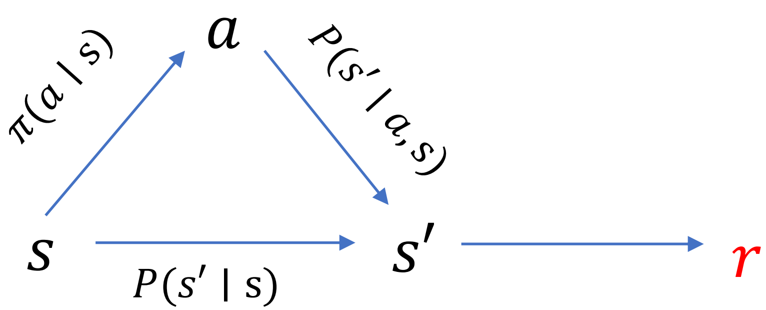
马尔可夫决策过程

$s_0, a_0, s_1, r_1, a_1, s_2, r_2 \ldots s_{t-1}, r_{t-1}, a_{t-1}, s_t, r_t \ldots,$

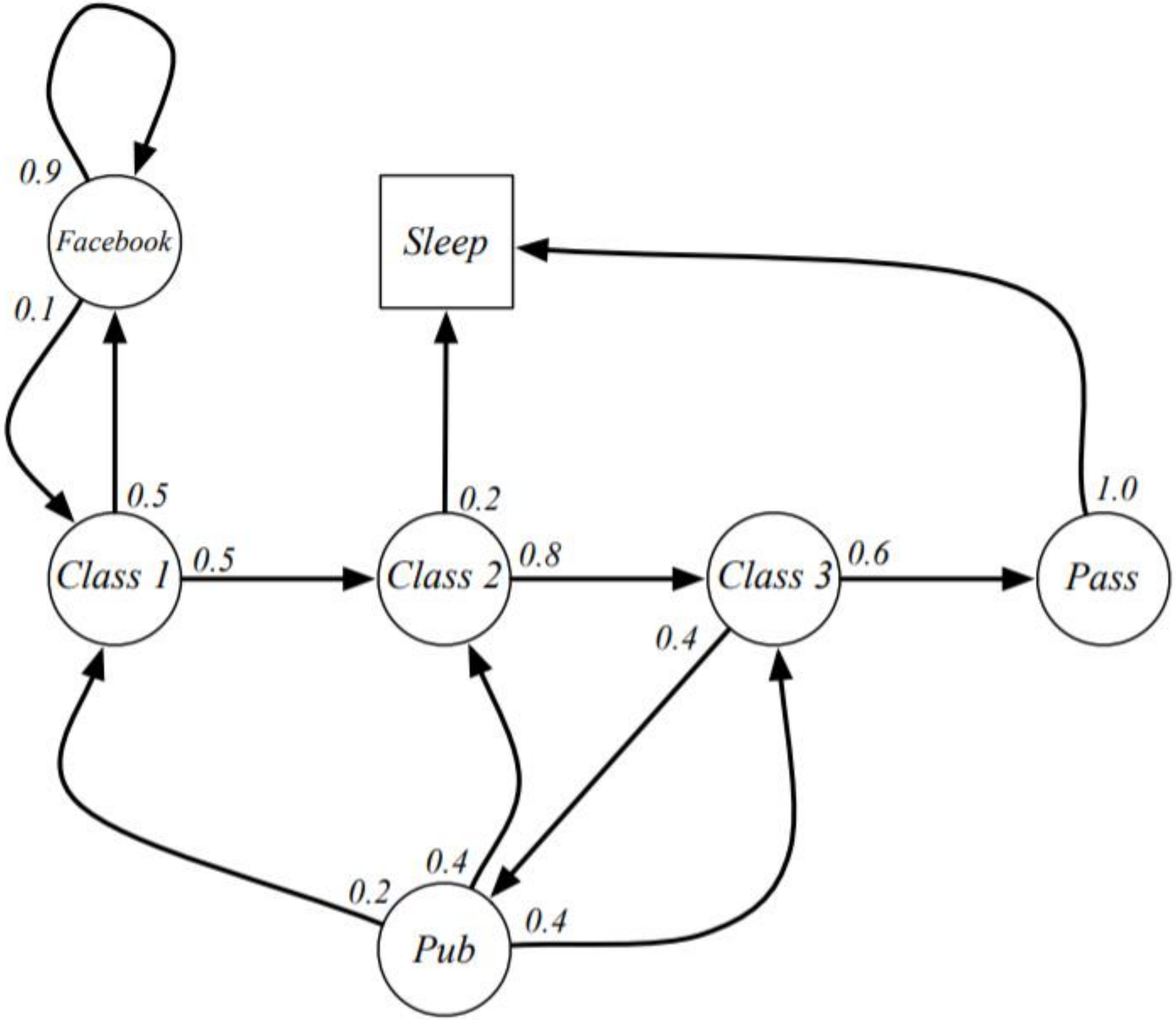
三个概率

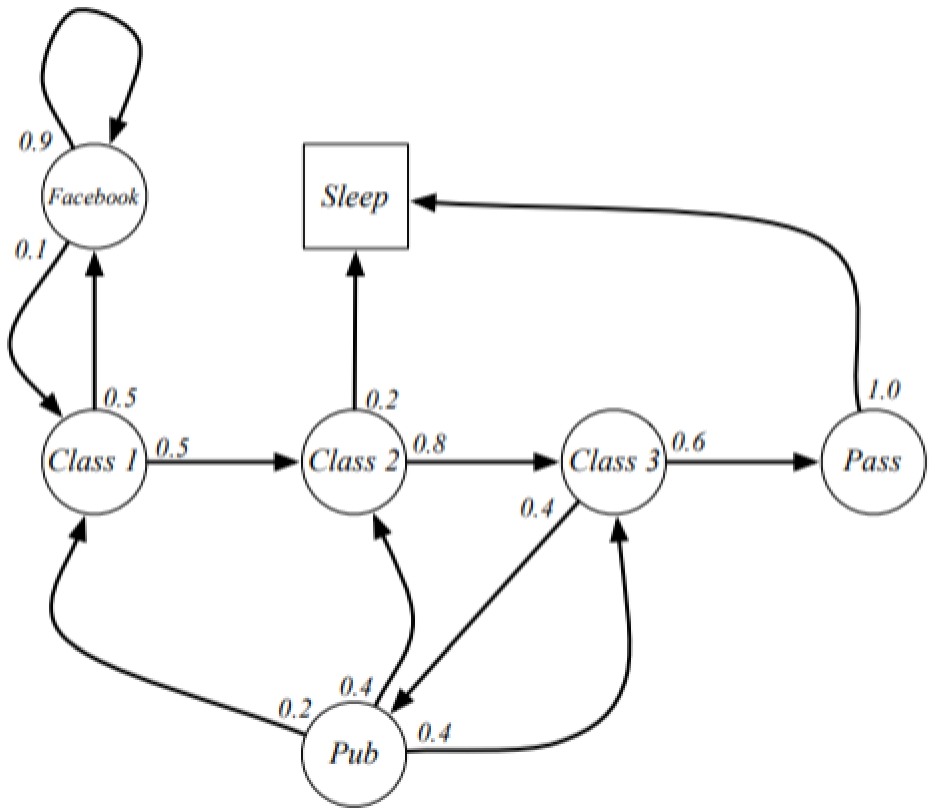
- $\pi(a \mid s) = P(a \mid s)$：表示从状态$s$采取动作$a$的概率，也称策略
- $P(s' \mid a, s)$：表示在状态s下采取动作a后转移到状态s'的概率
- $P(s' \mid s)$：表示从状态s下转移到状态s'的概率



一组序列为(s, a, s', r)，从状态s转移到状态s'后才能计算此时的回报。也可以写成$r_t = r(s_{t-1}, a_{t-1}, s_t)$

学生马尔可夫链，状态-状态（不包括动作a）



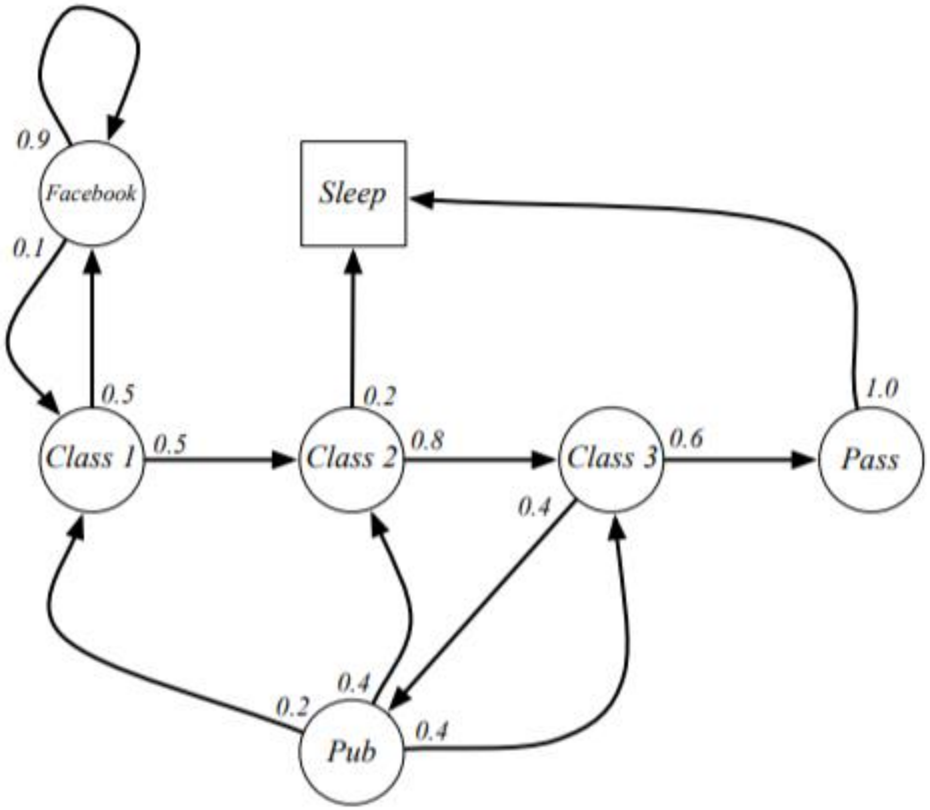


Sample episodes for Student Markov Chain starting from $S_1 = C1$

$$S_1, S_2, ..., S_T$$

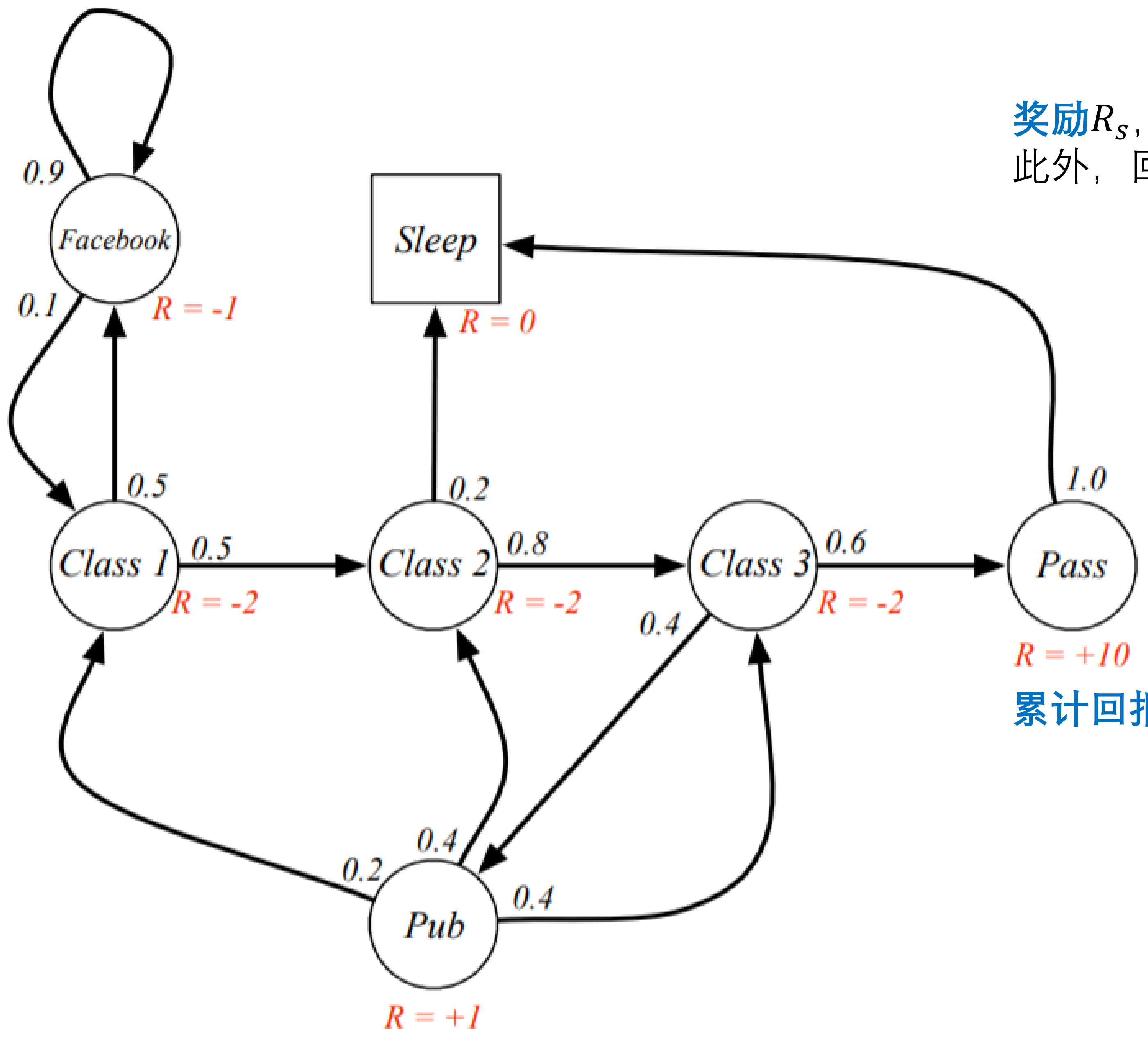
从C1开始，有如下4种方法到达最终状态Sleep(不只四种)

- 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



计算**状态转移概率**$P(s' \mid a, s)$，写成矩阵的形式

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

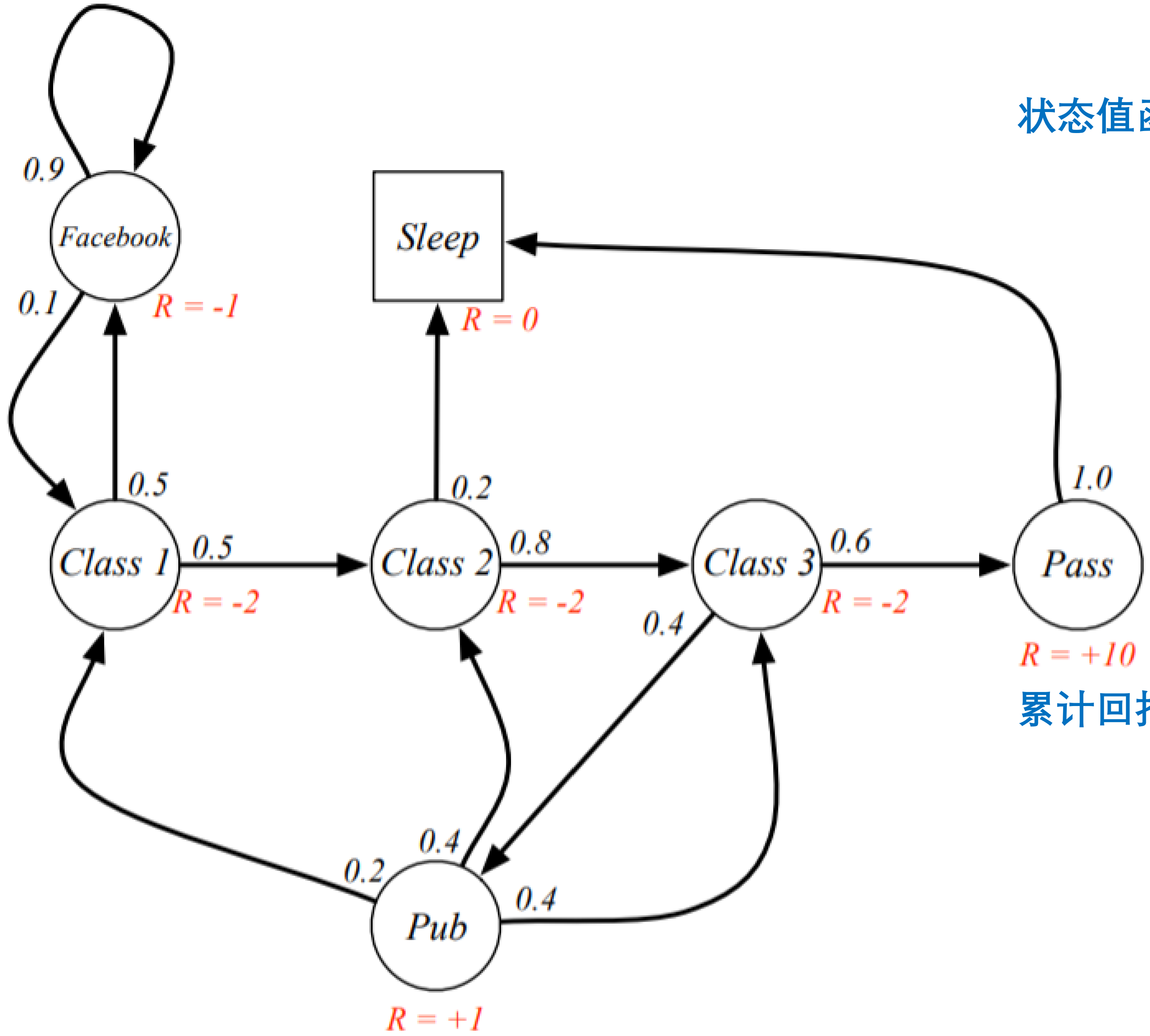


奖励R_s，表示从一种状态转移到另一种状态后的奖励。此外，回报应该随着时间递减，增加折扣系数γ

$$R_s = E[\, R_{t+1} \mid S_t = s\,]$$

累计回报G_t，智能体和环境一次交互过程收到的累计奖励

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

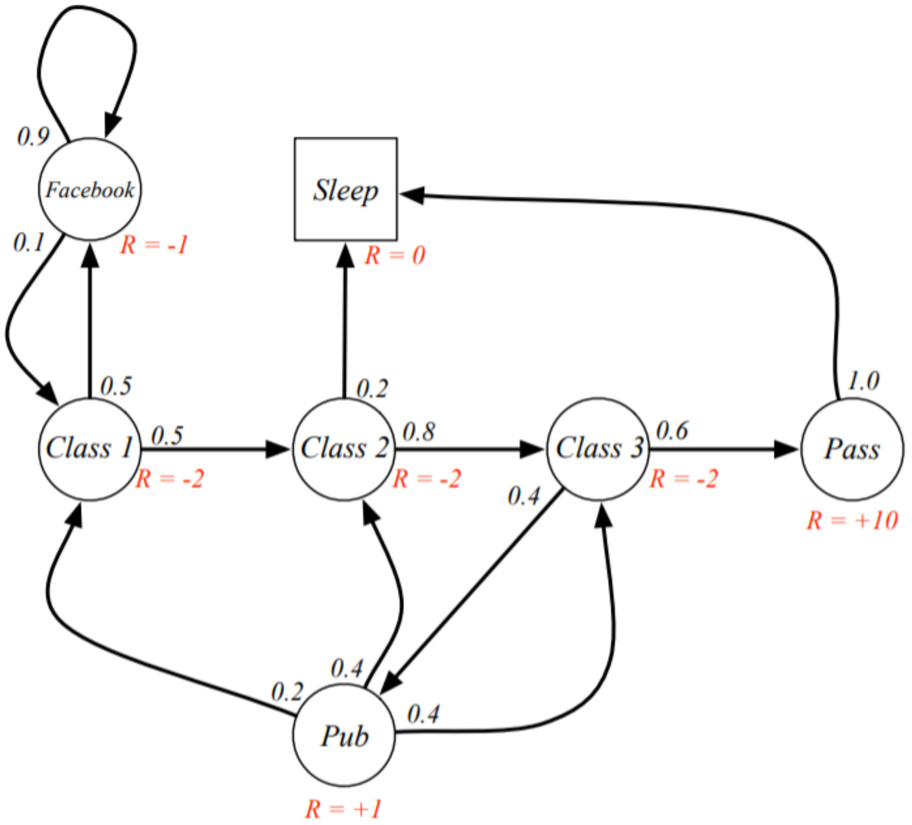


状态值函数$V(s)$

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

累计回报G_t，智能体和环境一次交互过程收到的累计奖励

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



Sample **returns** for Student MRP:
Starting from $S_1 = \text{C1}$ with $\gamma = \frac{1}{2}$

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

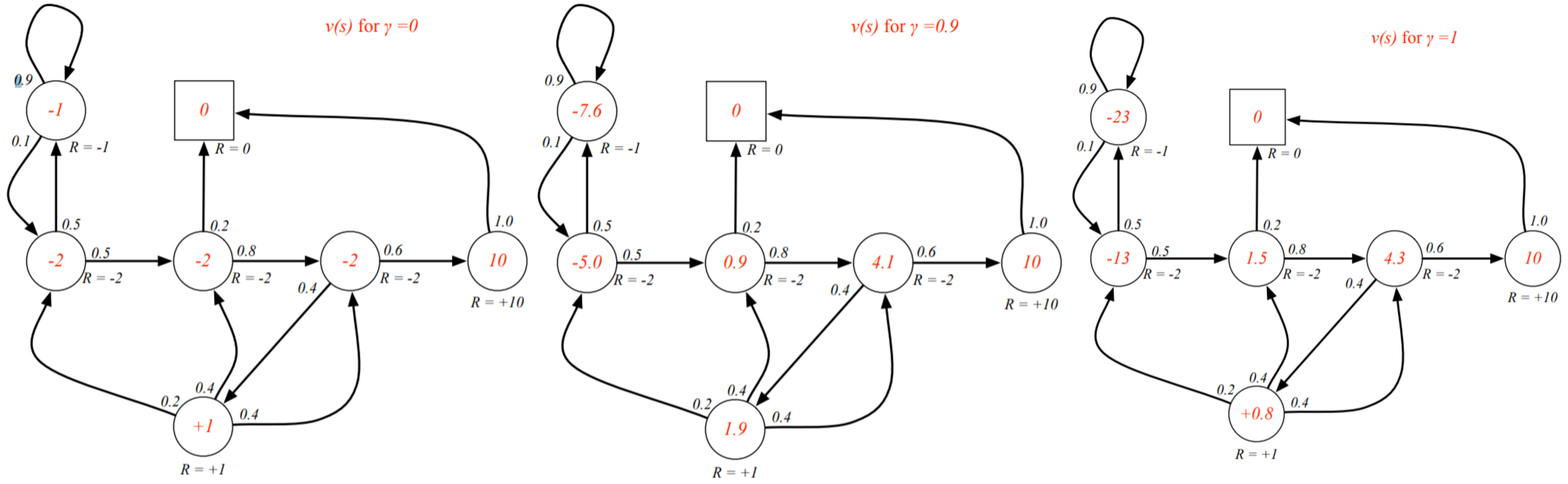
Sample	v_1		
C1 C2 C3 Pass Sleep	$v_1 = -2 - 2 * \frac{1}{2} - 2 * \frac{1}{4} + 10 * \frac{1}{8}$	$=$	-2.25
C1 FB FB C1 C2 Sleep	$v_1 = -2 - 1 * \frac{1}{2} - 1 * \frac{1}{4} - 2 * \frac{1}{8} - 2 * \frac{1}{16}$	$=$	-3.125
C1 C2 C3 Pub C2 C3 Pass Sleep	$v_1 = -2 - 2 * \frac{1}{2} - 2 * \frac{1}{4}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

累计回报G_t，智能体和环境一次交互过程收到的累计奖励

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

状态值函数$V(s)$

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



对于$\gamma = 0,0.9,1$，计算此时的$V_1(s), \ldots V_7(s)$，思考下**如何计算**?

$$V(s) = R_s + \gamma \sum_{s' \in S} P(s' \mid s) V(s')$$

Bellman Equation

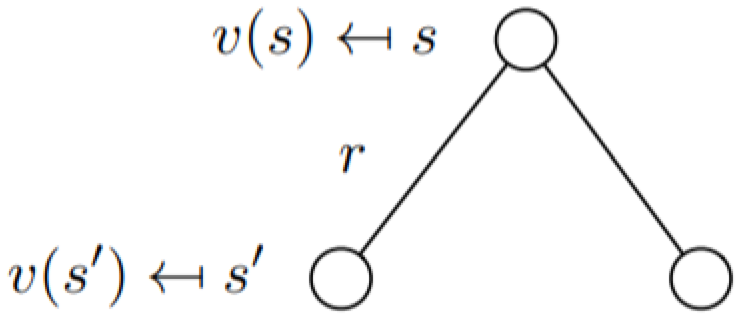
$$V(s) = E[R_{t+1} + \gamma V(S_{t+1}) \mid S_t = s]$$

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

也就是说，上一状态的$V(s)$的值通过下一个状态的$V(S_{t+1})$的值和该过程得到的奖励函数$r_{t+1} = r(s_t, a_t, s_{t+1})$来更新。

$$V(s) = R_s + \gamma \sum_{s' \in S} P(s' \mid s) V(s')$$

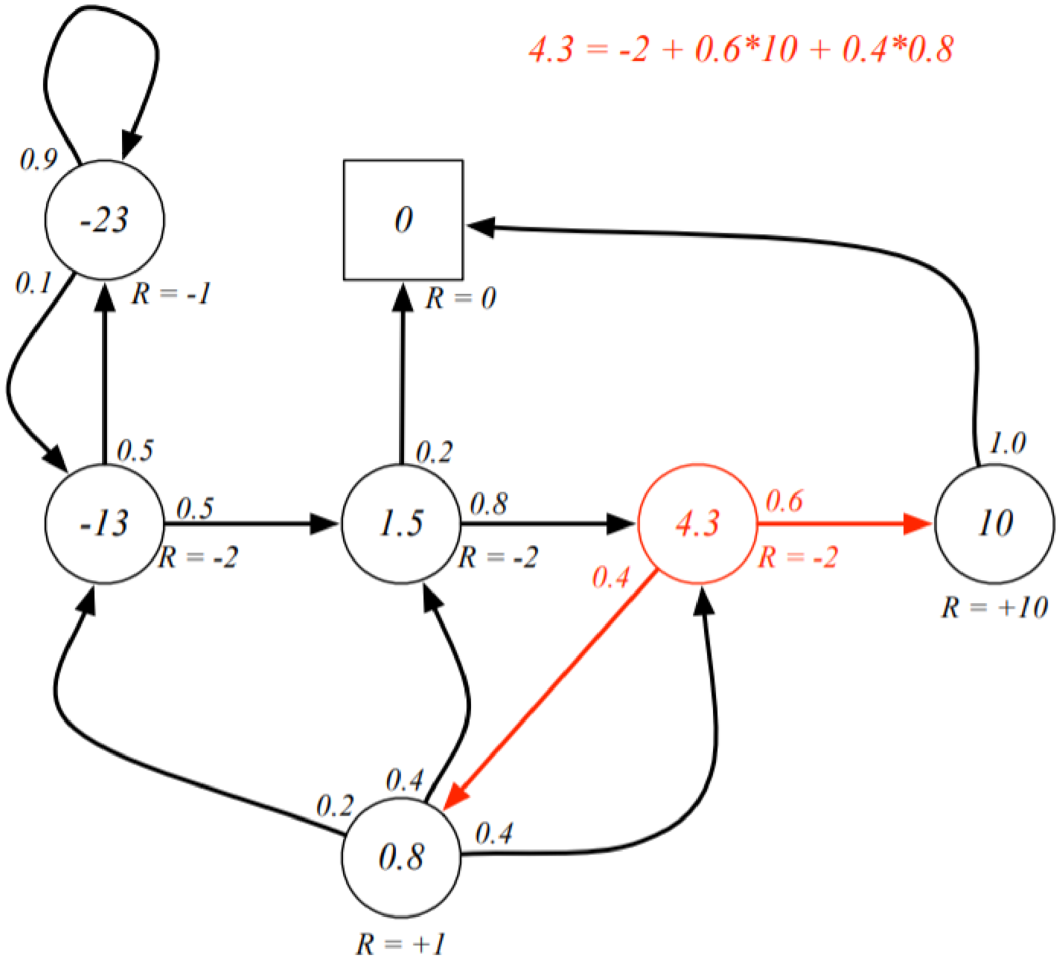
$$R_s = E[R_{t+1} \mid S_t = s]$$



Bellman Equation

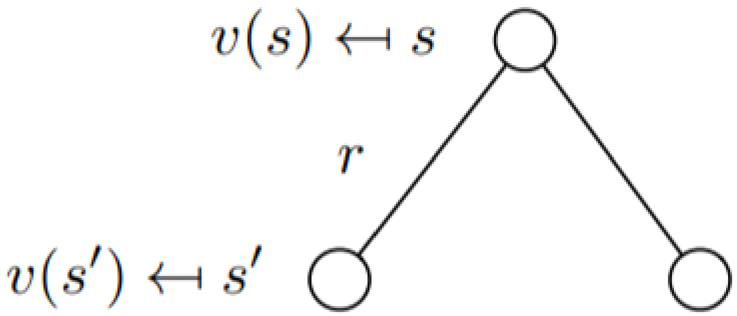
$$V(s) = E[R_{t+1} + \gamma V(S_{t+1}) \mid S_t = s]$$

也就是说，上一状态的$V(s)$的值通过下一个状态的$V(S_{t+1})$的值和该过程得到的奖励函数$r_{t+1} = r(s_t, a_t, s_{t+1})$来更新。



$$V(s) = R_s + \gamma \sum_{s' \in S} P(s' \mid s) V(s')$$

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



求解V值相当于解方程

The Bellman equation can be expressed concisely using matrices,

$$v = \mathcal{R} + \gamma \mathcal{P} v$$

where v is a column vector with one entry per state

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

求解V值相当于解方程

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

```
P=[0 0.5 0 0 0 0.5 0;
0 0 0.8 0 0 0 0.2;
0 0 0 0.6 0.4 0 0;
0 0 0 0 0 0 1;
0.2 0.4 0.4 0 0 0 0;
0.1 0 0 0 0 0.9 0;
0 0 0 0 0 0 1];
gamma=[0,0.9,0.99999];
R=[-2 -2 -2 10 1 -1 0]';
for i = 1:3
    V=inv(eye(7)-gamma(i)*P)*R;
end
```

还有很多方法用来求解该问题

- 动态规划
- 蒙特卡洛
- 时序差分学习

```
V =          V =           V =

  -2         -5.0127       -12.5407
  -2          0.9427         1.4569
  -2          4.0870         4.3212
  10         10.0000        10.0000
   1          1.9084         0.8031
  -1         -7.6376       -22.5386
   0               0              0
```

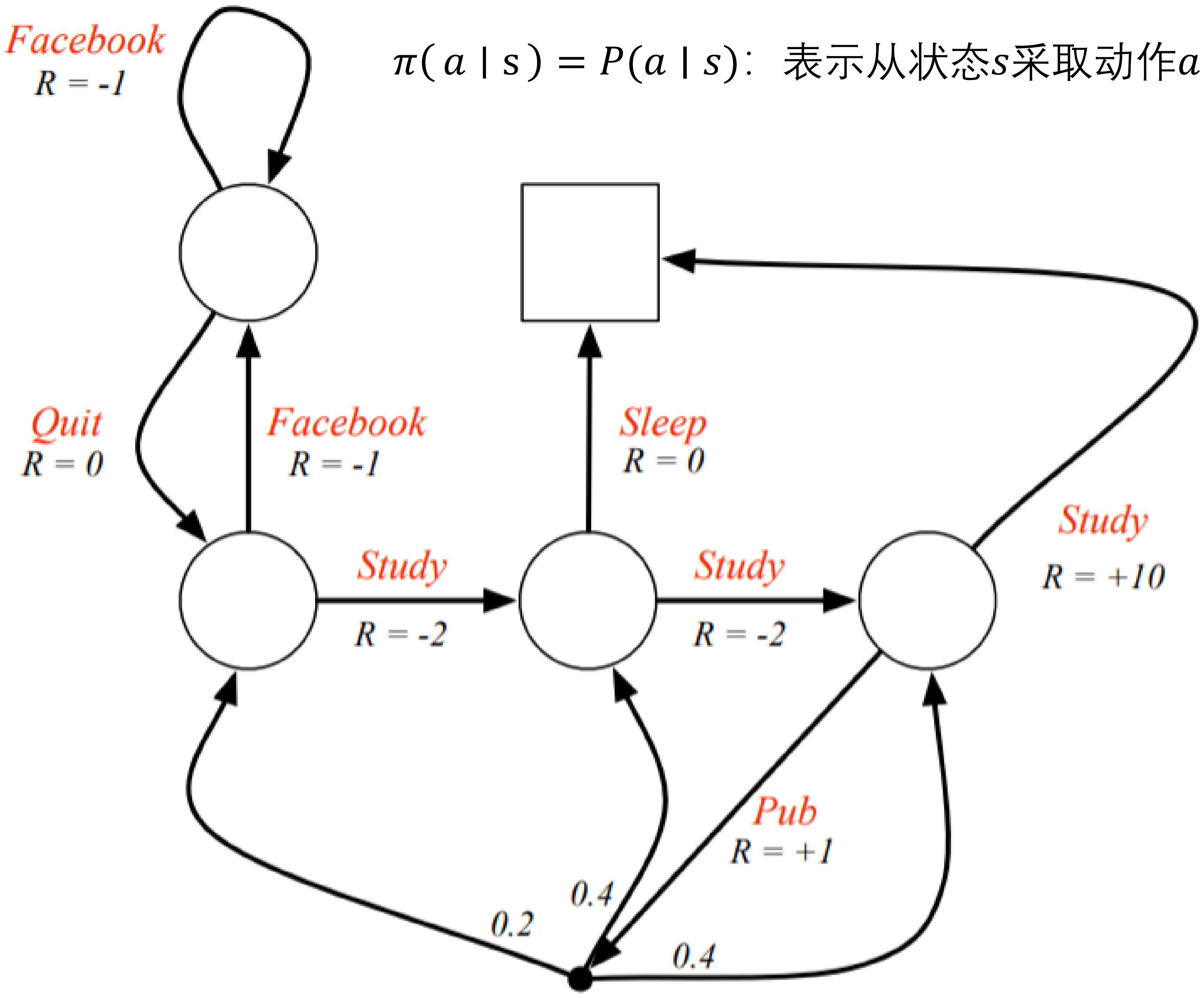
接下来添加动作a，也就是增加了决策过程π

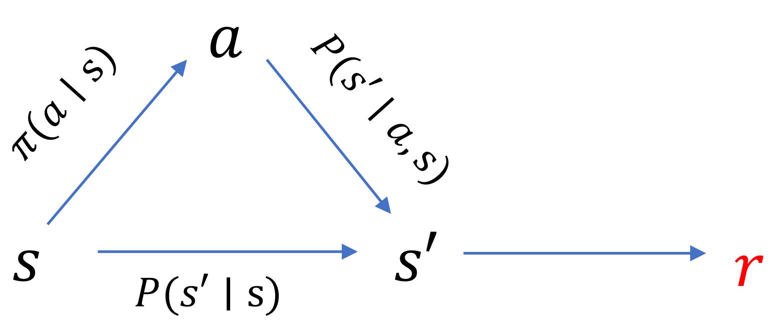
因此在符号上有以下改变：

$$R_s \to R_s^a, V(s) \to V_\pi(s), P(s' \mid s) \to P_\pi(s' \mid s)$$

学生马尔可夫链，状态-动作-状态（增加动作a）

$\pi(a \mid s) = P(a \mid s)$：表示从状态$s$采取动作$a$的概率，也称**策略**





三个**概率**

- $\pi(a \mid s) = P(a \mid s)$：表示从状态$s$采取动作$a$的概率，也称策略
- $P(s' \mid a, s)$：表示在状态s下采取动作a后转移到状态s'的概率
- $P(s' \mid s)$：表示从状态s下转移到状态s'的概率

类似于**全概率公式**

- $\pi(a \mid s) = P(a \mid s)$：表示从状态$s$采取动作$a$的概率，也称**策略**
- $P(s' \mid a, s)$：表示在状态s下采取动作a后转移到状态s'的概率
- $P(s' \mid s)$：表示从状态s下转移到状态s'的概率

在增加了**策略**后，状态转移概率的计算公式为(计算每一个动作的策略概率再求和)：

$$P_\pi(s' \mid s) = \sum_{a \in A} \pi(a \mid s) P(s' \mid a, s) = \pi(a_1 \mid s) P(s' \mid a_1, s) + \pi(a_2 \mid s) P(s' \mid a_2, s) + \cdots$$

在增加了**策略**后，奖励的计算公式为(计算每一个策略下的奖励，即期望)：

状态-动作值函数 $Q(s,a) = E[G_t \mid S_t = s, A_t = a]$ **状态值函数** $V(s) = E[G_t \mid S_t = s]$

Bellman Equation $Q_\pi(s,a) = E[R_{t+1} + \gamma Q(S_{t+1}) \mid S_t = s, A_t = a]$ $V(s) = E[R_{t+1} + \gamma V(S_{t+1}) \mid S_t = s]$

$$V(s) = R_s + \gamma \sum_{s' \in S} P(s' \mid s) V(s')$$

增加策略π

$$\boxed{\begin{aligned} P_\pi(s' \mid s) &= \sum_{a \in A} \pi(a \mid s) P(s' \mid a, s) \\ R_s^\pi &= \sum_{a \in A} \pi(a \mid s) R_s^a \end{aligned}}$$

$$\begin{aligned} V_\pi(s) &= R_s^\pi + \gamma \sum_{s \in s'} P_\pi(s' \mid s) V_\pi(s') \\ &= R_s^\pi + \gamma \sum_{a \in A} \pi(a \mid s) P(s' \mid a, s) V_\pi(s') \\ &= \sum_{a \in A} \pi(a \mid s) \boxed{[R_s^a + \gamma \sum_{s \in s'} P(s' \mid a, s) V_\pi(s')]} \end{aligned}$$

记为$Q_\pi(s,a)$

$$Q_\pi(s,a) = R_s^a + \gamma \sum_{s \in s'} P(s' \mid a, s) V_\pi(s') \qquad V_\pi(s) = \sum_{a \in A} \pi(a \mid s) Q_\pi(s,a)$$

Bellman Equation

$$\begin{aligned} Q_\pi(s,a) &= R_s^a + \gamma \sum_{s \in s'} P(s' \mid a, s) V_\pi(s') \\ &= R_s^a + \gamma \sum_{s \in s'} P(s' \mid a, s) \sum_{a \in A} \pi(a' \mid s') Q_\pi(s', a') \end{aligned}$$

$$\begin{aligned} V_\pi(s) &= \sum_{a \in A} \pi(a \mid s) Q_\pi(s,a) \\ &= \sum_{a \in A} \pi(a \mid s) [R_s^a + \gamma \sum_{s \in s'} P(s' \mid a, s) V_\pi(s')] \end{aligned}$$

状态-动作值函数 $Q(s,a) = E[G_t \mid S_t = s, A_t = a]$ 状态值函数 $V(s) = E[G_t \mid S_t = s]$

Bellman Equation $Q_\pi(s,a) = E[R_{t+1} + \gamma Q(S_{t+1}) \mid S_t = s, A_t = a]$ $V(s) = E[R_{t+1} + \gamma V(S_{t+1}) \mid S_t = s]$

	状态值函数$V^\pi(s)$	状态-动作值函数$Q^\pi(s,a)$
含义	从状态s开始，执行策略π后得到的期望总回报	从初始状态为s执行动作a，然后执行策略π得到的总回报
计算公式	$V^\pi(s) = E_{\tau\sim p(\tau)}[\sum_{t=0}^{T-1}\gamma^t r_{t+1} \vert \tau_{s_0} = s]$	$Q^\pi(s,a) = E_{s'\sim p(s'\vert s,a)}[r(s,a,s') + \gamma V^\pi(s')]$
贝尔曼	$V^\pi(s) = E_{a\sim\pi(a\vert s)} E_{s'\sim p(s'\vert s,a)}[r(s,a,s') + \gamma V^\pi(s')]$	$Q^\pi(s,a) = E_{s'\sim p(s'\vert s,a)}[r(s,a,s') + \gamma E_{a'\sim\pi(a'\vert s')} Q^\pi(s',a')]$
关系	状态值函数$V^\pi(s)$是Q函数$Q^\pi(s,a)$关于动作a的期望	$V^\pi(s) = E_{a\sim\pi(a\vert s)} Q^\pi(s,a)$

Bellman Equation

$$Q_\pi(s,a) = R_s^a + \gamma \sum_{s\in s'} P(s' \mid a,s) V_\pi(s')$$

$$= R_s^a + \gamma \sum_{s\in s'} P(s' \mid a,s) \sum_{a\in A} \pi(a' \mid s') Q_\pi(s',a')$$

$$V_\pi(s) = \sum_{a\in A} \pi(a \mid s) Q_\pi(s,a)$$

$$= \sum_{a\in A} \pi(a \mid s)[R_s^a + \gamma \sum_{s\in s'} P(s' \mid a,s) V_\pi(s')]$$

状态-动作值函数Q(s,a) $Q(s,a) = E[G_t \mid S_t = s, A_t = a]$

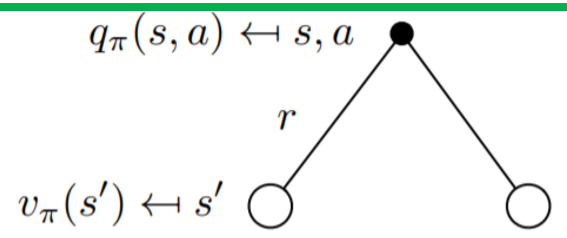
状态值函数V(s) $V(s) = E[G_t \mid S_t = s]$

Bellman Equation

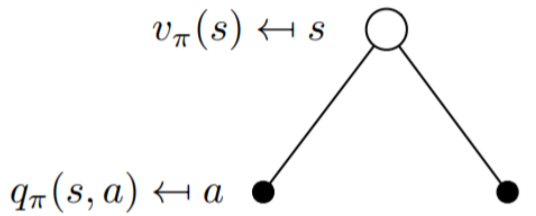
$$Q_\pi(s,a) = E[R_{t+1} + \gamma Q(S_{t+1}) \mid S_t = s, A_t = a]$$

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

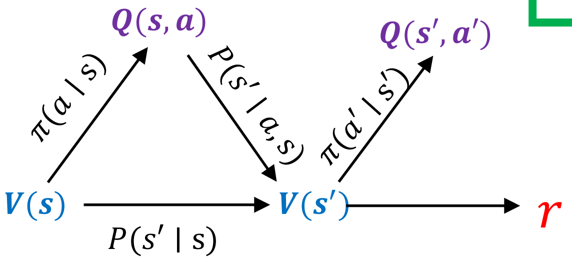
与箭头相反的方向计算即从后续状态开始，计算先前的状态值。

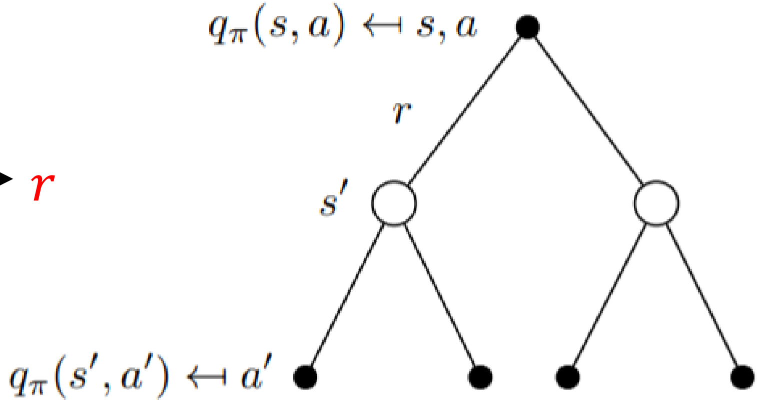


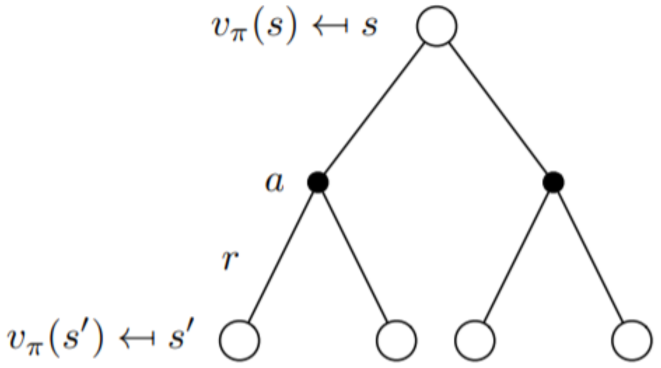
$$Q_\pi(s,a) = R_s^a + \gamma \sum_{s \in s'} P(s' \mid a, s) V_\pi(s')$$



$$V_\pi(s) = \sum_{a \in A} \pi(a \mid s) Q_\pi(s,a)$$







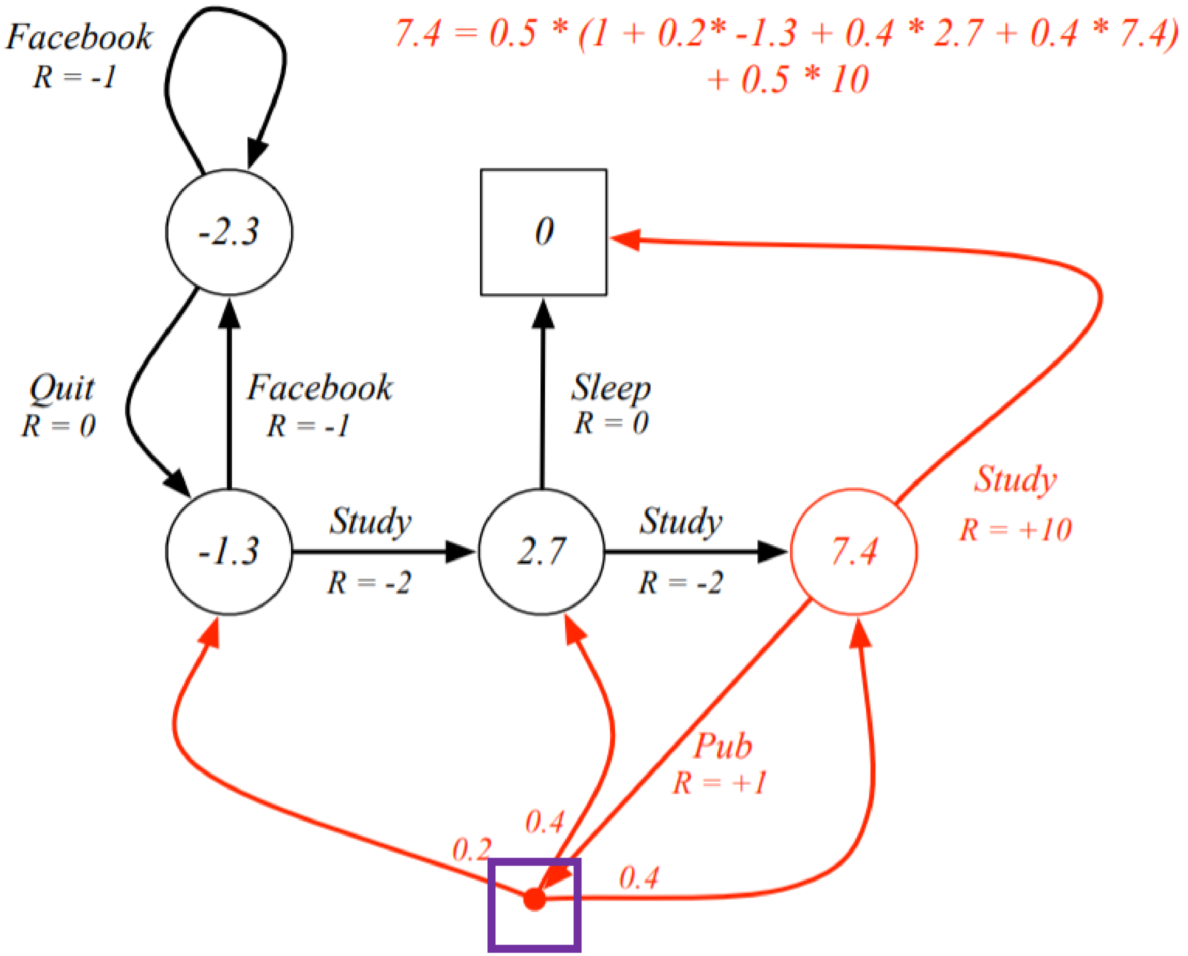
$$Q_\pi(s,a) = R_s^a + \gamma \sum_{s \in s'} P(s' \mid a, s) V_\pi(s')$$

$$= R_s^a + \gamma \sum_{s \in s'} P(s' \mid a, s) \sum_{a \in A} \pi(a' \mid s') Q_\pi(s', a')$$

$$V_\pi(s) = \sum_{a \in A} \pi(a \mid s) Q_\pi(s,a)$$

$$= \sum_{a \in A} \pi(a \mid s) [R_s^a + \gamma \sum_{s \in s'} P(s' \mid a, s) V_\pi(s')]$$

学生马尔可夫链，状态-动作-状态



该位置是一个新的状态，转移的类型为状态-状态。

验证此时的V值，注意包括两种类型，一种是状态-动作-状态，另一种是状态-状态。

而且计算的终止应该为下一个动作前。

比如$s, a(study), s$或$s, \mathrm{a}(Pub), s', s''$,由状态直接转移成另一个状态，相当于策略概率的值为$\pi(a \mid s) = 1$。

$P(s' \mid a, s)$决定着有执行动作后状态的各种可能

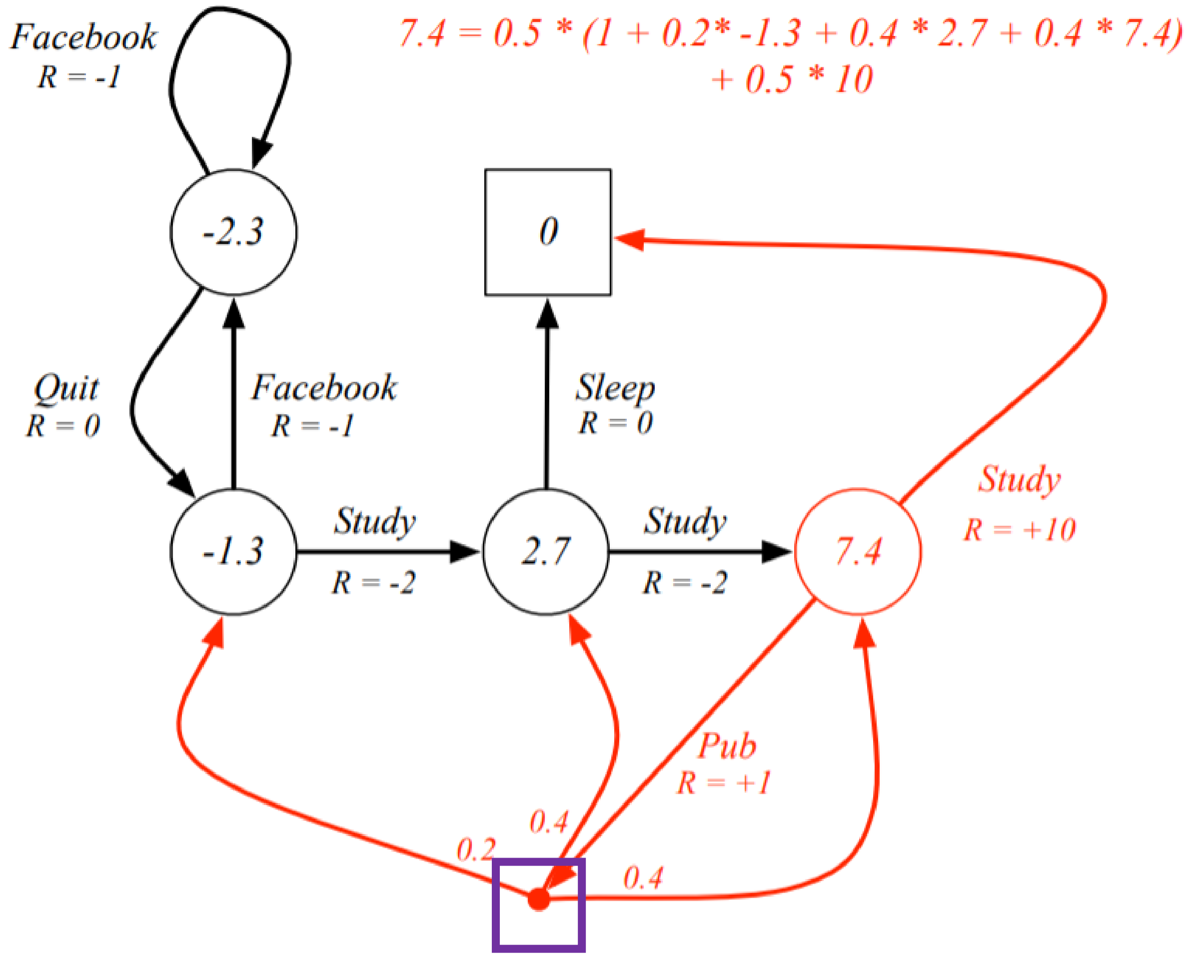
$\pi(a \mid s)$决定着选择动作的可能

$$V_\pi(s) = \sum_{a \in A} \pi(a \mid s) Q_\pi(s, a)$$

$$= \sum_{a \in A} \pi(a \mid s)[R_s^a + \gamma \sum_{s \in s'} P(s' \mid a, s) V_\pi(s')]$$

$$V(s) = R_s + \gamma \sum_{s' \in S} P(s' \mid s) V(s')$$

学生马尔可夫链，状态-动作-状态



$$V_\pi(s) = \sum_{a\in A} \pi(a \mid s) Q_\pi(s,a)$$

$$= \sum_{a\in A} \pi(a \mid s)[R_s^a + \gamma \sum_{s\in s'} P(s' \mid a,s) V_\pi(s')]$$

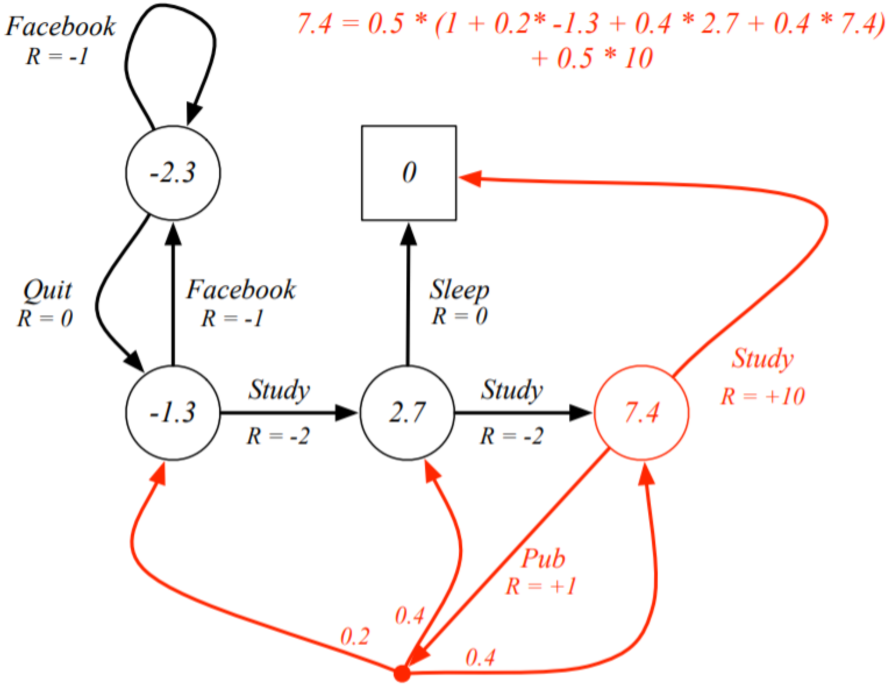
$$V(s) = R_s + \gamma \sum_{s'\in S} P(s' \mid s) V(s')$$

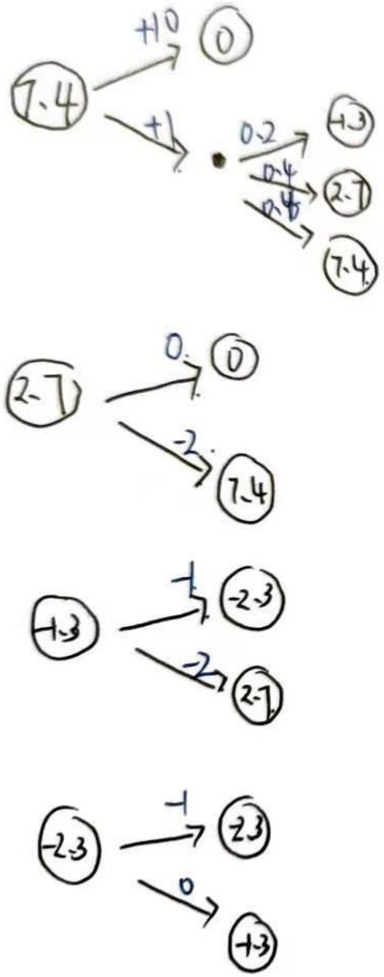
该位置是一个新的状态，转移的类型为状态-状态。

学生马尔可夫链，状态-动作-状态

$$V_\pi(s) = \sum_{a \in A} \pi(a \mid s)[R_s^a + \gamma \sum_{s \in s'} P(s' \mid a, s) V_\pi(s')]$$

$$V(s) = R_s + \gamma \sum P(s' \mid s) V(s')$$





π R_s P $V(s')$

$0.5 \times [10 + 1 \times 0] = 5$

$0.5 \times [1 + 0.2 \times (-1.3) + 0.4 \times 2.7 + 0.4 \times 7.4] = 2.39$

则 $7.4 \approx 5 + 2.39 = 7.39$

$0.5 \times [0 + 1 \times 0] = 0$

$0.5 \times [-2 + 1 \times 7.4] = 2.7$

则 $2.7 \approx 2.7$

$0.5 \times [-1 + 1 \times (-2.3)] = -1.65$

$0.5 \times [-2 + 1 \times 2.7] = 0.35$

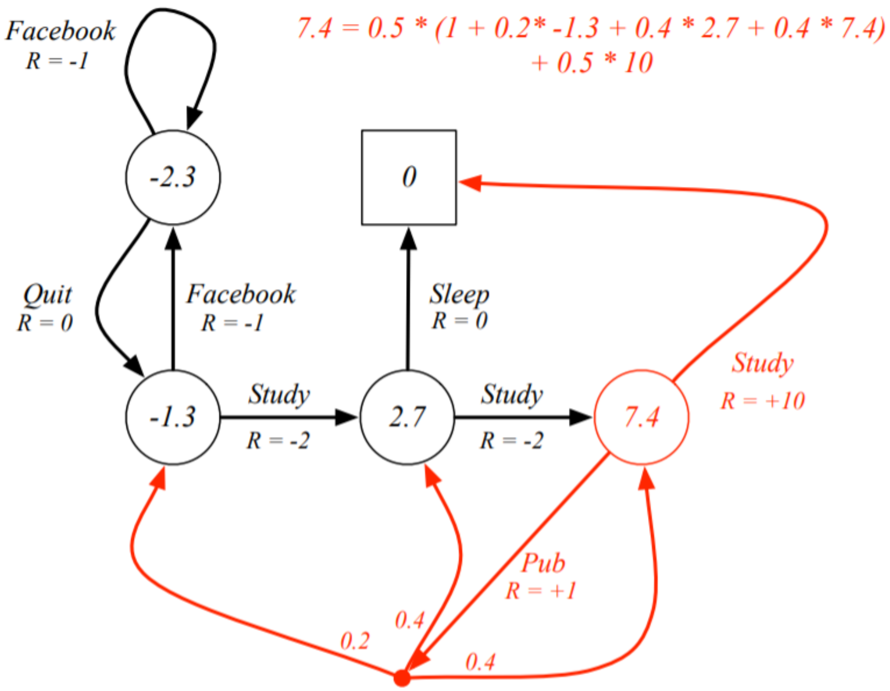
$0.5 \times [-1 + 1 \times (-2.3)] = \frac{-3.3}{2}$

$0.5 \times [0 + 1 \times (-1.3)] = \frac{-1.3}{2}$

学生马尔可夫链，状态-动作-状态

$$V_\pi(s) = \sum_{a\in A}\pi(a\mid s)[R_s^a + \gamma\sum_{s\in s'}P(s'\mid a,s)V_\pi(s')]$$

$$V(s) = R_s + \gamma\sum_{s'\in S}P(s'\mid s)V(s')$$




```
P1=[0.5 0.5 0 0 0;
    0.5 0 0.5 0 0;
    0 0 0 0.5 0.5;
    0 0.1 0.2 0.2 0.5;
    0 0 0 0 0];
R1=[-0.5 -1.5 -1 5.5 0]';
V1=inv(eye(5)-P1)*R1;
```

一共有5个状态，在策略下的状态转移还应该乘以策略的概率，策略下的奖励为不同方案奖励乘积的和。这里选择动作策略均为0.5。
如S_4来说，有两种选择策略各为0.5，因此$R = 0.5\times 1 + 0.5\times 10 = 5.5$.

在增加了**策略**后，状态转移概率的计算公式为(计算每一个动作的策略概率再求和)：

$$P_\pi(s'\mid s) = \sum_{a\in A}\pi(a\mid s)P(s'\mid a,s) = \pi(a_1\mid s)P(s'\mid a_1,s) + \pi(a_2\mid s)P(s'\mid a_2,s) + \cdots$$

在增加了**策略**后，奖励的计算公式为(计算每一个策略下的奖励，即期望)：

求解V_π值相当于**解方程**

The Bellman expectation equation can be expressed concisely using the induced MRP,

$$v_\pi = \mathcal{R}^\pi + \gamma \mathcal{P}^\pi v_\pi$$

with direct solution

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

```
P1=[0.5 0.5 0 0 0;
    0.5 0 0.5 0 0;
    0 0 0 0.5 0.5;
    0 0.1 0.2 0.2 0.5;
    0 0 0 0 0];
R1=[-0.5 -1.5 -1 5.5 0]';
V1=inv(eye(5)-P1)*R1;
```

```
V1 =

   -2.3077
   -1.3077
    2.6923
    7.3846
         0
```

在增加了**策略**后，状态转移概率的计算公式为(计算每一个动作的策略概率再求和):

$$P_\pi(s' \mid s) = \sum_{a \in A} \pi(a \mid s) P(s' \mid a, s) = \pi(a_1 \mid s) P(s' \mid a_1, s) + \pi(a_2 \mid s) P(s' \mid a_2, s) + \cdots$$

在增加了**策略**后，奖励的计算公式为(计算每一个策略下的奖励，即期望):

$$R_s^\pi = \sum_{a \in A} \pi(a \mid s) R_s^a = \pi(a_1 \mid s) R_s^{a_1} + \pi(a_2 \mid s) R_s^{a_2} + \cdots$$

最优值函数$V^*(s), Q^*(s,a)$

$V^*(s) = max\ V_\pi(s)$，对于所有的策略，找到一个策略使得V值函数最大

$Q^*(s,a) = max\ Q_\pi(s,a)$，对于所有的策略，找到一个策略使得Q值函数最大

一个很显然的方法是通过迭代的方式寻找最优策略，如果使得值函数最大，那么就选这个策略，把概率值设为1

$$\pi'(a|s) = \begin{cases} 1 & \text{if } a = \arg\max_a Q^\pi(s,\hat{a}) \\ 0 & \text{otherwise} \end{cases}$$

比如这个策略好，将其值设置为1
说明在分支中只选择这个动作后转移到其他状态。

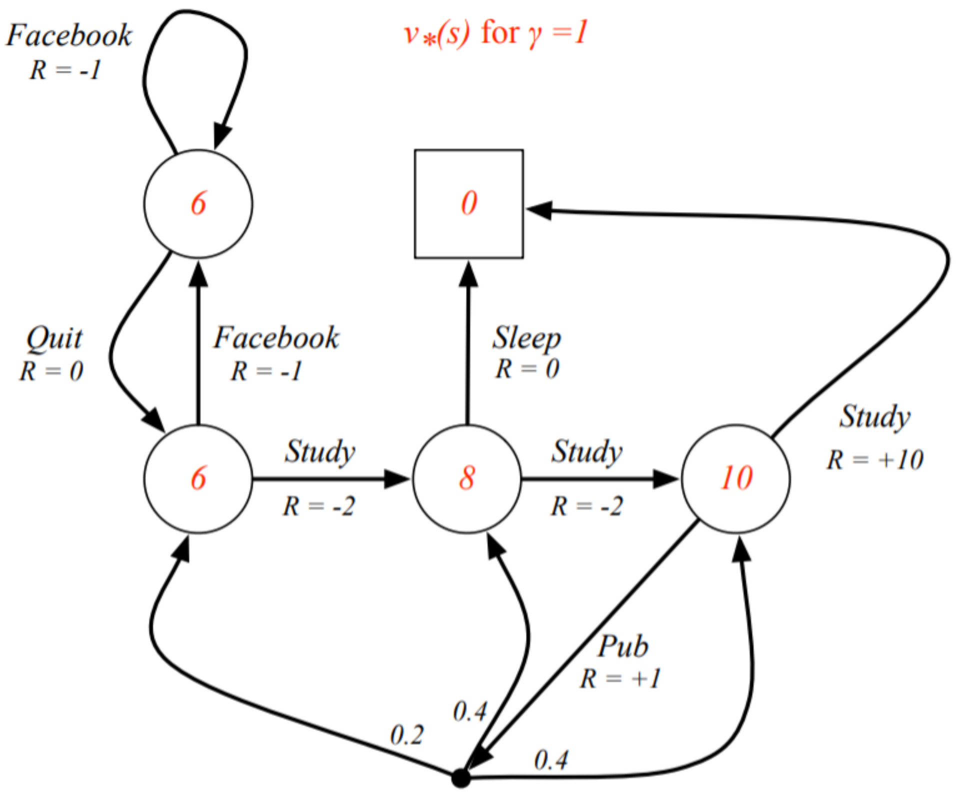
在增加了策略后，状态转移概率的计算公式为(计算每一个动作的策略概率再求和):

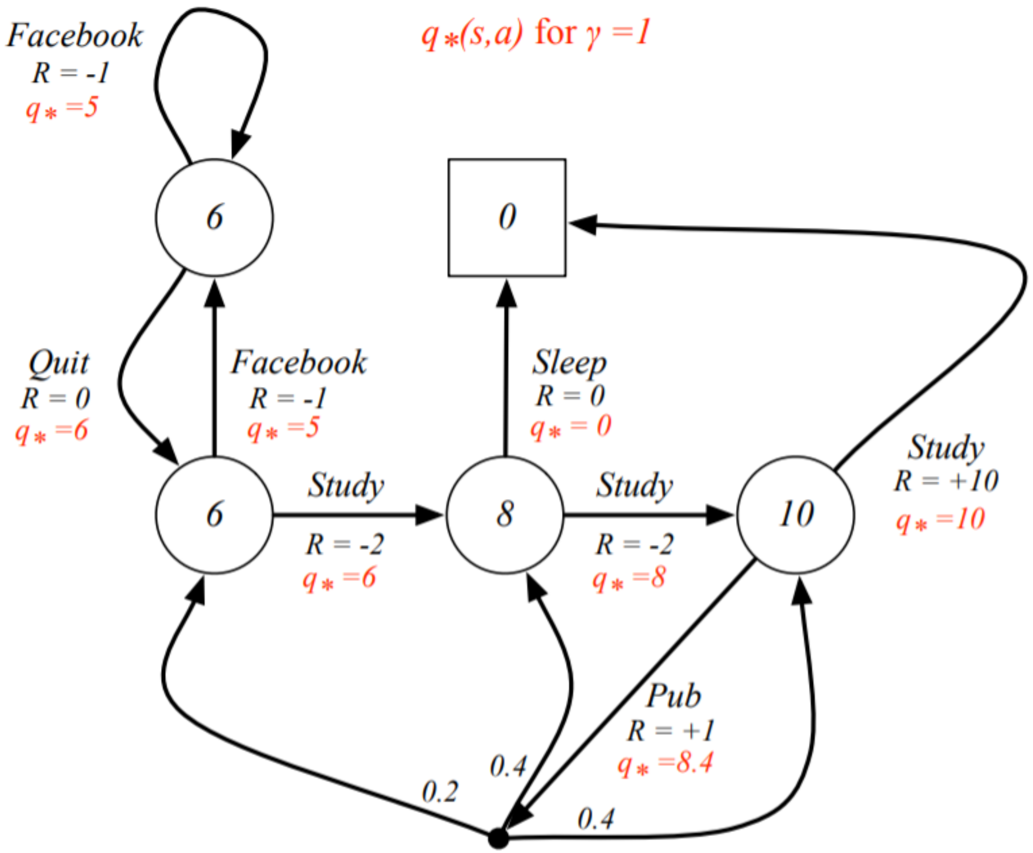
$$P_\pi(s' \mid s) = \sum_{a \in A} \pi(a \mid s) P(s' \mid a, s) = \boxed{\pi(a_1 \mid s)} P(s' \mid a_1, s) + \pi(a_2 \mid s) P(s' \mid a_2, s) + \cdots$$

最优值函数$V^*(s), Q^*(s,a)$

$V^*(s) = max\ V_\pi(s)$，对于所有的策略，找到一个策略使得V值函数最大

$Q^*(s,a) = max\ Q_\pi(s,a)$，对于所有的策略，找到一个策略使得Q值函数最大



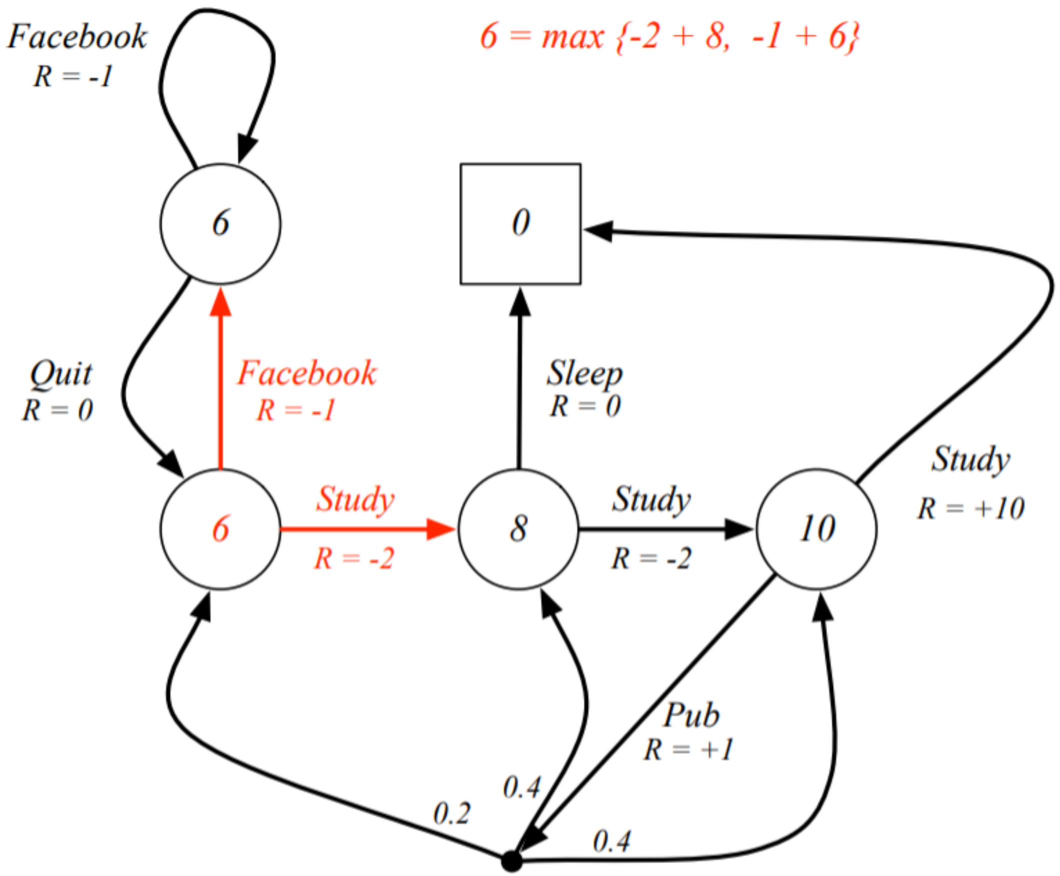


思考下如何计算?

最优值函数$V^*(s), Q^*(s,a)$

$V^*(s) = max\ V_\pi(s)$，对于所有的策略，找到一个策略使得V值函数最大

$Q^*(s,a) = max\ Q_\pi(s,a)$，对于所有的策略，找到一个策略使得Q值函数最大



$$V_\pi(s) = \sum_{a \in A} \pi(a \mid s)[R_s^a + \gamma \sum_{s \in s'} P(s' \mid a, s) V_\pi(s')]$$

验证方法：假设第二个状态转移到其他状态后的V值已知

那么$V_\pi(s_2) = \pi_1[-1+6] + (1-\pi_1)[-2+8] = 6 - \pi_1$
显然，最大值为6，即不选择π_1策略，直接向右转移到下一个状态。

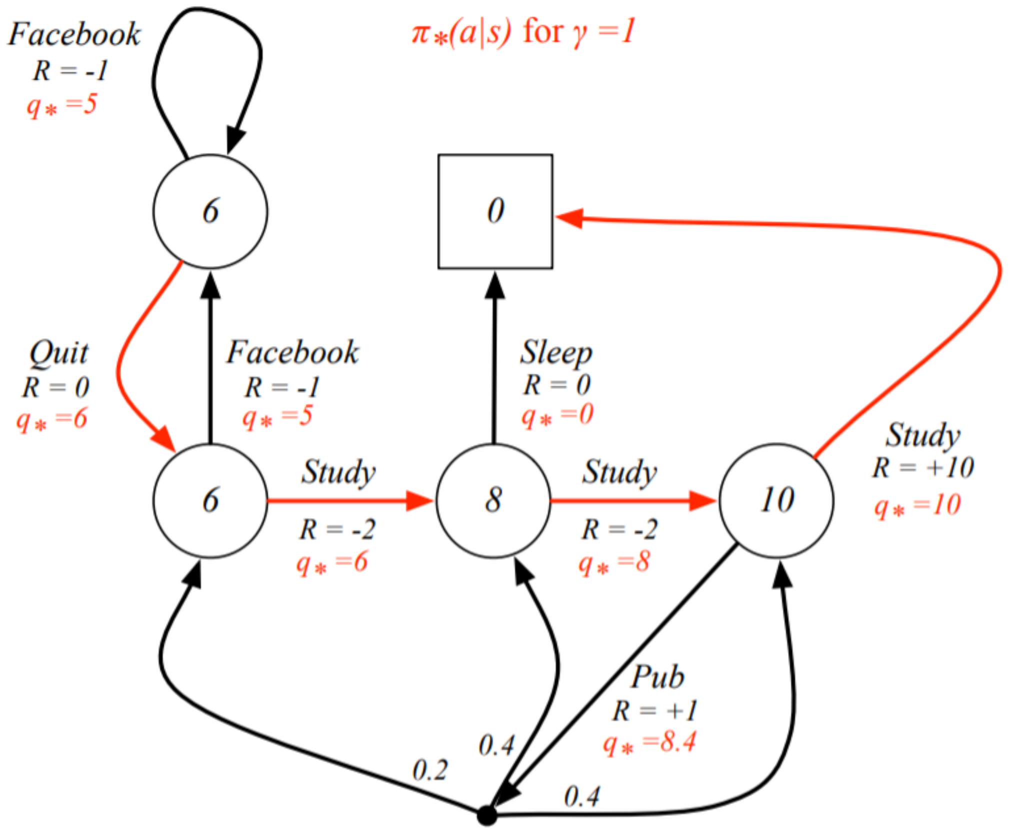
思考下**如何计算**？

最优值函数$V^*(s), Q^*(s,a)$

$V^*(s) = max\ V_\pi(s)$，对于所有的策略，找到一个策略使得V值函数最大

$Q^*(s,a) = max\ Q_\pi(s,a)$，对于所有的策略，找到一个策略使得Q值函数最大

$$Q_\pi(s,a) = R_s^a + \gamma \sum_{s \in s'} P(s' \mid a,s) V_\pi(s')$$



验证方法：假设第二个状态转移到其他状态后的V值已知

$$Q_\pi(s_2, FB) = -1 + 6p$$
$$Q_\pi(s_2, Study) = -2 + 8(1-p)$$

显然，各个动作的最大Q值分别5和6。那么采取最大的Q值依然是向右转移到下一个状态。

那么对应四个公式的**最优值**如何改变？推导**贝尔曼最优方程**。

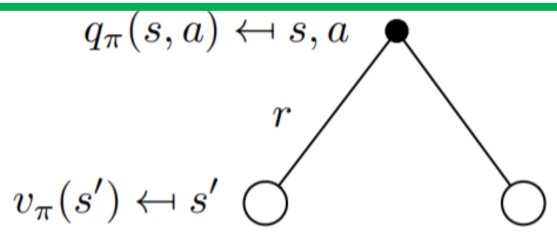
状态-动作值函数Q(s, a) $Q(s,a) = E[G_t \mid S_t = s, A_t = a]$

状态值函数V(s) $V(s) = E[G_t \mid S_t = s]$

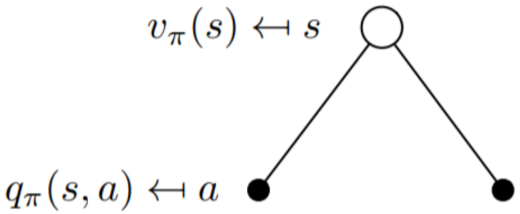
Bellman Equation

$$Q_\pi(s,a) = E[R_{t+1} + \gamma Q(S_{t+1}) \mid S_t = s, A_t = a]$$

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



$$Q_\pi(s,a) = R_s^a + \gamma \sum_{s \in s'} P(s' \mid a, s) V_\pi(s')$$

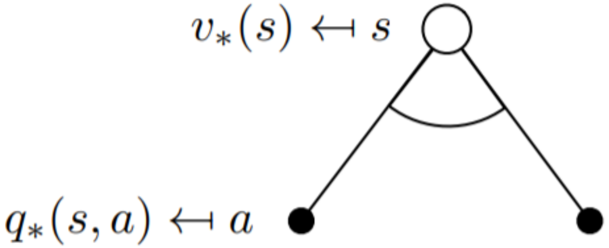


$$V_\pi(s) = \boxed{\sum_{a \in A} \pi(a \mid s)} Q_\pi(s,a)$$

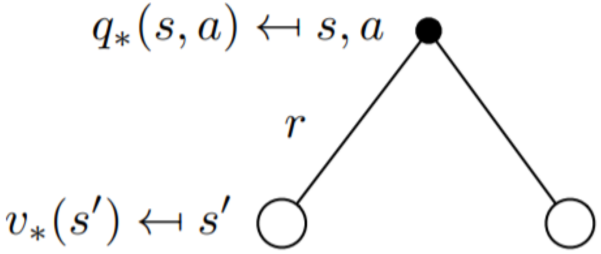
最优值只有一种决策，故取1

$$\boldsymbol{V^*(s) = max\, V_\pi(s)}$$

$$\boldsymbol{Q^*(s,a) = max\, Q_\pi(s,a)}$$



$$Q_*(s,a) = R_s^a + \gamma \sum_{s \in s'} P(s' \mid a, s) V_*(s')$$



$$V_*(s) = max Q_*(s,a) \ , a \in A$$

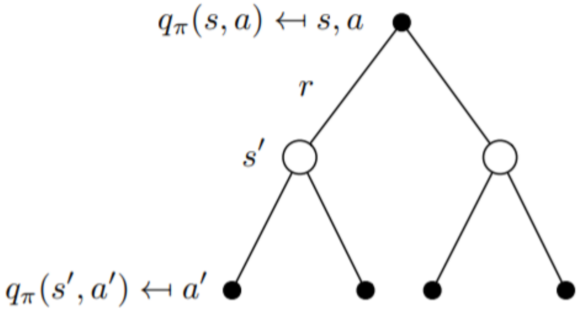
状态-动作值函数Q(s,a) $Q(s,a) = E[G_t \mid S_t = s, A_t = a]$

状态值函数V(s) $V(s) = E[G_t \mid S_t = s]$

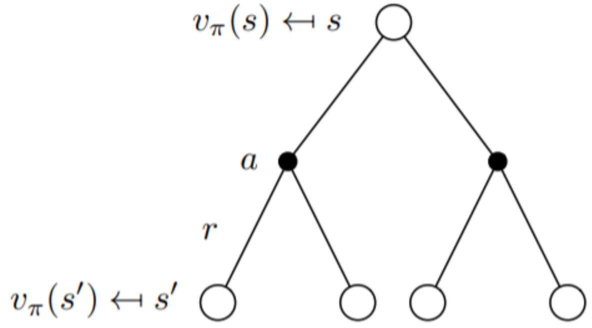
Bellman Equation

$$Q_\pi(s,a) = E[R_{t+1} + \gamma Q(S_{t+1}) \mid S_t = s, A_t = a]$$

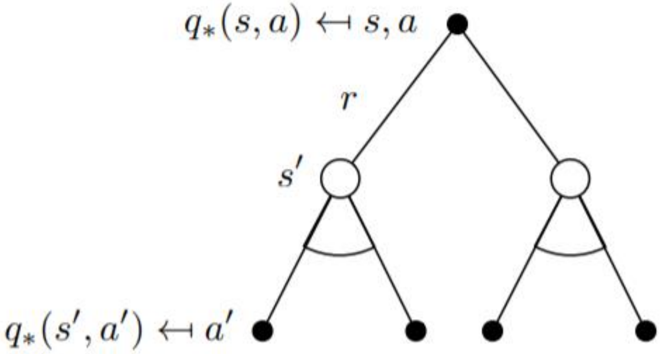
$$V(s) = E[R_{t+1} + \gamma V(S_{t+1}) \mid S_t = s]$$



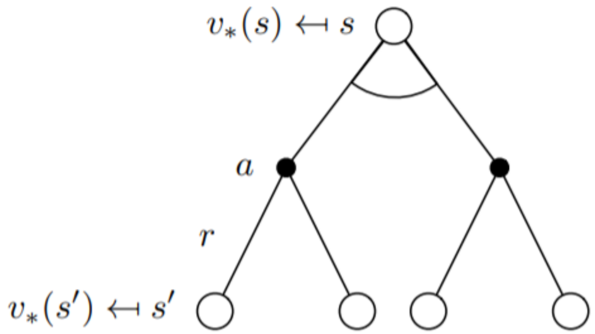
$$Q_\pi(s,a) = R_s^a + \gamma \sum_{s\in s'} P(s' \mid a, s) \boxed{\sum_{a\in A} \pi(a' \mid s')} Q_\pi(s',a')$$



$$V_\pi(s) = \boxed{\sum_{a\in A} \pi(a \mid s)} [R_s^a + \gamma \sum_{s\in s'} P(s' \mid a, s) V_\pi(s')]$$



$$Q_*(s,a) = R_s^a + \gamma \sum_{s\in s'} P(s' \mid a, s) max Q_*(s',a'), a' \in A$$



$$V_*(s) = max R_s^a + \gamma \sum_{s\in s'} P(s' \mid a, s) V_*(s')]$$

Bellman Optimality Equation