➢ 马尔科夫奖励过程 Markov Reward Process：

收获 Return

定义：收获 Gt 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总和。也有翻译成“收益”或"回报"。公式如下：。

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

其中衰减系数体现了未来的奖励在当前时刻的价值比例，在k+1时刻获得的奖励R在t时刻的体现出的价值是 Rt+1 ，γ接近0，则表明趋向于“近视”性评估；γ接近1则表明偏重考虑远期的利益。

马尔科夫奖励过程 Markov Reward Process：

价值函数 Value Function
价值函数给出了某一状态或某一行为的长期价值。

定义：一个马尔科夫奖励过程中某一状态的价值函数为从该状态开始的马尔可夫链收获的期望：

$$v(s) = E[G_t | S_t = s]$$

注：价值可以仅描述状态，也可以描述某一状态下的某个行为，在一些特殊情况下还可以仅描述某个行为。除了特别指出，约定用状态价值函数或价值函数来描述针对状态的价值；用行为价值函数来描述某一状态下执行某一行为的价值，严格意义上说行为价值函数是“状态行为对”价值函数的简写。

马尔科夫奖励过程 Markov Reward Process：

为方便计算，把“学生马尔科夫奖励过程”示例图表示成下表的形式。表中第二行对应各状态的即时奖励值，蓝色区域数字为状态转移概率，表示为从所在行状态转移到所在列状态的概率：

States	C1	C2	C3	Pass	Pub	FB	Sleep
Rewards	-2	-2	-2	10	1	-1	0
C1		0.5				0.5	
C2			0.8				0.2
C3				0.6	0.4		
Pass							1
Pub	0.2	0.4	0.4				
FB	0.1					0.9	
Sleep							1

马尔科夫奖励过程 Markov Reward Process：

考虑如下4个马尔科夫链。现计算当γ= 1/2时，在t=1时刻时状态 S1 的收获分别为：

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

马尔科夫奖励过程 Markov Reward Process：

考虑如下4个马尔科夫链。现计算当γ= 1/2时，在t=1时刻（S1=C1）时状态 S1 的收获分别为：

C1 C2 C3 Pass Sleep	G1 = -2 + (-2)*1/2 + (-2)*1/4 + 10*1/8 + 0*1/16 = -2.25
C1 FB FB C1 C2 Sleep	G1 = -2 + (-1)*1/2 + (-1)*1/4 + (-2)*1/8 + (-2)*1/16 + 0*1/32 = -3.125
C1 C2 C3 Pub C2 C3 Pass Sleep	G1 = -2 + (-2)*1/2 + (-2)*1/4 + (1)*1/8 + (-2)*1/16 + … = -3.41
C1 FB FB C1 C2 C3 Pub C1 FB FB FB C1 C2 C3 Pub C2 Sleep	G1 = -2 + (-1)*1/2 + (-1)*1/4 + (-2)*1/8 + (-2)*1/16 + (-2)*1/32 + … = -3.20

➢ 马尔科夫奖励过程 Markov Reward Process：

考虑如下4个马尔科夫链。现计算当γ= 1/2时，在t=1时刻（[公式]）时状态 S1 的收获分别为：

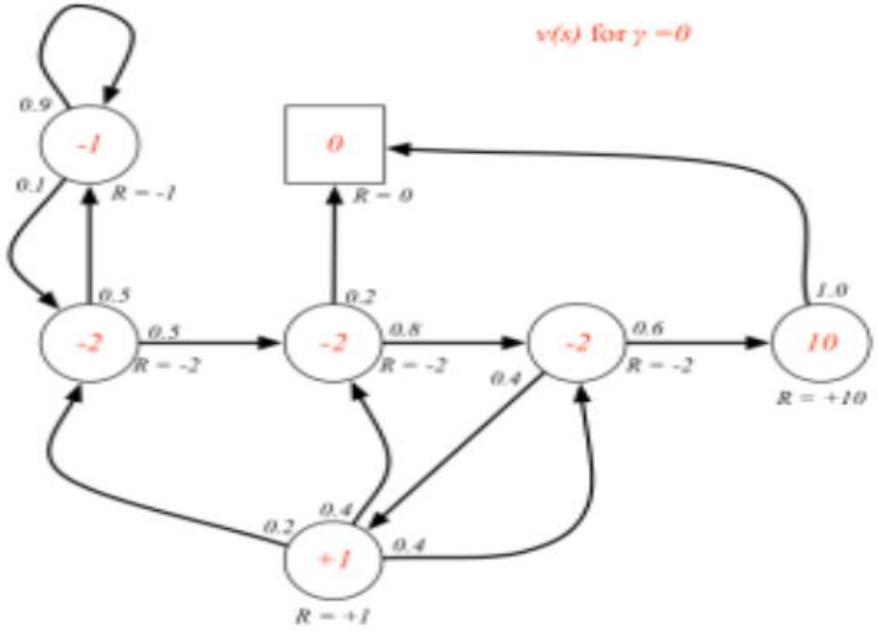
C1 C2 C3 Pass Sleep	G1 = -2 + (-2)*1/2 + (-2)*1/4 + 10*1/8 + 0*1/16 = -2.25
C1 FB FB C1 C2 Sleep	G1 = -2 + (-1)*1/2 + (-1)*1/4 + (-2)*1/8 + (-2)*1/16 + 0*1/32 = -3.125
C1 C2 C3 Pub C2 C3 Pass Sleep	G1 = -2 + (-2)*1/2 + (-2)*1/4 + (1)*1/8 + (-2)*1/16 + … = -3.41
C1 FB FB C1 C2 C3 Pub C1 FB FB FB C1 C2 C3 Pub C2 Sleep	G1 = -2 + (-1)*1/2 + (-1)*1/4 + (-2)*1/8 + (-2)*1/16 + (-2)*1/32 + … = -3.20

从上表也可以理解到，收获是针对一个马尔科夫链中的某一个状态来说的。

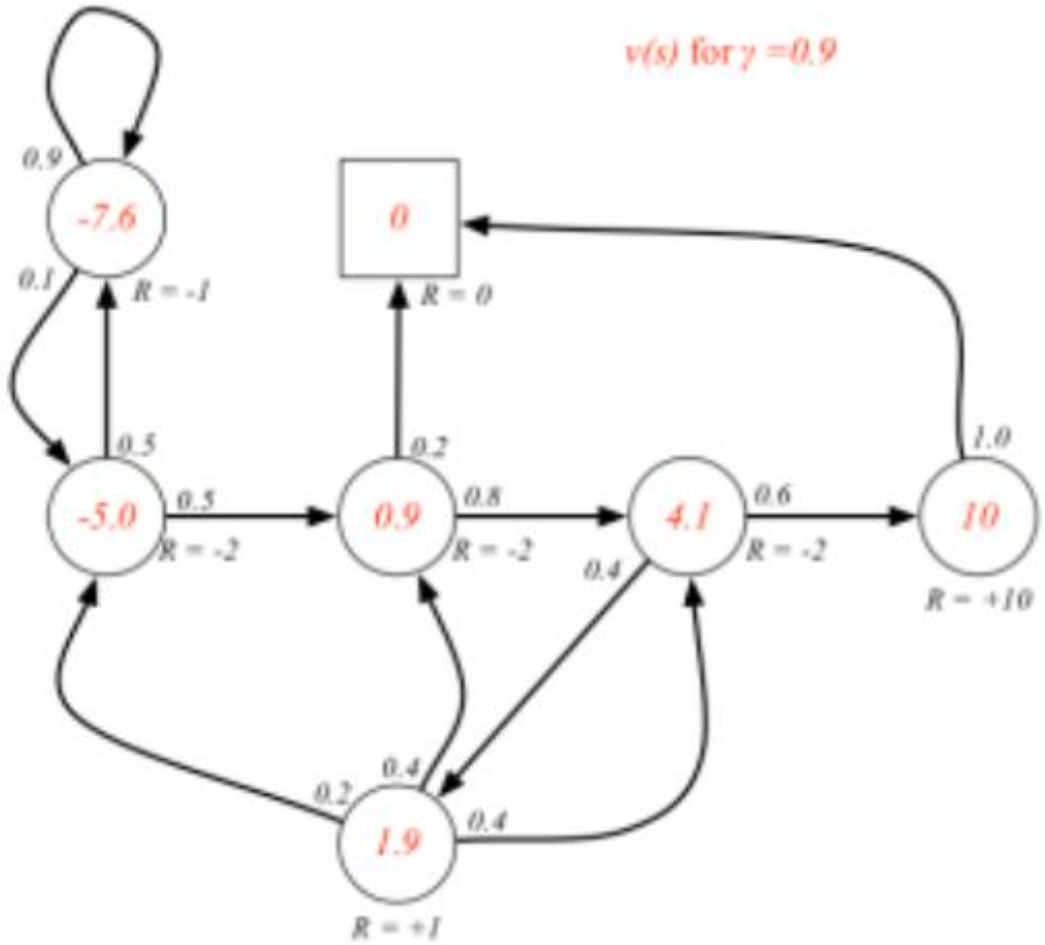
马尔科夫奖励过程 Markov Reward Process：

当γ= 0时，上表描述的MRP中，各状态的即时奖励就与该状态的价值相同。当γ≠ 0时，各状态的价值需要通过计算得到，这里先给出γ分别为0, 0.9,和1三种情况下各状态的价值，如下图所示。

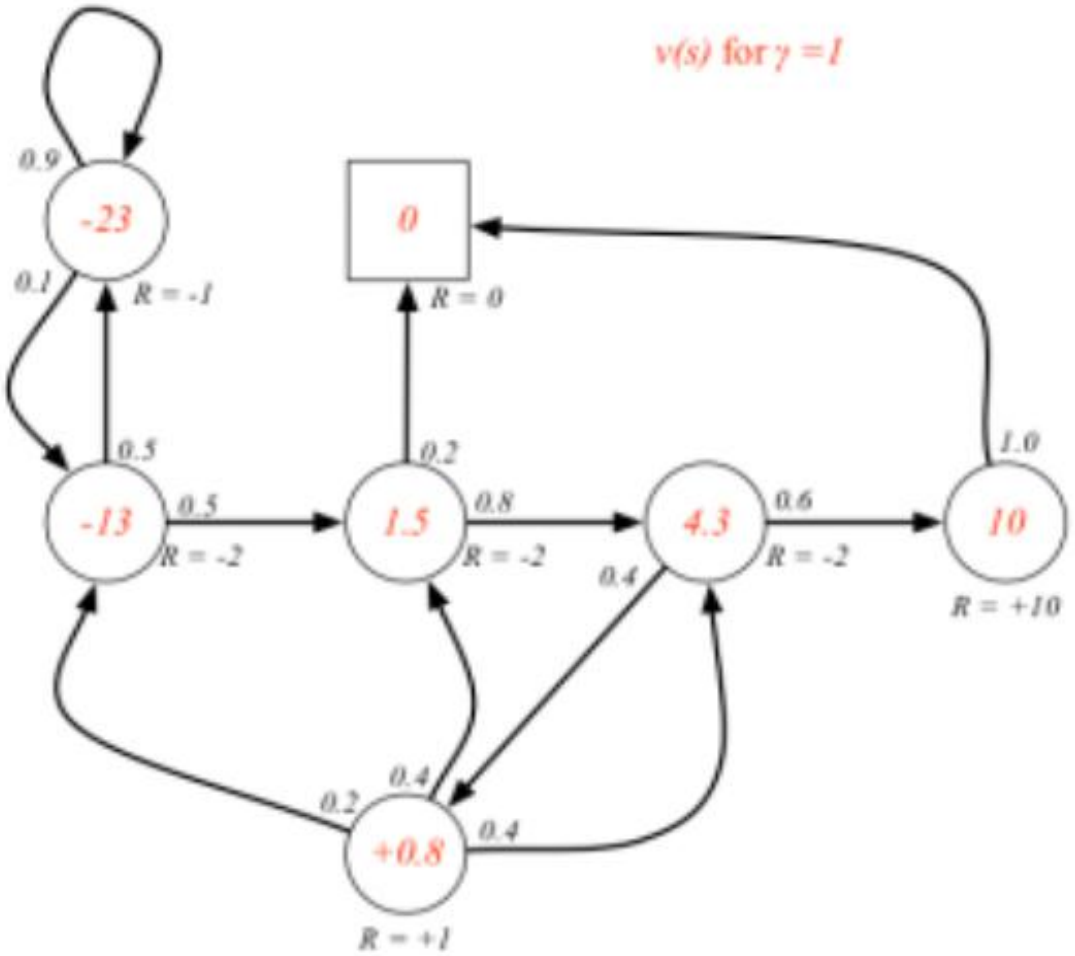
各状态圈内的数字表示该状态的价值，圈外的R=-2等表示的是该状态的即时奖励。



- 马尔科夫奖励过程 Markov Reward Process：



- 马尔科夫奖励过程 Markov Reward Process：



- 马尔科夫奖励过程 Markov Reward Process：

  各状态价值的确定是很重要的，RL的许多问题可以归结为求状态的价值问题。因此如何求解各状态的价值，也就是寻找一个价值函数（从状态到价值的映射）就变得很重要了。

- 价值函数的推导：

Bellman方程 - MRP

先尝试用价值的定义公式来推导看看能得到什么：

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

- 价值函数的推导：

这个推导过程相对简单，仅在导出最后一行时，将 G t+1变成了 V（S t+1）。其理由是收获的期望等于收获的期望的期望。下式是针对MRP的Bellman方程：

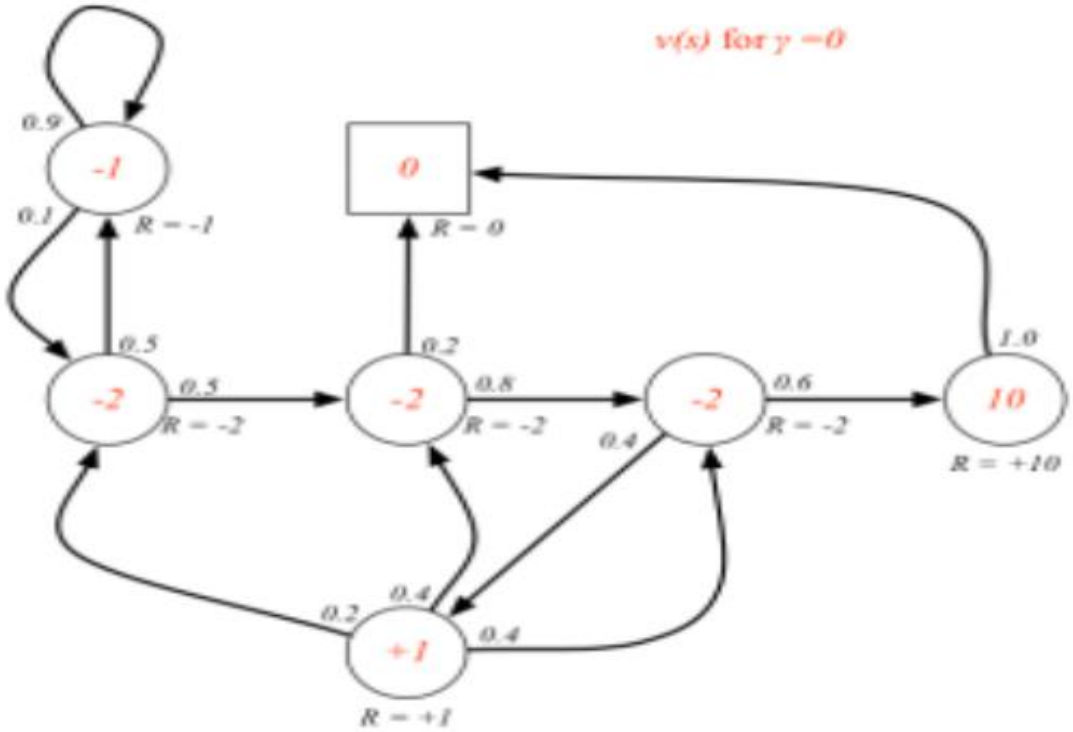
$$v(s) = \mathbb{E}\left[R_{t+1} + \gamma v(S_{t+1}) \mid S_t = s\right]$$

通过方程可以看出 v(s)由两部分组成，一是该状态的即时奖励期望，即时奖励期望等于即时奖励，因为根据即时奖励的定义，它与下一个状态无关；另一个是下一时刻状态的价值期望，可以根据下一时刻状态的概率分布得到其期望。如果用s’表示s状态下一时刻任一可能的状态，那么Bellman方程可以写成：

$$v(s) = \mathcal{R}_s + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'} v(s')$$

- 方程的解释：

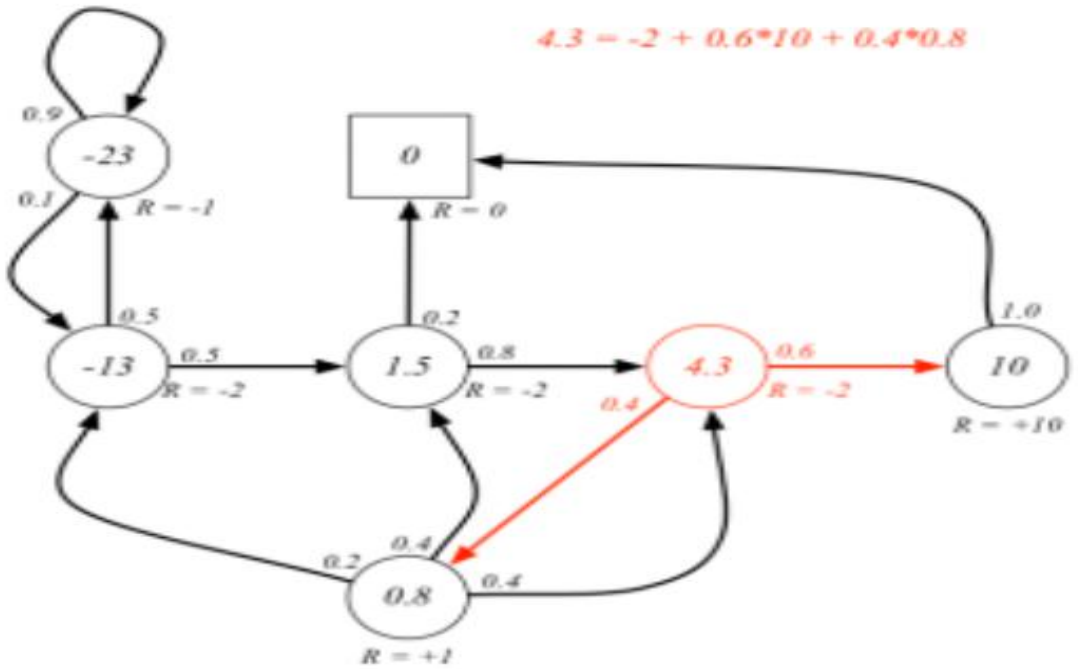
根据下图已经给出了γ=1时各状态的价值，状态 C3的价值可以通过状态Pub和Pass的价值以及他们之间的状态转移概率来计算：



- 方程的解释：

下图已经给出了γ=1时各状态的价值，状态 C3的价值可以通过状态Pub和Pass的价值以及他们之间的状态转移概率来计算：

$$4.3 = -2 + 1.0 * (0.6 * 10 + 0.4 * 0.8)$$



马尔科夫决定过程 Markov Decision Process：

相较于马尔科夫奖励过程，马尔科夫决定（决策）过程多了一个行为集合A，它是这样的一个元组: <S, A, P, R, γ>。看起来很类似马尔科夫奖励过程，但这里的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表示的是有限的行为的集合。具体的数学表达式如下：

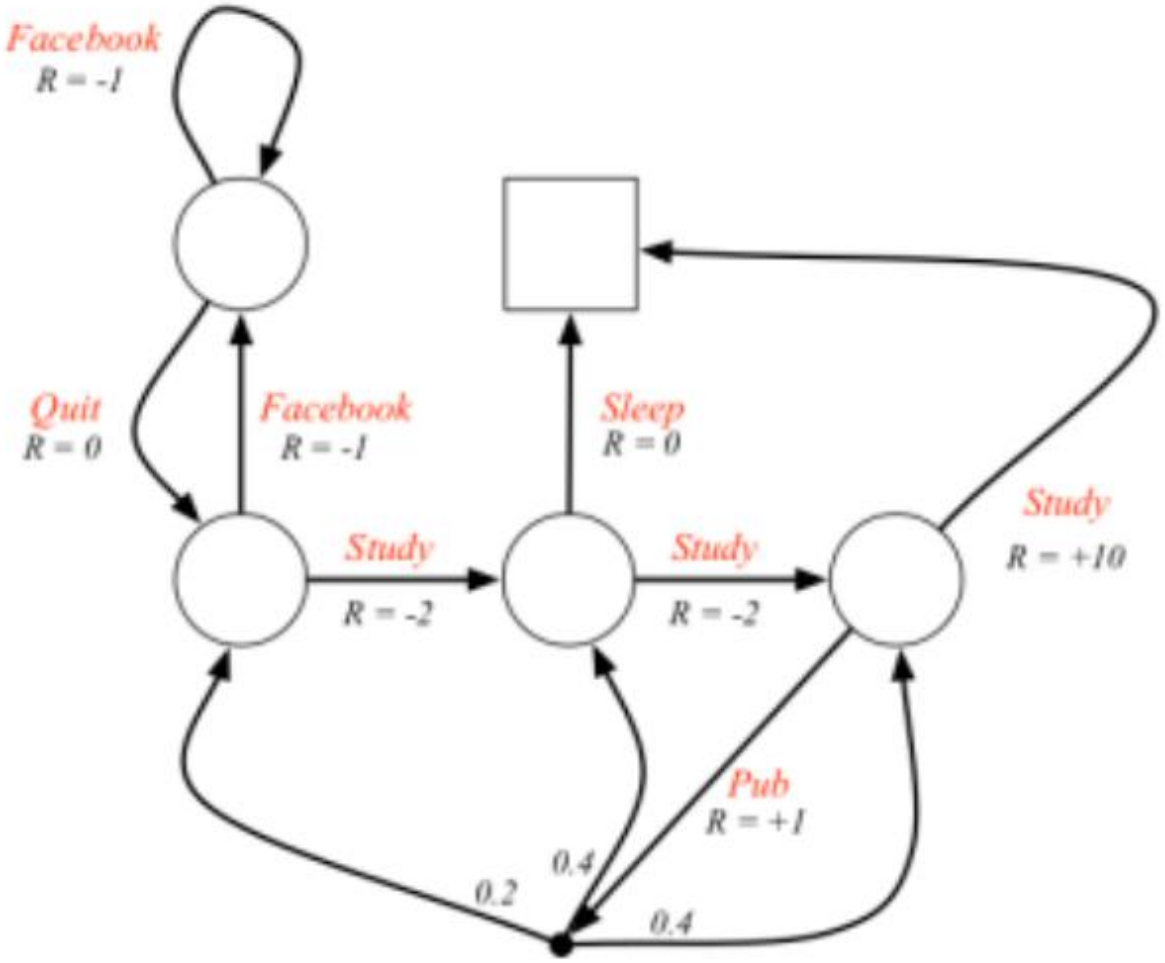
$$\mathcal{P}_{ss'}^{a} = \mathbb{P}\left[S_{t+1} = s' \mid S_t = s, A_t = a\right]$$

$$\mathcal{R}_{s}^{a} = \mathbb{E}\left[R_{t+1} \mid S_t = s, A_t = a\right]$$

马尔科夫决定过程 Markov Decision Process：

下图给出了一个可能的MDP的状态转化图。图中红色的文字表示的是采取的行为，而不是先前的状态名。对比之前的学生MRP示例可以发现，即时奖励与行为对应了，同一个状态下采取不同的行为得到的即时奖励是不一样的。由于引入了Action，容易与状态名混淆，因此此图没有给出各状态的名称；此图还把Pass和Sleep状态合并成一个终止状态；另外当选择”pub”这个动作时，主动进入了一个临时状态（图中用黑色小实点表示），随后被动的被环境按照其动力学分配到另外三个状态，也就是说此时Agent没有选择权决定去哪一个状态。

- 马尔科夫决定过程 Markov Decision Process：



策略Policy：

策略 π 是概率的集合或分布，其元素 $\pi(a|s)$ 为对过程中的某一状态s采取可能的行为a的概率。用 $\pi(a|s)$ 表示。

$$\pi(a|s) = \mathbb{P}\left[A_t = a \mid S_t = s\right]$$

一个策略完整定义了个体的行为方式，也就是说定义了个体在各个状态下的各种可能的行为方式以及其概率的大小。Policy仅和当前的状态有关，与历史信息无关；同时某一确定的Policy是静态的，与时间无关；但是个体可以随着时间更新策略。

马尔科夫决定过程 Markov Decision Process：

当给定一个MDP: $M =< S, A, P, R, \gamma >$ 和一个策略π，那么状态序列 $S_1, S_2, \ldots$ 是一个马尔科夫过程 $< S, P^{\pi} >$ ；同样，状态和奖励序列 $S_1, R_2, S_2, R_3, S_3, \ldots$ 是一个马尔科夫奖励过程 $< S, P^{\pi}, R^{\pi}, \gamma >$ ，并且在这个奖励过程中满足下面两个方程：

$$\mathcal{P}^{\pi}_{s,s'} = \sum_{a \in \mathcal{A}} \pi(a|s) \mathcal{P}^{a}_{ss'}$$

用文字描述是这样的，在执行策略 π 时，状态从s转移至 s' 的概率等于一系列概率的和，这一系列概率指的是在执行当前策略时，执行某一个行为的概率与该行为能使状态从s转移至s′ 的概率的乘积。

马尔科夫决定过程 Markov Decision Process：

奖励函数表示如下：

$$\mathcal{R}_s^\pi = \sum_{a \in \mathcal{A}} \pi(a|s)\mathcal{R}_s^a$$

用文字表述是这样的：当前状态s下执行某一指定策略得到的即时奖励是该策略下所有可能行为得到的奖励与该行为发生的概率的乘积的和。

策略在MDP中的作用相当于agent可以在某一个状态时做出选择，进而有形成各种马尔科夫过程的可能，而且基于策略产生的每一个马尔科夫过程是一个马尔科夫奖励过程，各过程之间的差别是不同的选择产生了不同的后续状态以及对应的不同的奖励。

➢ 马尔科夫决定过程 Markov Decision Process：

基于策略π的价值函数：

定义 $v_\pi(s)$ 是在MDP下的基于策略π的**状态价值函数**，表示从状态s开始，**遵循当前策略**时所获得的收获的期望；或者说在执行当前策略π时，衡量个体处在状态s时的价值大小。数学表示如下：

$$v_\pi(s) = \mathbb{E}_\pi [G_t \mid S_t = s]$$

注意策略是静态的、关于整体的概念，不随状态改变而改变；变化的是在某一个状态时，依据策略可能产生的具体行为，因为具体的行为是有一定的概率的，策略就是用来描述各个不同状态下执行各个不同行为的概率。

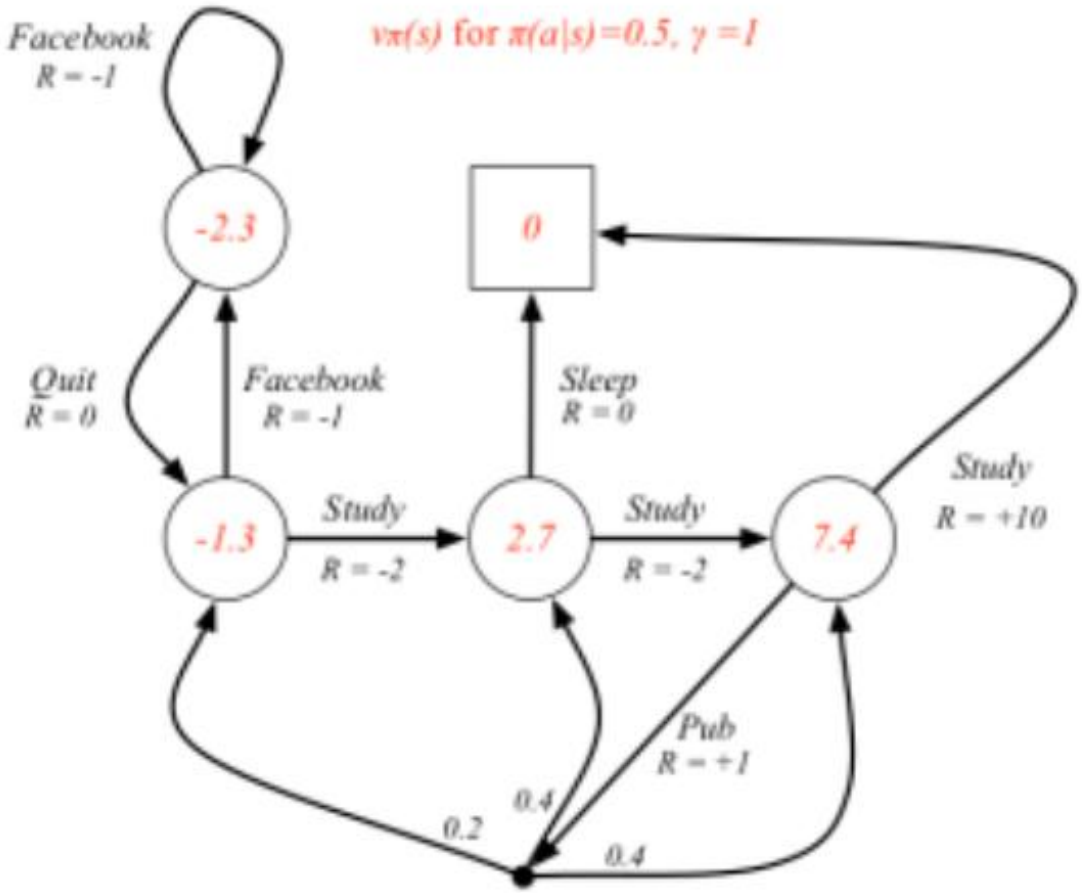
马尔科夫决定过程 Markov Decision Process：

定义 $q_\pi(s,a)$ 为**行为价值函数，表示**在执行策略π时，对当前状态s执行某一具体行为a所能的到的收获的期望；或者说在遵循当前策略π时，衡量对当前状态执行行为a的价值大小。行为价值函数一般都是与某一特定的状态相对应的，更精细的描述是**状态行为对**价值函数。行为价值函数的公式描述如下：

$$q_\pi(s,a) = \mathbb{E}_\pi\left[G_t \mid S_t = s, A_t = a\right]$$

- 马尔科夫决定过程 Markov Decision Process：

下图用例子解释了行为价值函数



➢ 马尔科夫决定过程 Markov Decision Process：

Bellman期望方程 Bellman Expectation Equation

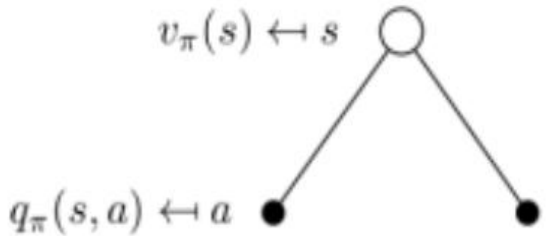
MDP下的状态价值函数和行为价值函数与MRP下的价值函数类似，可以改用下一时刻状态价值函数或行为价值函数来表达，具体方程如下：

$$v_\pi(s) = \mathbb{E}_\pi \left[R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = s \right]$$

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

马尔科夫决定过程 Markov Decision Process：

- $v_\pi(s)$ **和** $q_\pi(s,a)$ **的关系**



上图中，空心较大圆圈表示状态，黑色实心小圆表示的是动作本身，连接状态和动作的线条仅仅把该状态以及该状态下可以采取的行为关联起来。可以看出，在遵循策略π时，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。

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

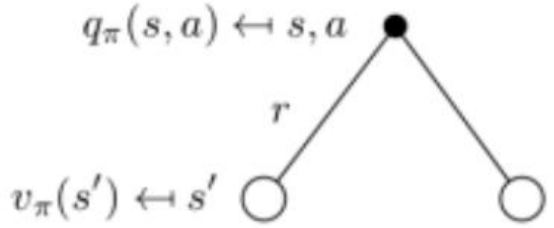
- 马尔科夫决定过程 Markov Decision Process：

类似的，一个行为价值函数也可以表示成状态价值函数的形式：

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

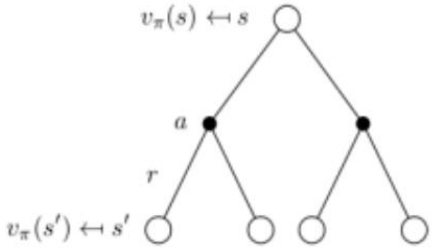
马尔科夫决定过程 Markov Decision Process：

它表明，一个某一个状态下采取一个行为的价值，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。



如果组合起来，可以得到下面的结果：

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



- 马尔科夫决定过程 Markov Decision Process：

学生MDP示例：