 $s \in \mathcal{S}$ 和动作 $a \in \mathcal{A}$，有

$$
\begin{aligned}
&\mathrm{E}_\pi[G_{t+1} \mid S_t=s, A_t=a] \\
&= \sum_g g \Pr[G_{t+1}=g \mid S_t=s, A_t=a] \\
&= \sum_g g \sum_{s'} \Pr[S_{t+1}=s', G_{t+1}=g \mid S_t=s, A_t=a] \\
&= \sum_g g \sum_{s'} \Pr[S_{t+1}=s' \mid S_t=s, A_t=a] \Pr[G_{t+1}=g \mid S_t=s, A_t=a, S_{t+1}=s'] \\
&= \sum_g g \sum_{s'} \Pr[S_{t+1}=s' \mid S_t=s, A_t=a] \Pr[G_{t+1}=g \mid S_{t+1}=s'] \\
&= \sum_{s'} \Pr[S_{t+1}=s' \mid S_t=s, A_t=a] \sum_g g \Pr[G_{t+1}=g \mid S_{t+1}=s'] \\
&= \sum_{s'} \Pr[S_{t+1}=s' \mid S_t=s, A_t=a] \mathrm{E}_\pi[G_{t+1} \mid S_{t+1}=s'] \\
&= \sum_{s'} p(s' \mid s,a) v_\pi(s')
\end{aligned}
$$

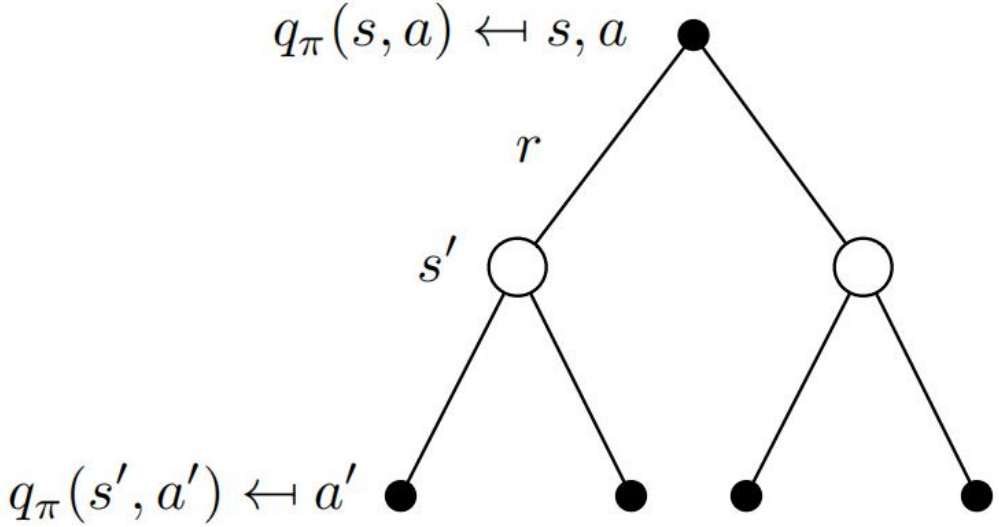
# 2.3马尔可夫决策过程

## 关于 $v_\pi$ 的贝尔曼期望方程（2）




$$v_\pi(s) = \sum_{a \in A} \pi(a \mid s) \left( R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a v_\pi(s') \right)$$

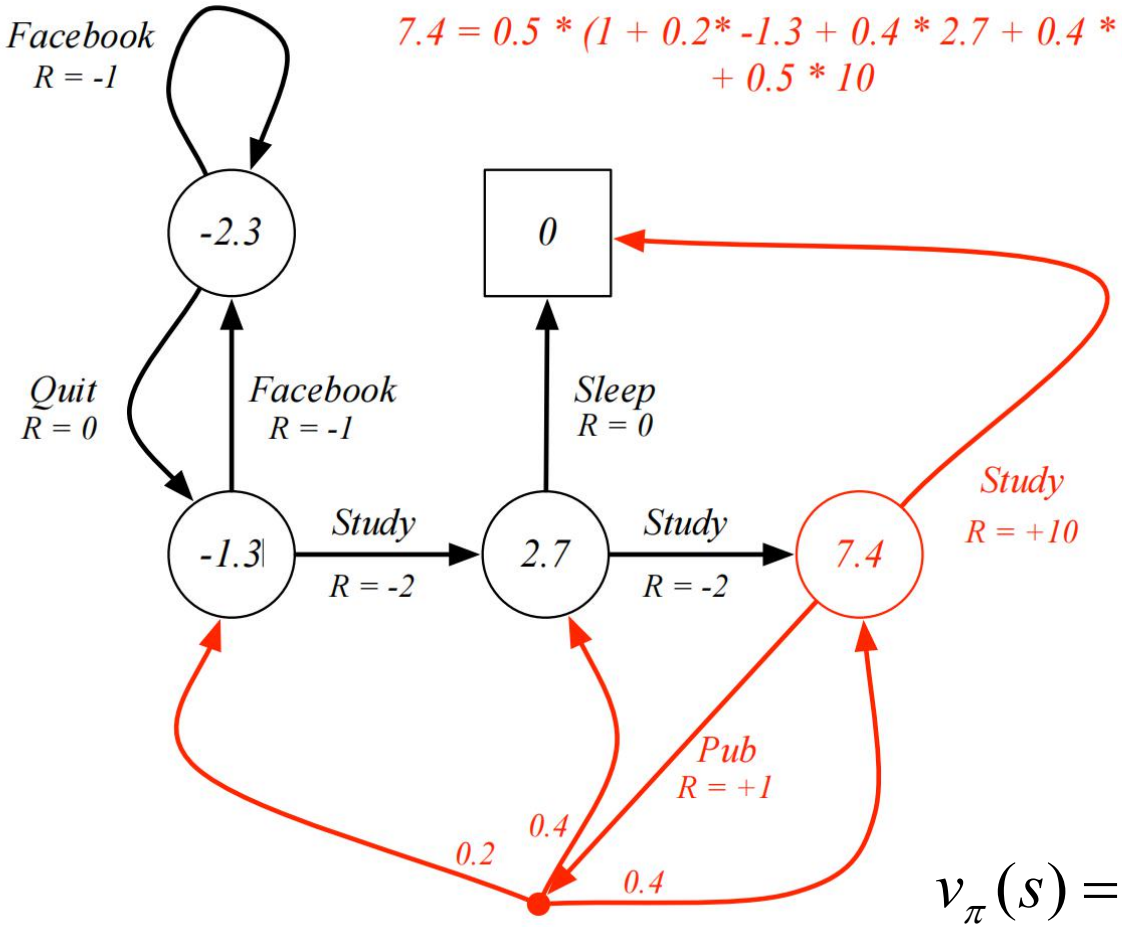
## 关于 $q_\pi$ 的贝尔曼期望方程（2）




$$q_\pi(s,a) = R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a \sum_{a' \in A} \pi(a' \mid s') q_\pi(s',a')$$

# 2.3马尔可夫决策过程

- 例：学生MDP的贝尔曼期望方程




$$v_\pi(s) = \sum_{a \in A} \pi(a \mid s) \left( R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a v_\pi(s') \right)$$

# 2.3马尔可夫决策过程

- 矩阵形式的贝尔曼期望方程

  贝尔曼期望方程可以用induced MRP简洁地表达出来

$$v_\pi = R^\pi + \gamma P^\pi v_\pi$$

  其直接解为

$$v_\pi = (I - \gamma P^\pi)^{-1} R^\pi$$

# 2.3马尔可夫决策过程

- **最优价值函数**

**定义**

最优状态价值函数$v_*(s)$是从所有策略中选出的最大价值函数

$$v_*(s) = \max_{\pi} v_\pi(s)$$

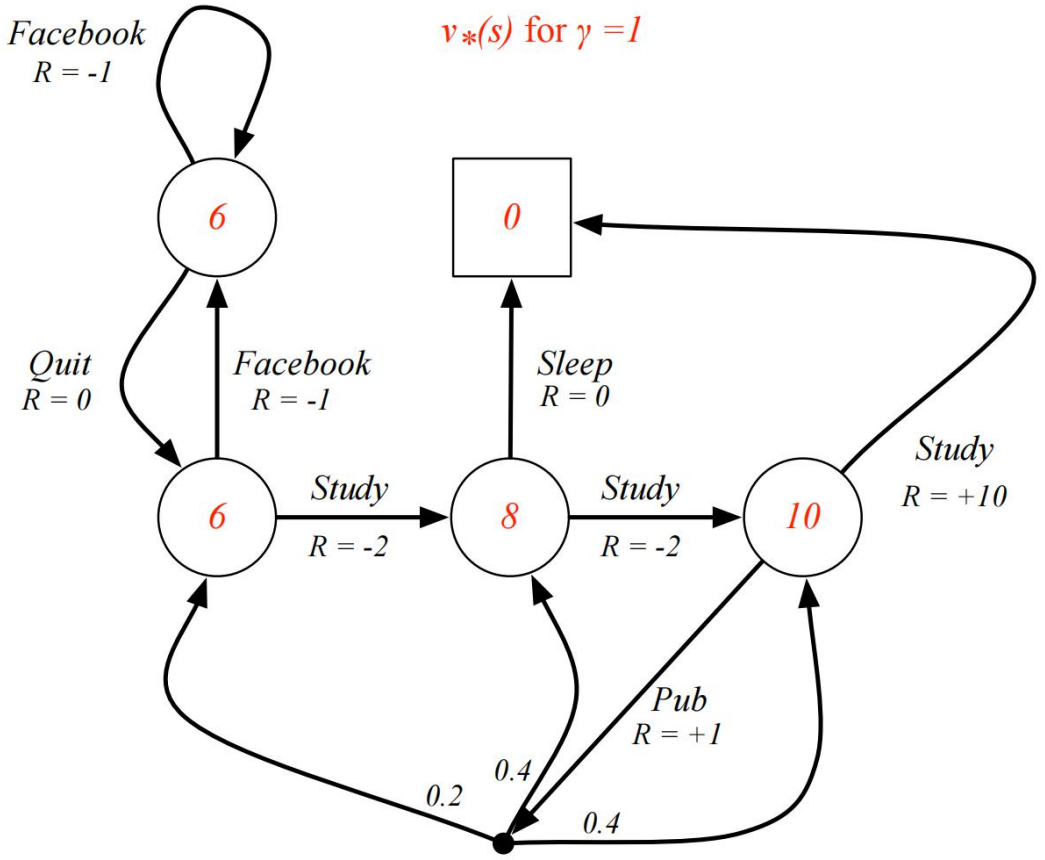
最优动作价值函数$q_*(s,a)$是从所有策略中选出的最大动作价值函数

$$q_*(s,a) = \max_{\pi} q_\pi(s,a)$$

- 最优值函数特指MDP中可能的最佳性能
- 当我们知道了最优价值函数，一个MDP问题就被解决了

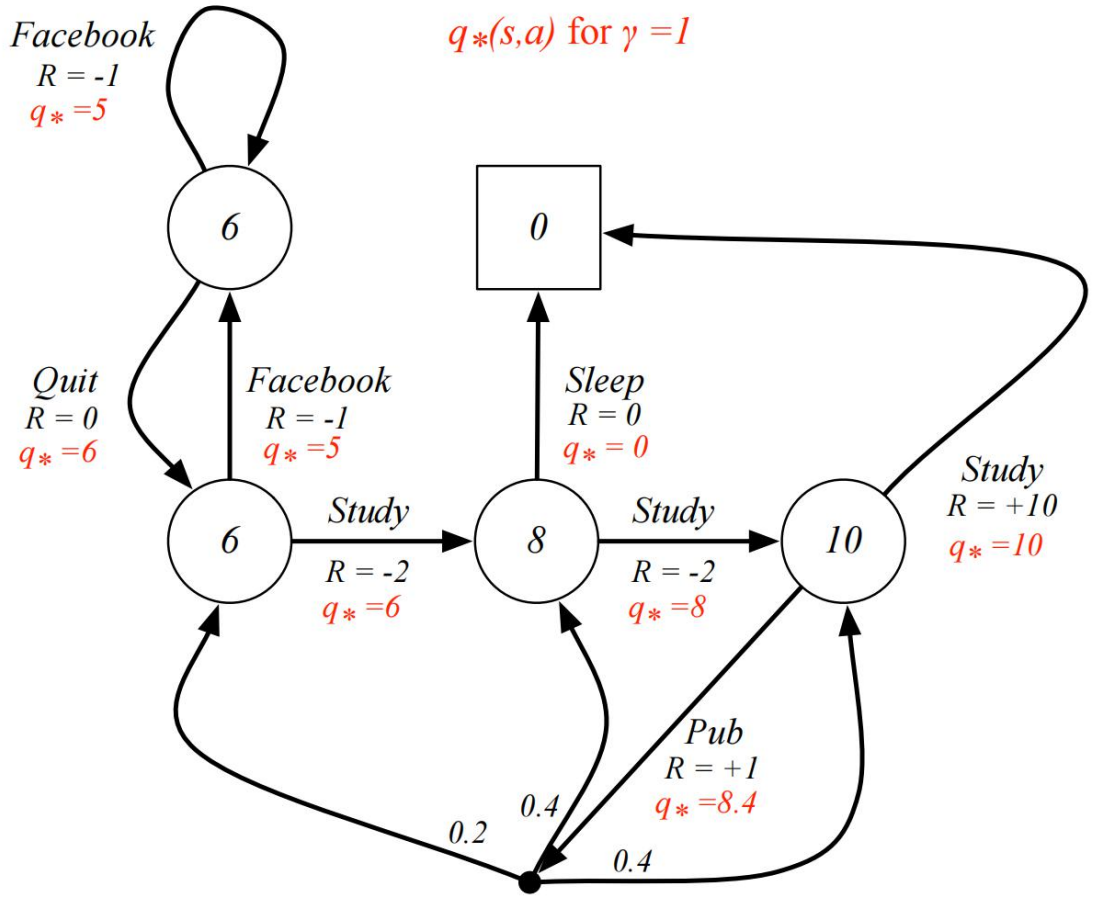
# 2.3马尔可夫决策过程

- 例：学生MDP的最优价值函数



# 2.3马尔可夫决策过程

- 例：学生MDP的最优动作价值函数



# 2.3马尔可夫决策过程

- **最优策略**

定义策略的偏序

$$\pi > \pi' \quad if \quad v_\pi(s) > v_{\pi'}(s), \forall s$$

**定理**

对于任何马尔可夫过程

- 一定存在一个优于或等于其他策略的最优策略 $\pi_*$ ， 其中 $\pi_* \geq \pi, \forall \pi$
- 所有的最优策略一定符合最优价值函数

$$v_{\pi^*}(s) = v_*(s)$$

- 所有的最优策略一定符合最优动作价值函数

$$q_{\pi^*}(s,a) = q_*(s,a)$$

# 2.3马尔可夫决策过程

- **寻找最优策略**

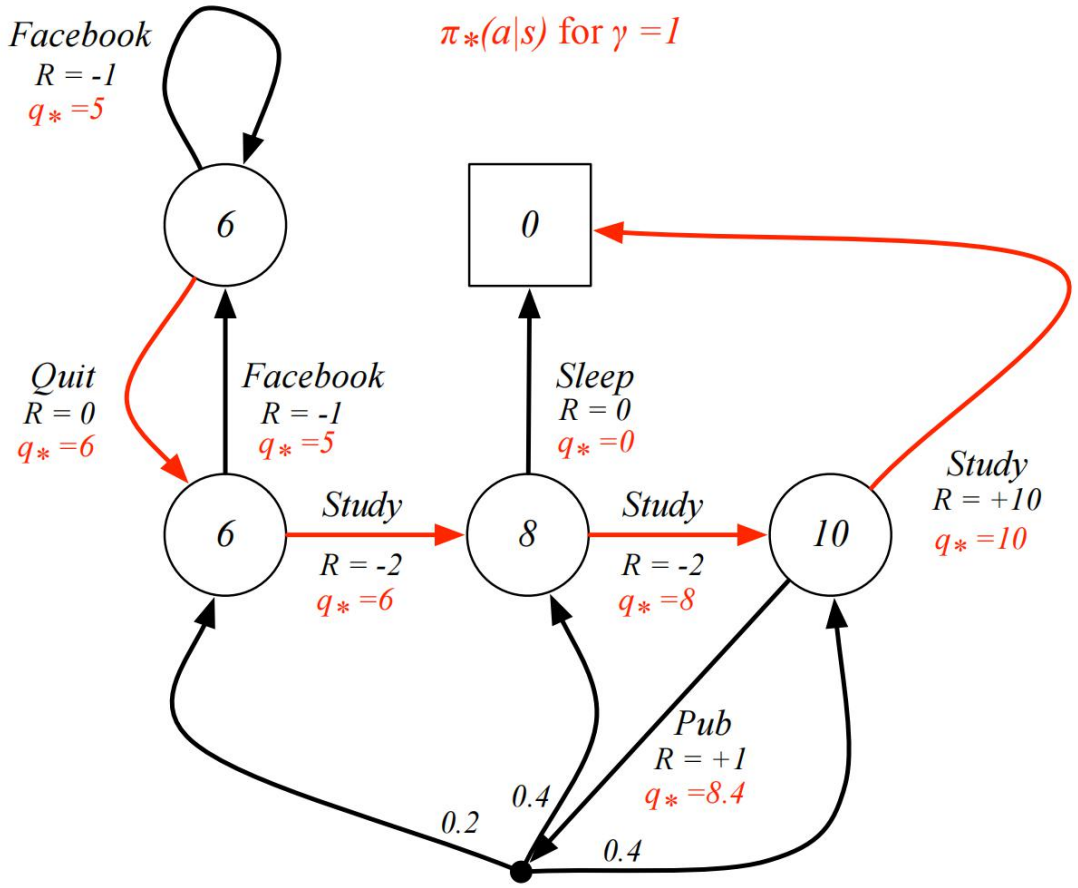
一个最优策略可以通过最大化 $q_*(s,a)$ 来寻找，

$$\pi_*(a|s) = \begin{cases} 1 & \text{if } a = \underset{a \in \mathcal{A}}{\text{argmax}}\ q_*(s,a) \\ 0 & otherwise \end{cases}$$

- 任何MDP都有一个确定的最优策略
- 如果我们知道$q_*(s,a)$，我们立即得到最优策略

# 2.3马尔可夫决策过程

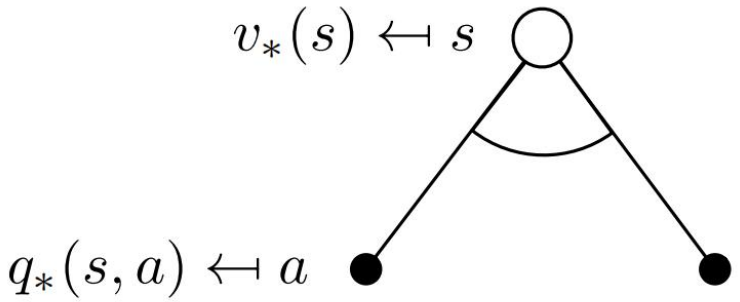
- 例：学生MDP的最优策略



# 2.3马尔可夫决策过程

- $v_*$ 的贝尔曼最优方程

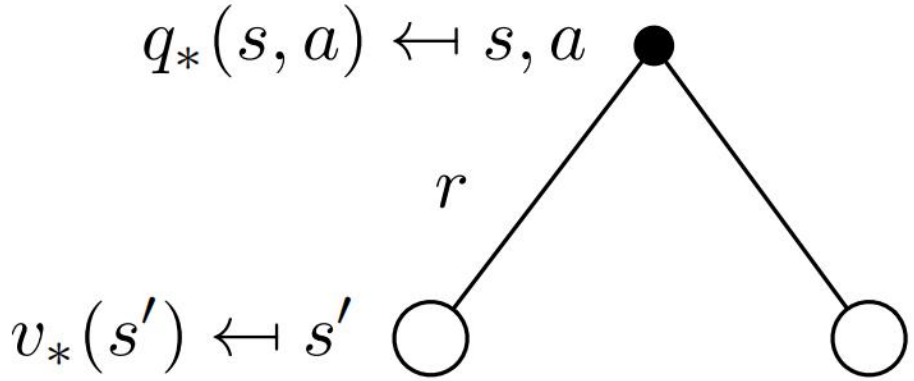
最优价值函数通过贝尔曼最优方程递归关联：




$$v_*(s) = \max_a q_*(s,a)$$

# 2.3马尔可夫决策过程

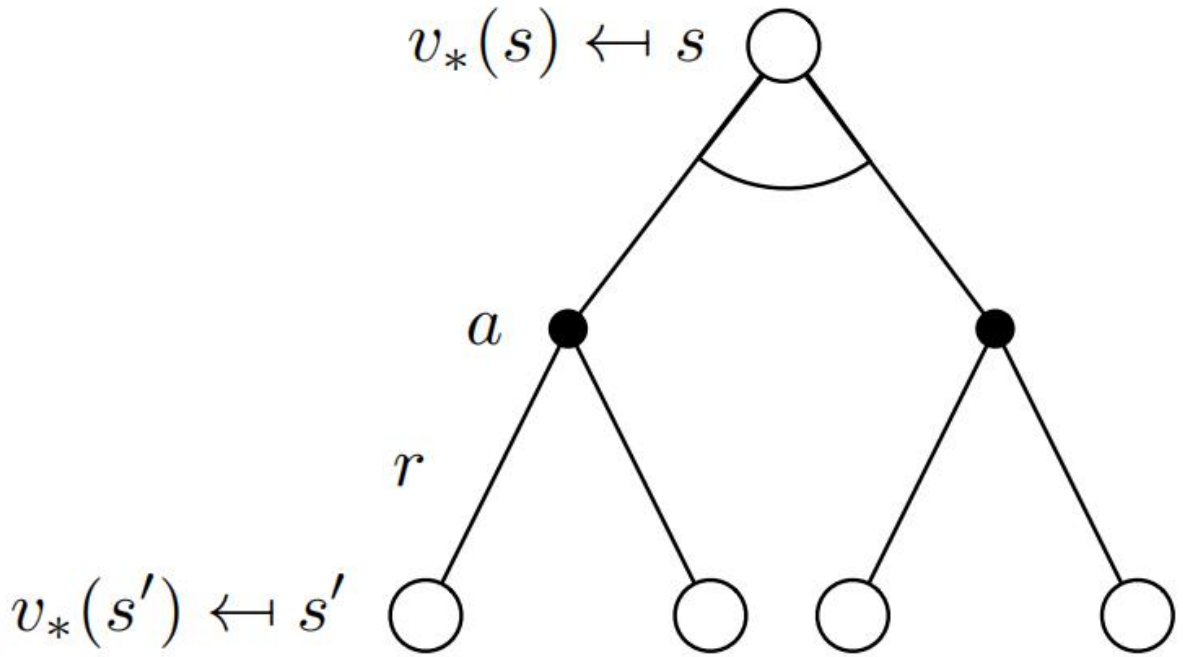
- $q_*$ 的贝尔曼最优方程




$$q_*(s,a) = R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a v_*(s')$$

# 2.3马尔可夫决策过程

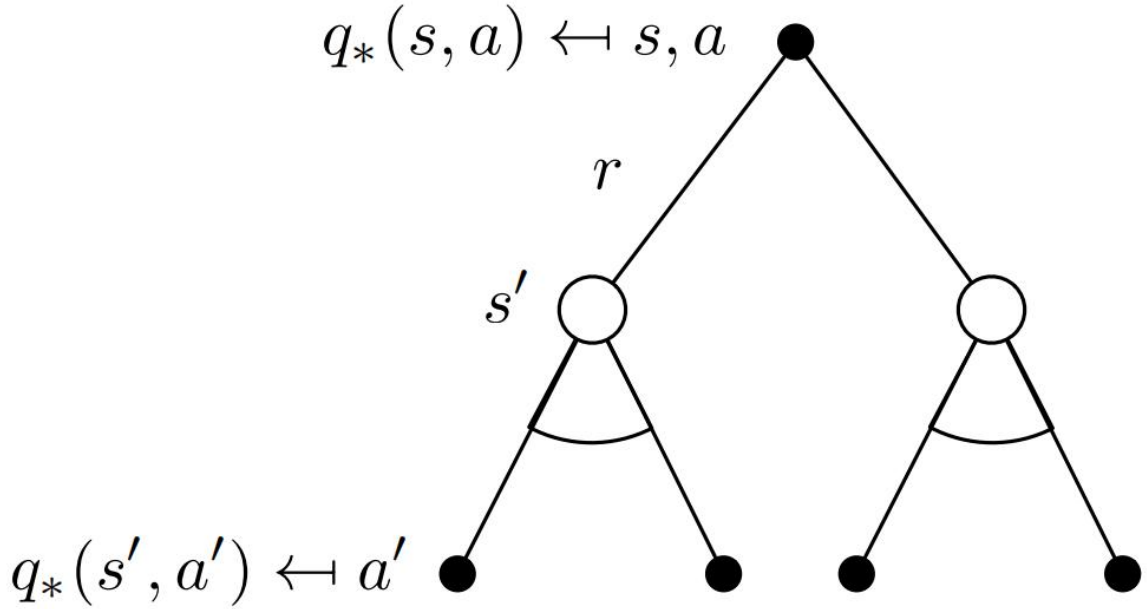
- $v_*$ 的贝尔曼最优方程（2）




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

# 2.3马尔可夫决策过程

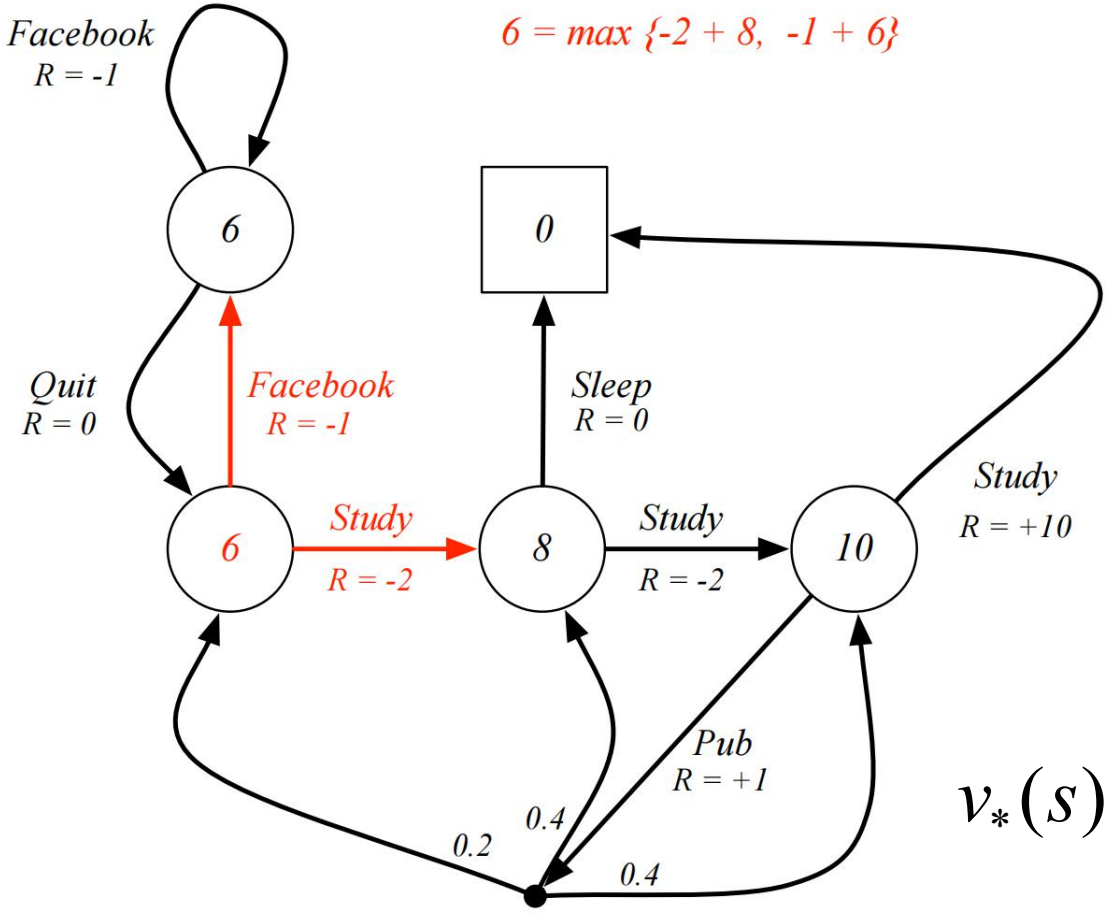
- $q_*$ 的贝尔曼最优方程（2）




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

# 2.3马尔可夫决策过程

- 例：学生MDP的贝尔曼方程




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

# 2.3马尔可夫决策过程

- 解贝尔曼方程
  - 贝尔曼最优方程不是线性方程
  - 一般来说无闭式解
  - 但是有许多间接的方法去求解：
    - 价值迭代
    - 策略迭代
    - 增强学习
    - Sarsa算法

# 2.4强化学习的拓展（了解）

- 无限连续MDP
- 部分可观测MDP
- 未折扣，平均奖励MDP

# 2.4强化学习的拓展（了解）

- 无限MDPs

以下扩展都是可能的

- 可数的无限状态和/或动作空间
  - Straightforward
- 连续的状态和/或动作空间

  线性二次模型（LQR）的闭式
- 连续的时间
  - 需要偏微分方程
  - Hamilton-Jacobi-Bellman（HJB）方程
  - 贝尔曼方程作为时间步长的极限情形→ 0

# 2.4强化学习的拓展（了解）

- 部分可观测马尔可夫决策过程（POMDPs）

部分可观测MDP是一个带有隐状态的MDP，它是一个带有动作的隐马尔可夫模型。

定义

一个POMDP是一个带有 $<S, A, O, P, R, Z, \gamma>$ 的元组

- S是一个有限的状态集
- A是一个有限的动作集
- O是一个有限的观测集
- P是一个状态转移可能性矩阵

  $P_{ss'}^{a} = P[S_{t+1} = s' \mid S_t = s, A_t = a]$
- R是收益方程 $R_s^a = E[R_{t+1} \mid S_t = s, A_t = a]$
- Z是观测函数

  $Z_{s'o}^{a} = P[O_{t+1} = o \mid S_{t+1} = s', A_t = a]$
- $\gamma$是折扣率

# 2.4强化学习的拓展（了解）

- 信念状态（Belief States）

**定义**

一个历史 $H_t$ 是包含一系列动作，观测和收益的序列

$$H_t = A_0, O_1, R_1, ..., A_{t-1}, O_t, R_t$$

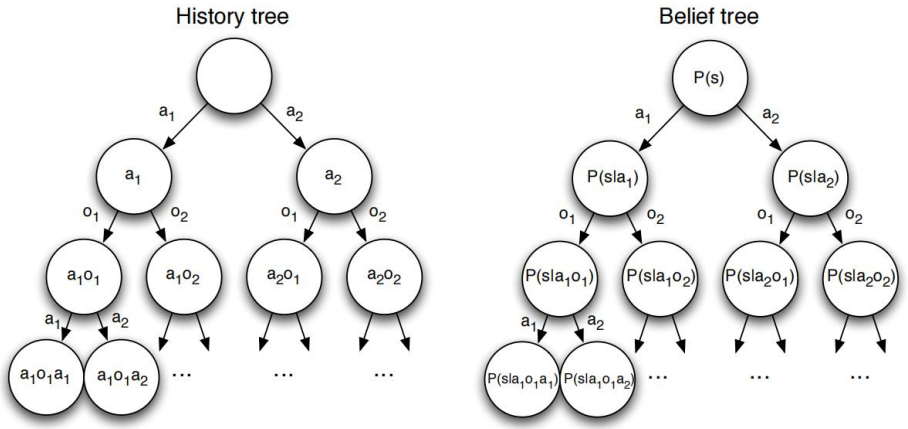
**定义**

信念状态 $b(h)$ 是状态的概率分布，取决于历史h

$$b(h) = (P[S_t = s^1 \mid H_t = h], ..., P[S_t = s^n \mid H_t = h])$$

# 2.4强化学习的拓展（了解）

- 简化POMDPs
  - 历史满足马尔可夫性
  - 信念状态满足马尔可夫性




  - POMDP可以简化为历史树（无限）
  - POMDP可以简化为信念状态树（无限）

# 2.4强化学习的拓展（了解）

- 遍历马尔可夫过程

遍历马尔可夫过程具有两个性质：

循环性：每个状态被访问的次数是无限的

非周期性：访问每个状态没有任何系统周期

**定理**

遍历马尔可夫过程的极限平稳分布$d^{\pi}(s)$具有如下性质

$$d^{\pi}(s)=\sum_{s'\in S}d^{\pi}(s')P_{s's}$$

# 2.4强化学习的拓展（了解）

- 遍历MDP

**定义**

如果由任何策略引起的马尔可夫链是可遍历的，则MDP是可遍历的

对于任何策略 $\pi$，遍历MDP的每个时间步 $p^{\pi}$ 的平均报酬与开始状态无关。

$$p^{\pi}=\lim_{T\to\infty}\frac{1}{T}E[\sum_{t=1}^{T}R_t]$$

## 平均收益价值方程

无折扣遍历MDP的值函数可以用平均收益来表示

是从状态s开始的额外奖励，

$$\tilde{v}_\pi(s) = E_\pi[\sum_{k=1}^{\infty}(R_{t+k} - \rho^\pi) \mid S_t = s]$$

有一个相应的平均收益贝尔曼方程

$$\tilde{v}_\pi(s) = E_\pi[(R_{t+1} - \rho^\pi) + \sum_{k=1}^{\infty}(R_{t+k+1} - \rho^\pi) \mid S_t = s]$$

$$= E_\pi[(R_{t+1} - \rho^\pi) + \tilde{v}_\pi(S_{t+1}) \mid S_t = s]$$

# 2.4强化学习的拓展（了解）

- 提问

- The only stupid question is the one you were afraid to ask but never did.
- -Rich Sutton

# The End