# Highlights of Calculus

FrankZhou-jun*

2019 年 11 月 23 日

## 1 马尔科夫决策过程

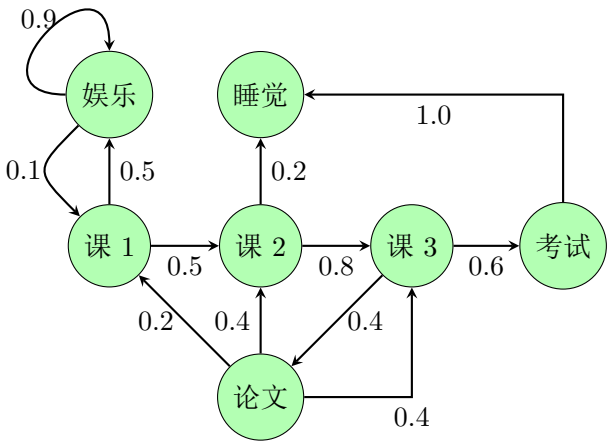
### 1.1 马尔科夫性

马尔科夫性是指环境的下一个状态 $s_{t+1}$ 只与当前的状态相关 $s_t$，即：

$$P[s_{t+1}|s_t] = P[s_{t+1}|s_1, s_1, ..., s_n]$$

### 1.2 马尔科夫过程

下图为一个学生的 7 中状态 {娱乐，课程 1，课程 2，课程 3，考过，睡觉，论文}，7 种状态之间转移的概率已在图中标出。马尔科夫过程是一个包含状态和状态转移概率矩阵的二元组 $(\mathbf{S}, \mathbf{P})$。




该生从课程 1 开始的一天的学习状态序列可能为：

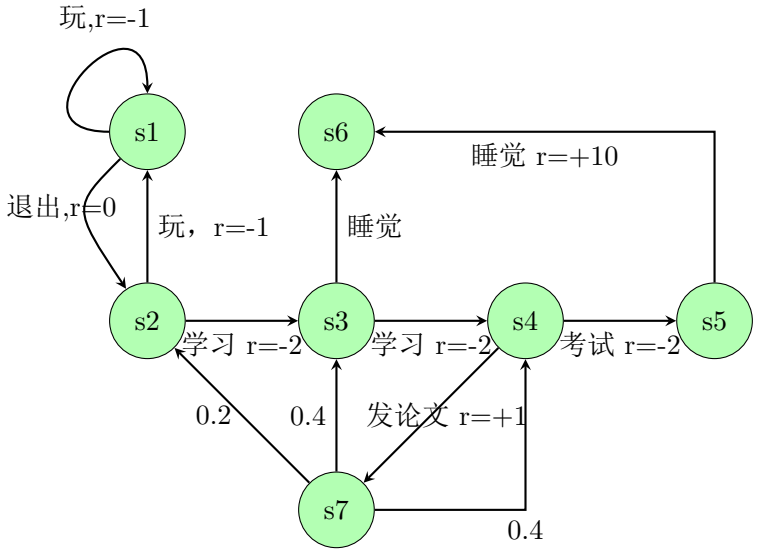
1. { 课程 1 – 课程 2 –课程 3 – 考过 – 睡觉 }
2. { 课程 1 – 课程 2 – 睡觉 }
3. ...

以上的状态序列称为**马尔科夫链** ，可以看到从某个状态出发可能存在多条的马尔科夫链。

### 1.3 马尔科夫决策过程

与马尔科夫过程不同的是，马尔科夫决策过程由有限状态集，有限动作集，状态转移概率，回报函数，折扣因子 $(\mathbf{S}, \mathbf{A}, \mathbf{P}, r, \gamma)$ 组成。则之前的马尔科夫过程转变为马尔科夫决策过程如下：

*研究方向：信号处理，机械故障诊断，深度学习，强化学习，邮箱:zhoujun14@yeah.net




强化学习的目标是给定一个马尔科夫决策过程，然后得到最优策略，策略其实就是给定一个状态，然后得到动作的分布，是一个状态到动作的映射，用符号 $\pi$ 表示，公式如下：

$$\pi(a|s) = p[\mathbf{A}_t = a|\mathbf{S}_t = s]$$

假设重状态 $s_1$ 出发，则该生的状态序列可能为：

1. $s_1 - -s_2 - -s_3 - -s_6$
2. $s_1 - -s_2 - -s_3 - -s_4 - -s_5 - -s_6$
3. ...

可以看到由于策略 $\pi$ 是随机的，可能存在多条马尔科夫决策链，在计算积累回报：

$$\mathbf{G}_t = \mathbf{R}_{t+1} + \gamma\mathbf{R}_{t+1} + \gamma^2\mathbf{R}_{t+2} + ... = \sum_{k=0}^{\infty}\gamma^k\mathbf{R}_{t+k+1}$$

## 1.4 状态值函数与状态动作值函数

由于存在多个马尔科夫决策链，则会产生多个积累回报值，因为策略 $\pi$ 是随机的，所以产生的积累回报也是随机的。但是在评定状态的价值时是需要一个确定的量才可以的，所以使用当前状态下产生的所有马尔科夫决策链的期望来代表当前的状态值。状态值的计算函数如下：

$$v_\pi(s) = E_\pi[\sum_{k=0}^{\infty}\gamma^k\mathbf{R}_{t+k+1}|\mathbf{S}_t = s]$$

还可以得到:

$$q_\pi(s,a) = E_\pi[\sum_{k=0}^{\infty}\gamma^k\mathbf{R}_{t+k+1}|\mathbf{S}_t = s, \mathbf{A}_t = a]$$

状态值函数与状态动作值函数的贝曼方程：

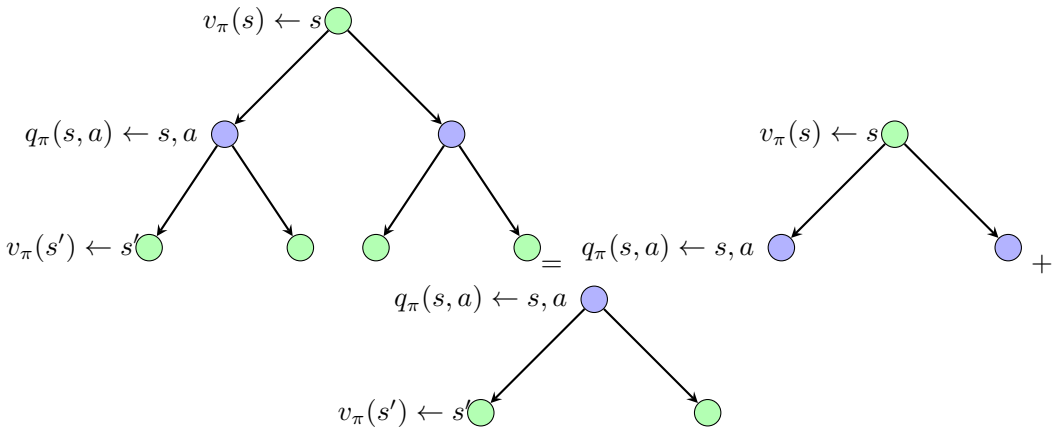
$$\begin{aligned} v_\pi(s) &= E_\pi[\mathbf{G}_t|\mathbf{S}_t = s] \\ &= E_\pi[\mathbf{R}_{t+1} + \gamma v_\pi(\mathbf{S}_{t+1})|\mathbf{S}_t = s] \end{aligned}$$

状态动作值的贝曼方程：

$$q_\pi(s,a) = E_\pi[\mathbf{R}_{t+1} + \gamma q_\pi(\mathbf{S}_{t+1}, \mathbf{A}_{t+1})|\mathbf{S}_t = s, \mathbf{A}_t = a]$$

状态值函数与状态动作值函数的关系如下，状态值可以用动作值函数的期望表示：

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




状态动作值函数也可以用状态值函数的关系表示

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

结合上两式

$$v_\pi(s) = \sum_{a \in \mathbf{A}} \pi(a|s) \big[ \mathbf{R}_s^a + \gamma \sum_{s' \in \mathbf{S}} \mathbf{P}_{ss'}^a v_\pi(s') \big]$$

下一次的状态值函数可表示为：

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

由此可以得到

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

在后面的 Qlearning 过程中可以看到此公式的应用。

## 1.5 最优策略

计算状态值函数或动作值函数的目的是从数据中得到最优策略，由于存在多条马尔科夫决策链，而每一条马尔科夫决策链对应一个策略，每一个策略的好坏是由值函数来衡量的，所以最优策略对应的是所有策略中最大的值函数。可以得到最优状态值函数：

$$v^*(s) = \max_\pi \ v_\pi(s)$$

最优状态动作值函数：

$$q^*(s,a) = \max_\pi \ q_\pi(s,a)$$