强化学习入门基础-马尔达夫决策过程（MDP）

作者：YJLAugus 博客： https://www.cnblogs.com/yjlaugus

MDP背景介绍

Random Variable

随机变量（Random Variable），通常用大写字母来表示一个随机事件。比如看下面的例子：

X: 河水是咸的

Y: 井水是甜的

很显然，X, Y两个随机事件是没有关系的。也就是说X和Y之间**是相互独立**的。记作：

$$X \perp Y$$

Stochastic Process

对于一类随机变量来说，它们之间存在着某种关系。比如：

S_t：表示在 t 时刻某支股票的价格，那么 S_{t+1} 和 S_t 之间一定是有关系的，至于具体什么样的关系，这里原先不做深究，但有一点可以确定，两者之间一定存在的一种关系。随着时间 t 的变化，可以写出下面的形式：

$$\dots S_t, S_{t+1}, S_{t+2} \dots$$

这样就生成了一组随机变量，它们之间存在着一种相当复杂的关系，也就是说，各个随机变量之间存在着关系，即不相互独立。由此，我们会把按照某个时间或者次序上的一组不相互独立的随机变量的这样一个整体作为研究对象。这样的话，也就引出了另外的一个概念：**随机过程（Stochastic Process）**。也就是说随机过程的研究对象不在是单个的随机变量，而是一组随机变量，并且这一组随机变量之间存在着一种非常紧密的关系（不相互独立）。记作：

$$\{S_t\}_{t=1}^{\infty}$$

Markov Chain/Process

马尔科夫链（Markov Chain）即马尔可夫过程，是一种特殊的随机过程——具备马尔可夫性的随机过程。

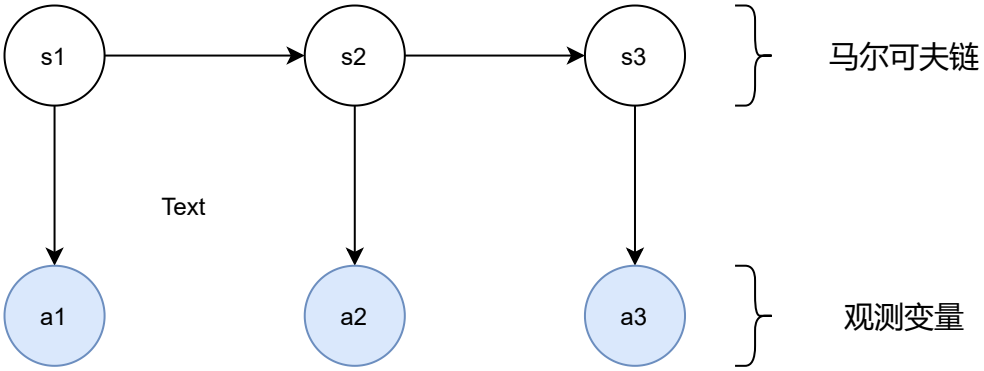
- 马尔可夫性：（Markov Property）：还是上面股票的例子，如果满足 $P(S_{t+1} \mid S_t, S_{t-1} \dots S_1) = P(S_{t+1} \mid S_t)$，即具备了马尔可夫性。简单来说，$S_{t+1}$ 和S_t之间存在关系，和以前的时刻的没有关系，即只和“最近的状态”有关系。
- 现实例子：下一个时刻仅依赖于当前时刻，跟过去无关。比如：一个老师讲课，明天的讲课状态一定和今天的状态最有关系，和过去十年的状态基本就没关系了。
- 最主要考量：为了简化计算。$P(S_{t+1} \mid S_t, S_{t-1} \dots S_1) = P(S_{t+1} \mid S_t)$ 如果 S_{t+1} 和 $S_t, S_{t-1} \dots S_1$ 都有关系的话，计算的话就会爆炸了。

马尔可夫链/过程 即满足马尔可夫性质的随机过程，记作：

$$P(S_{t+1}) \mid S_t, S_{t-1} \dots S_1) = P(S_{t+1} \mid S_t)$$

State Space Model

状态空间模型（State Space Model），常应用于 HMM,Kalman Filterm Particle Filter，关于这几种这里不做讨论。在这里就是指马尔可夫链 + 观测变量，即 `Markov Chain + Obervation`

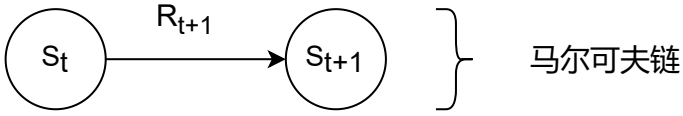


如上图所示，s1-s2-s3为马尔可夫链，a1, a2, a3为观测变量，以a2为例，a2只和s2有关和s1, s3无关。状态空间模型可以说是由马尔可夫链演化而来的模型。记作：

Markov Chain + Obervation

Markov Reward Process

马尔可夫奖励过程（Markov Reward Process），即马尔可夫链+奖励，即：`Markov Chain + Reward`。如下图：



举个例子，比如说你买了一支股票，然后你每天就会有“收益”，当然了这里的收益是泛化的概念，收益有可能是正的，也有可能是负的，有可能多，有可能少，总之从今天的状态S_t 到明天的状态 S_{s+1} ，会有一个 `reward`。记作：

Markov Chain + Reward

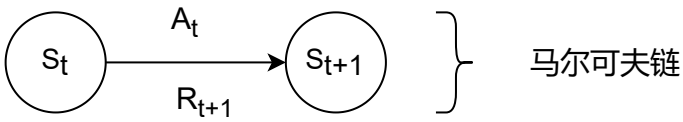
Markov Decision Process

马尔可夫决策过程（Markov Decision Process），即马尔可夫奖励过程的基础上加上 `action`，即：`Markov Chain + Reward + action`。如果还用刚才的股票为例子的话，我们只能每天看到股票价格的上涨或者下降，然后看到自己的收益，但是无法操作股票的价格的，只有看到份，只是一个“小散户”。这里的马尔可夫决策过程相当于政策的制定者，相当于一个操盘手，可以根据不同的状态而指定一些政策，也就相当于 action。

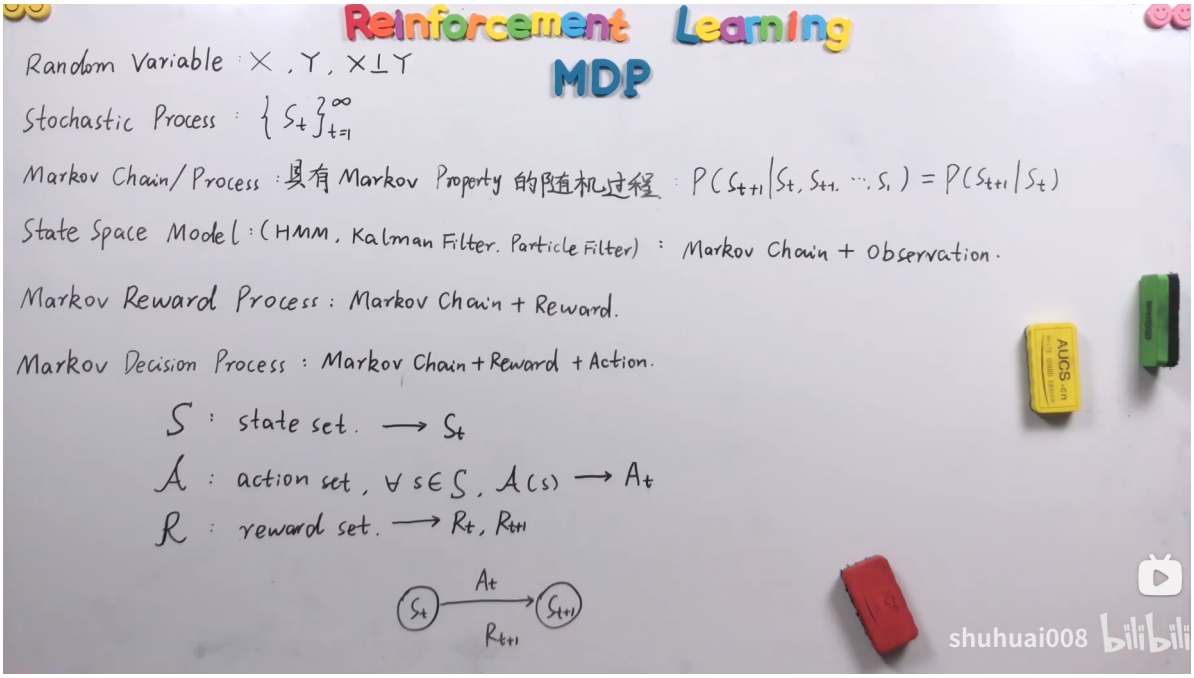
在马尔可夫决策过程中，所有的**状态**是我们看成离散的，有限的集合。所有的**行为**也是离散有限的集合。记作：

$$
\begin{array}{lll}
S: & state \quad set & S_t \\
A: & action \quad set, \quad \forall s \in S, A_{(s)} & A_t \\
R: & reward \quad set & R_t, R_{(t+1)}
\end{array}
$$

对于上述公式简单说明，S_t 用来表示某一个时刻的状态。$A_{(s)}$ 表示在**某一个状态**时候的行为，这个行为一定是基于某个状态而言的，假设在t 时刻的状态为S 此时的 `action` 记作 A_t。R_t 和 $R_{(t+1)}$ 只是记法不同，比如下面的例子：从S_t状态经过 A_t 到S_{t+1}状态，获得的奖励一般记作$R_{(t+1)}$。 也就是说S_t， A_t ， $R_{(t+1)}$ 是配对使用的。



Summary



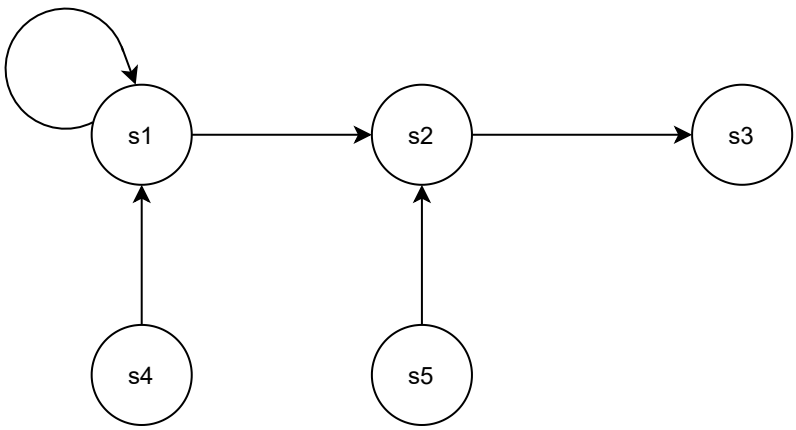
Reference

https://www.bilibili.com/video/BV1RA411q7wt?p=1

MDP动态特性

Markov Chain

马尔可夫链只有一个量——**状态**。比如 $S \in (s_1, s_2, s_3, s_4, s_5, s_6, s_7, s_8, s_9, s_{10})$，在状态集合中一共有十个状态，每个状态之间可以互相转化，即可以从一个状态转移到另外一个状态，当然，“另外的状态”也有可能是当前状态本身。如下图所示，s1状态到可以转移到s2状态，s1状态也可以转移到自己当前的状态，当然s1也有可能转移到s3，s4，s5，状态，下图中没有给出。



根据上面的例子，我们可以把所有的状态写成矩阵的形式，就成了**状态转移矩阵**。用状态转移矩阵来描述马尔可夫链的**动态特性**。以上面的状态集合$S \in (s_1, s_2, s_3, s_4, s_5, s_6, s_7, s_8, s_9, s_{10})$为例，可得一个 10×10 的矩阵。如下图所示：

$$\begin{bmatrix} s_1 s_1 & \dots & s_1 s_{10} \\ \vdots & \vdots & \vdots \\ s_{10} s_1 & \dots & s_{10} s_{10} \end{bmatrix}$$

由上面的例子可知，在状态转移的过程中，对于下一个状态转移是有概率的，比如说s1转移到到s1状态的概率可能是0.5，s1有0.3的概率转移到s2状态。马尔科夫过程是一个二元组（S， P），且满足： **S是有限状态集合， P是状态转移概率。** 可得：

$$P = \begin{bmatrix} P_{11} & \dots & P_{1n} \\ \vdots & \vdots & \vdots \\ P_{n1} & \dots & P_{nn} \end{bmatrix}$$

一个简单的例子：如图2.2所示为一个学生的7种状态{娱乐， 课程1， 课程2， 课程3， 考过， 睡觉， 论文}， 每种状态之间的转换概率如图所示。 则该生从课程 1 开始一天可能的状态序列为：

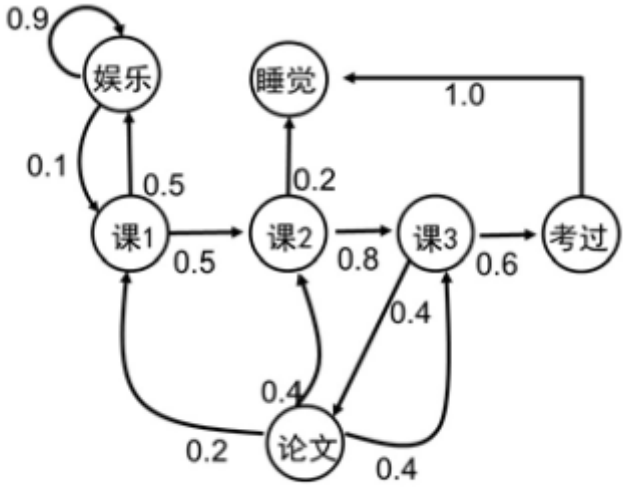


图2.2 马尔科夫过程示例图

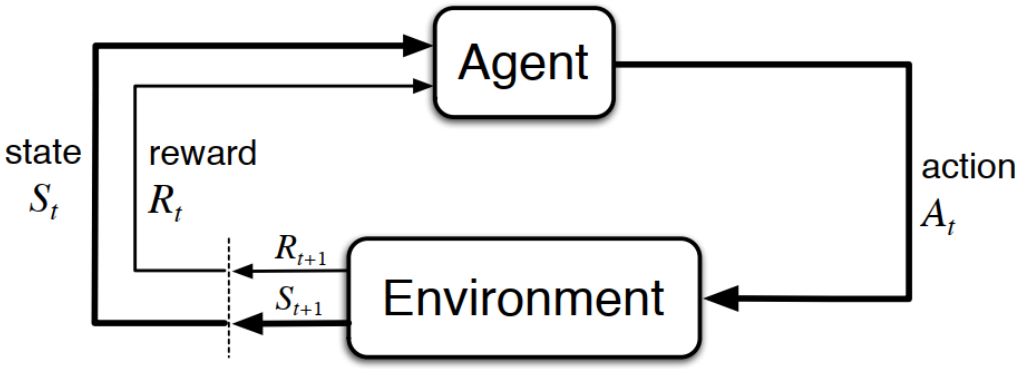
MRP

在 MPR 中，打个比喻，更像是随波逐流的小船，没有人为的干预，小船可以在大海中随波逐流。

MDP

在MDP中，打个比喻，更像是有人划的小船，这里相比较MRP中的小船来说，多了“人划船桨”的概念，可以认为控制船的走向。这里我们看下面的图：

s1状态到s2状态的过程，agent从s1发出action A1，使得s1状态转移到s2状态，并从s2状态得到一个R2的奖励。其实就是下图所示的一个过程。这是一个**动态的过程**，由此，引出**动态函数**。



- **动态函数：** The function p defines the dynamics of the MDP. 这是书上的原话，也就说，这个动态函数定义了MDP的 `动态特性` ，动态函数如下：

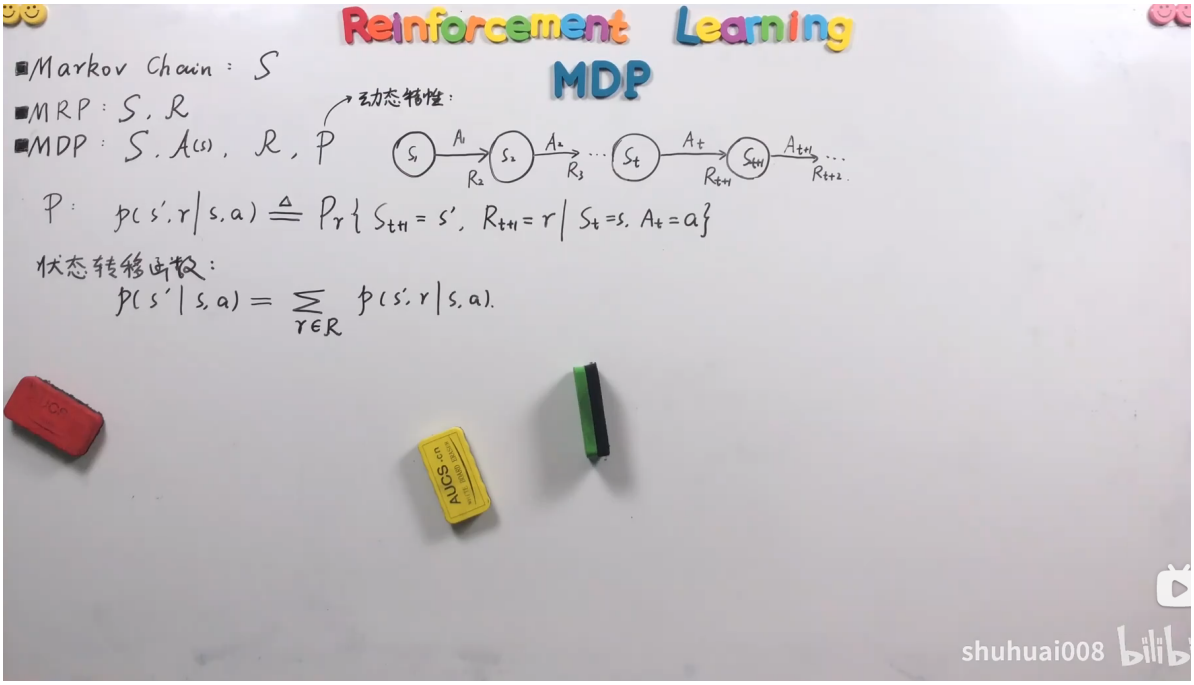
$$p(s', r \mid s, a) \doteq Pr\{S_{t+1} = s', R_{t+1} = r \mid S_t = s, A_t = a\}$$

- **状态转移函数：** 我们去掉 r ，也就是 `reward` ，动态函数也就成了状态转移函数。

$$p(s' \mid s, a) \doteq Pr\{S_{t+1} = s', \mid S_t = s, A_t = a\}$$
$$p(s' \mid s, a) = \sum_{r \in R} p(s' \mid s, a)$$

- `reward` **的动态性：** 在 s 和 a 选定后，r 也有可能是不同的，即 r 也是随机变量。但是，大多数情况在 s 和 a 选定后 r 是相同的，这里只做简单的介绍。

Summary



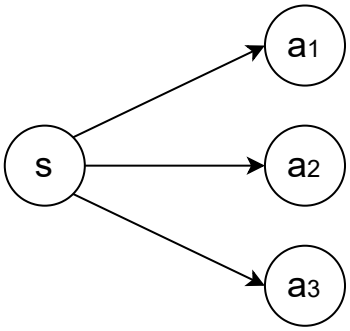
Reference

https://www.bilibili.com/video/BV1RA411q7wt?t=775&p=2

MDP价值函数

策略的定义

在MDP中，即马尔可夫决策过程，最重要的当然是**策略（Policy），用 π 来表示**。在策略中其主要作用的就是`action`，也即 A_t，需要指出的一定是，action 一定是基于某一状态 S 时。看下面的例子：



即，当 $S_t = S$ 状态时，无论 t 取何值，只要遇到 S 状态就 选定 a_1 这个 action ，这就是一种策略，并且是确定性策略。

策略的分类

- **确定性策略**：也就是说和时间 t 已经没有关系了，只和这个状态有关，只要遇到这个状态，就做出这样的选择。
- **随机性策略**：与确定性策略相对，当遇到 S 状态时，可能选择 a_1 ，可能选择 a_2，也可能选择 a_3。只是选择action的概率不同。如下图，就是**两种不同**的策略:

Policy1	
A	P(s)
a1	**0.8**
a2	**0.1**
a3	**0.1**

Policy2	
A	P(s)
a1	**0.5**
a2	**0.3**
a3	**0.2**

从上面两天图中，因为一个策略是基于一个状态而言的，在 S 状态，可能选择 a_1 ，可能选择 a_2，也可能选择 a_3，故三个 `action` 之间是**或**的关系，所以说以上是两个策略，而不要误以为是6个策略。

故策略可分为确定性策略和随机性策略两种。

$$Policy = \begin{cases} 确定性策略， & a \doteq \pi(s) \\ 随机性策略， & \pi(a \mid s) \doteq P\{A_t = a \mid S_t = s\} \end{cases}$$

对于随机性策略而言，给定一个 s ，选择**一个** a ，也就是条件概率了。

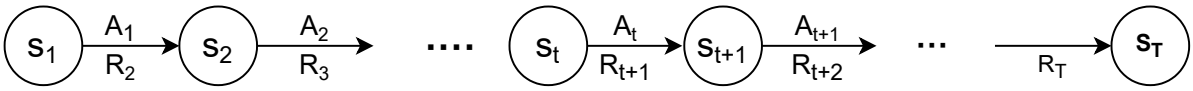
> 确定性策略可以看作是一种特殊的随机性策略，以上表-Policy1为例，选择a1的概率为1，选择a2，a3的概率都为0。

最优策略

在所有的策略中一定存在至少一个**最优策略**，而且在强化学习中，`reward`的获得有**延迟性（delay）**，举雅达利游戏中，很多游戏只有到结束的时候才会知道是否赢了或者输了，才会得到反馈，也就是`reward`，所以这就是奖励获得延迟。当选定一个状态 S_t 时，选定action A_t ，因为奖励延迟的原因可能对后续的 S_{t+1} 等状态都会产生影响。这样，就不能用当前的reward来衡量策略的好坏、优劣。这里引入了回报和价值函数的概念。

回报（G_t）

而是后续多个reward的加和，也就是**回报**，用 G_t 表示 t 时刻的回报。



如上图所示，此时的“回报”可以表示为：$G_t = R_{t+1} + R_{t+2} + \ldots + R_T$

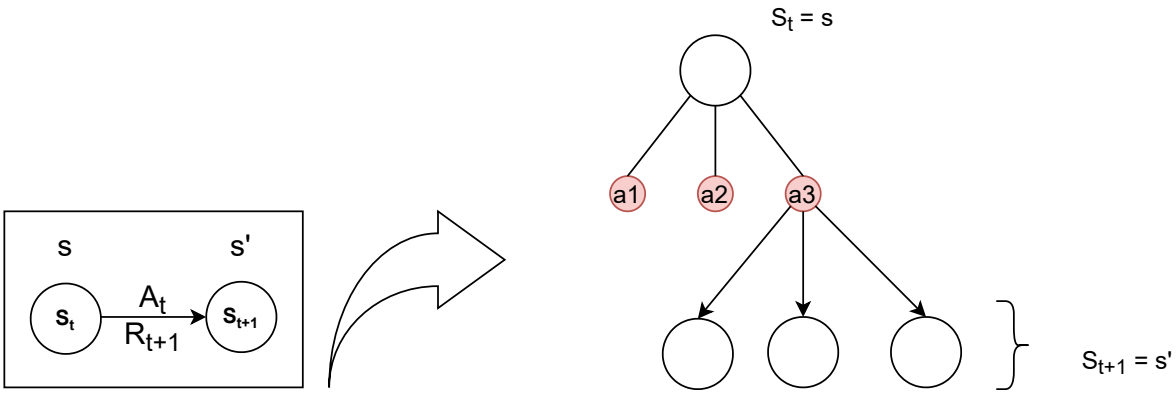
值得注意的是，T 可能是有限的，也有可能是无限的。

举例：张三对李四做了行为，对李四造成了伤害，李四在当天就能感受到伤害，而且，这个伤害明天，后头都还是有的，但是，时间是最好的良药，随着时间的推移，李四对于张三对自己造成的伤害感觉没有那么大了，会有一个wei折扣，用 γ 来表示。故**真正的回报**表示为：

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} \ldots + \gamma^{T-t-1} R_T = \sum_{i=0}^{\infty} \gamma^i R_{t+i+1} \qquad \gamma \in [0,1], \quad (T \to \infty)$$

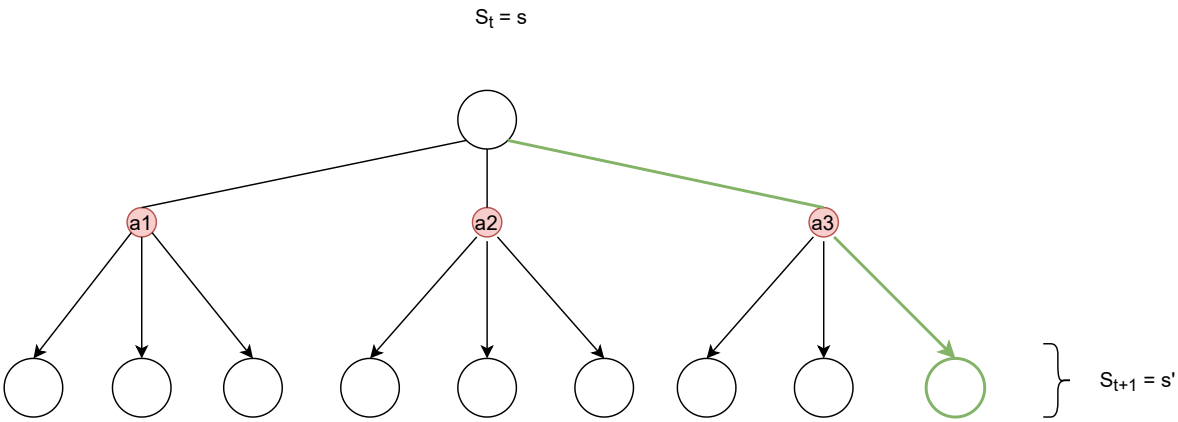
用 G_t 来衡量一个策略的好坏，G_t 大的策略就好，反之。

但是使用 G_t 也不能很好的衡量策略的好坏，比如一个最大的问题是，在给定一个状态后，选择一个确定的action后（这里还是在随机策略的角度），进入的下一个状态也是随机的。如下图所示：把左侧的过程放大，只给出a3下的随机状态，a1，a2也是有同样的情况，这里胜率。



举个例子，就像我们给一盆花浇水，水是多了还是少了，对于这盆花来说我们是不得知的，可能会少，也可能会多。这个花的状态变化也就是随机的了。从上面的例子得知，如果还是用 G_t 来对一个策略来进行评估的话，至少有9中情况（随机策略，3个action选一种）。

G_t 只能评估的是一个“分叉”而已（图中 绿色分支 ）。而不能更好的进行评估。如下图所示：



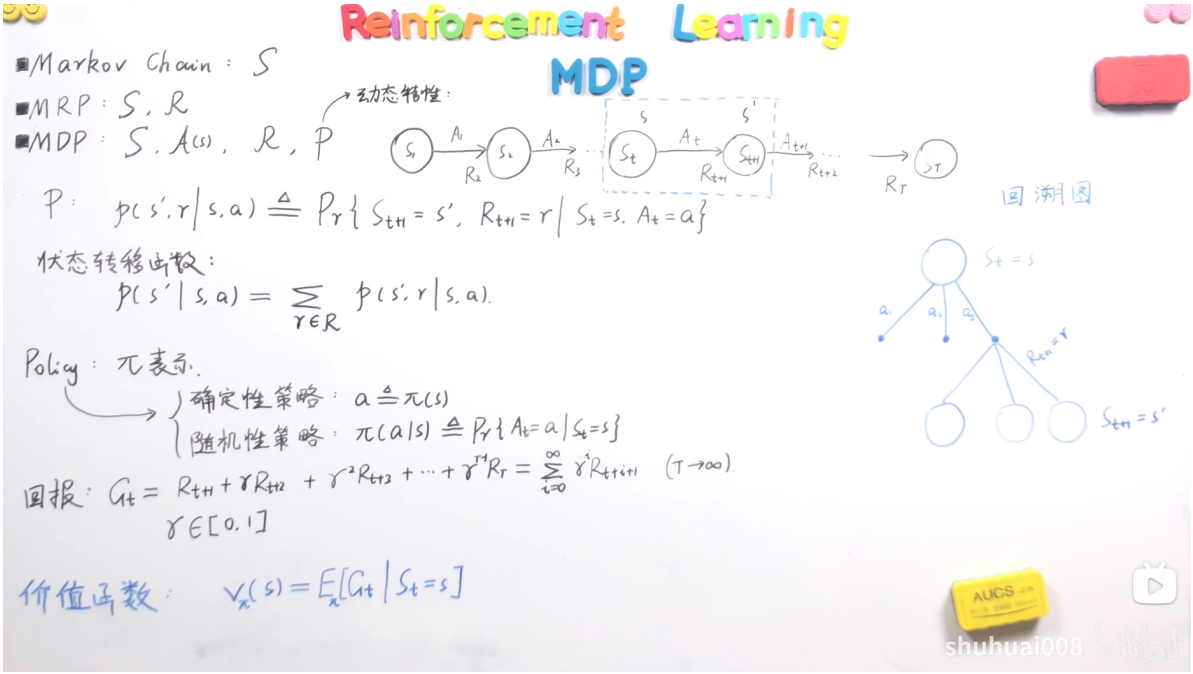
因为回报不能很好的对策略进行一个评估，由此引入了另外一个概念——价值函数。

价值函数（Value Function）

在指定一个状态 s，采取一个 随机策略 π，然后加权平均，以上图为例，把9 个分叉(G_t)加权平均。也就是**期望** E。故得出价值函数：

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

Summary



Reference

https://www.bilibili.com/video/BV1RA411q7wt?t=1200&p=3

MDP贝尔曼期望方程

价值函数分类

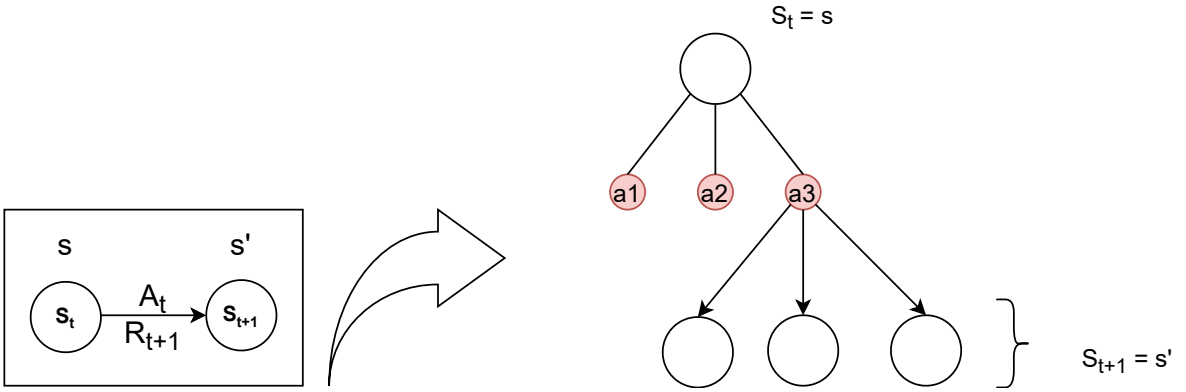
上面提到的的价值函数其实是其中的一种，确切的可以称为 状态价值函数，**用$v_\pi(s)$来表示**只和状态有关系。初次之外还有另外一种价值函数，那就是 状态行为价值函数，用$q_\pi(s,a)$这里引入的 `action`。故价值函数可以分为下面的两种:

$$Value \quad Function = \begin{cases} v_\pi(s) = E_\pi[G_t \mid S_t = s], & \text{only } s \text{ is independent variable} \\ q_\pi(s,a) = E_\pi[G_t \mid S_t = s, A_t = a], & \text{Both } s \text{ and a are independent variable} \end{cases}$$

从上面的公式中，我们可以得知，在 $v_\pi(s)$ 中，只有 s 是自变量，一个 π 其实就是一个状态s 和一个action的一个映射。故，只要π 确定了，那么s,a 也就确定了，即此时的 π 对状态 s 是有限制作用的。但是，在 $q_\pi(s,a)$ 中，子变量为s,a 两个，这两个自变量之间是没有特定的关系的。也就是说，s和a 都在变，无法确定一个映射(策略) π ,那么也就是说在 q_π 中的π 对于s 是没有约束的。

两种价值函数之间的关系

$v_\pi(s)$ 和 $q_\pi(s,a)$ 之间的关系



还是以上图为例，对于 s 状态，在随机策略中有三种 action 选择，分别是 $\pi(a_1 \mid s)$，$\pi(a_1 \mid s)$，$\pi(a_1 \mid s)$，三种action(行为)对应的价值函数（此时为行为价值函数）为 $q_\pi(s,a_1)$， $q_\pi(s,a_2)$， $q_\pi(s,a_3)$。那么此时的 $v_\pi(s)$ 就等于各个action的行为状态价值函数的加和，即：

$$v_\pi(s) = \pi(a_1 \mid s) \cdot q_\pi(s,a_1) + \pi(a_2 \mid s) \cdot q_\pi(s,a_2) + \pi(a_3 \mid s) \cdot q_\pi(s,a_3)$$

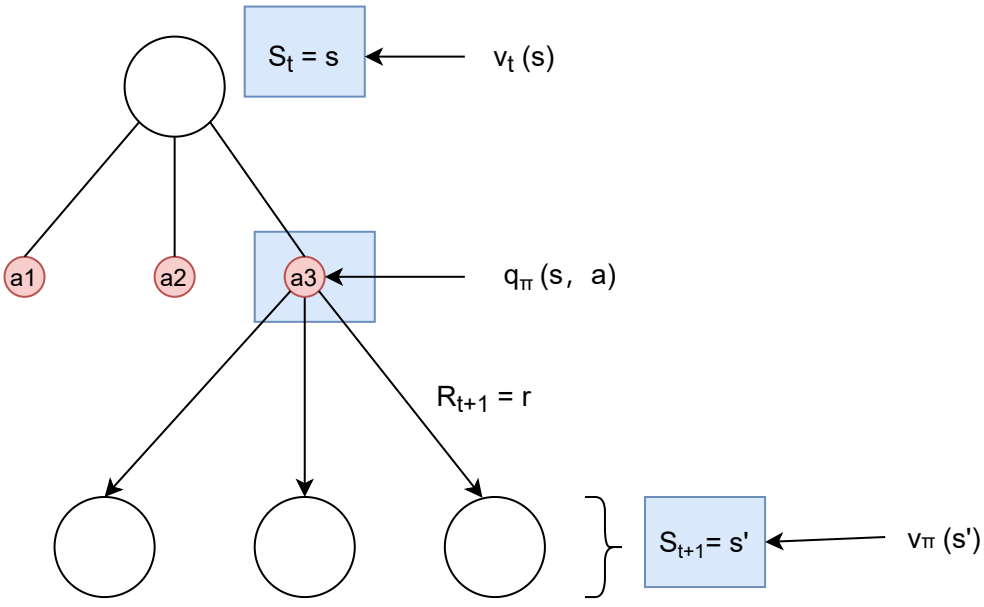
这样一来我们就得出了 $v_\pi(s)$ **和** $q_\pi(s,a)$ **之间的关系**，若条件已知，就可以直接计算出 v_π。

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

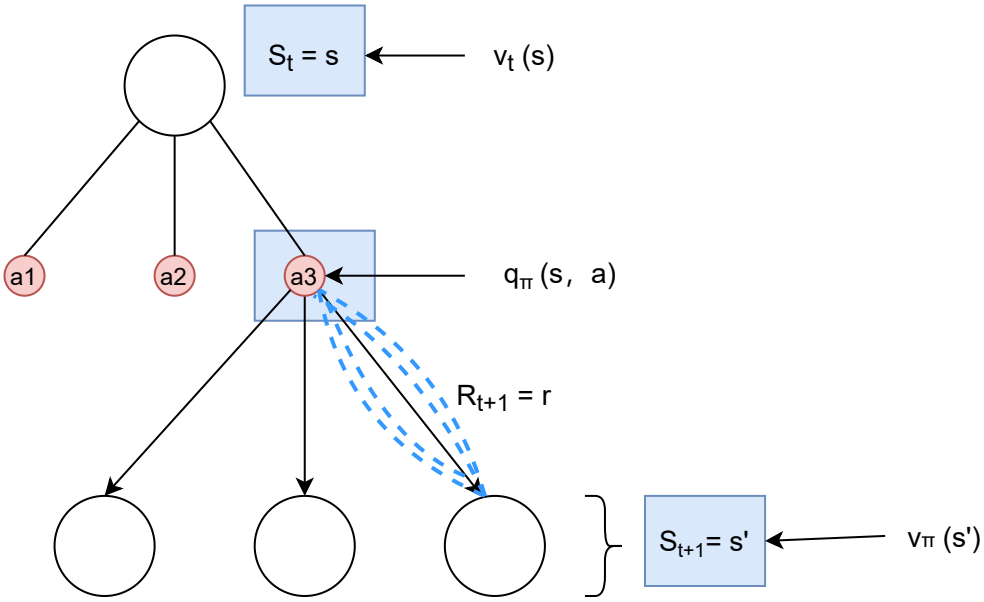
对于某个状态 s 来说，$v_\pi \le \max\limits_a q_\pi(s,a)$，$v_\pi(s)$ 是一个加权平均，实际上就是一个平均值，当然要小于等于 $q_\pi(s,a)$的最大值。$v_\pi(s)$只有全部是最大值的时候，两者才有可能相等。比如 5，5，5，平均值是 5，最大值也是5；3，4，5而言，平均值为4，但是最大值为5。注意的是，4是乘上权值后的值，换句话说也就是乘了一个概率（$\pi(a \mid s)$）。

$q_\pi(s,a)$ 和 $v_\pi(s')$ 之间的关系

从下面图中可得，在 $q_\pi(s,a)$ 位置，（一个action）状态转移只能向“箭头”方向转移，而不能向上。如果想从下面的状态转移到上面的状态，那必须还要另外一个action。情况是一样的，就按下图来说明，经过a3后到达状态s'，此时的状态函数就是 $v_\pi(s')$。



上面的图可知： 在确定了s 后，由随机策略action，引起“分叉”，同理，以a3为例，因为系统状态转移的随机性，也会引起分叉，也就是 s' 的状态也是不确定的。还有一点 r 也又不确定性，如下图蓝色虚线部分。



由我们前面提到的公式也可得知：s' 和 r 都是随机的。比如说s，a ，s' 都是给定的，r也是不确定的。

$$p(s',r \mid s,a) \doteq Pr\{S_{t+1}=s', R_{t+1}=r \mid S_t=s, A_t=a\}$$

这样一来，可得**一条蓝色通路的回报**：

$$q_\pi(s,a) = r + \gamma v_\pi(s') \qquad (1)$$

(1)式是怎么来的呢？以上图为例，在 $q_\pi(s,a)$ 处往下走，选定一个 r ，再往下到达一个状态s'， 此时在往下还是同样的状态，就是俄罗斯套娃，以此类推。关于其中的 $\gamma v_\pi(s')$ ，来自于 G_t。看下面的式子：

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} \ldots + \gamma^{T-t-1} R_T \qquad \gamma \in [0,1], \quad (T \to \infty)$$
$$G_t = R_{t+1} + \gamma(R_{t+2} + \gamma R_{t+3} + \gamma^2 R_{t+4} + \ldots)$$

因为 $v_\pi(s)$ 来自 G_t ，故类比得（1）式。

因为走每条蓝色的通路也都是由概率的，故我们需要乘上概率，同时累加求和，一个是多条蓝色通路另一个是多个s'。故得：$q_\pi(s,a)$ **和** $v_\pi(s')$ **之间的关系** 如下：

$$q_\pi(s,a) = \sum_{s',r} P(s',r \mid s,a)[r + \gamma v_\pi(s')] \qquad (2)$$

贝尔曼期望等式（方程）

这样我们得到两个式子：

$$v_\pi(s) = \sum_{a \in A} \pi(a \mid s) \cdot q_\pi(s,a) \qquad (3)$$
$$q_\pi(s,a) = \sum_{s',r} P(s',r \mid s,a)[r + \gamma v_\pi(s')] \qquad (4)$$

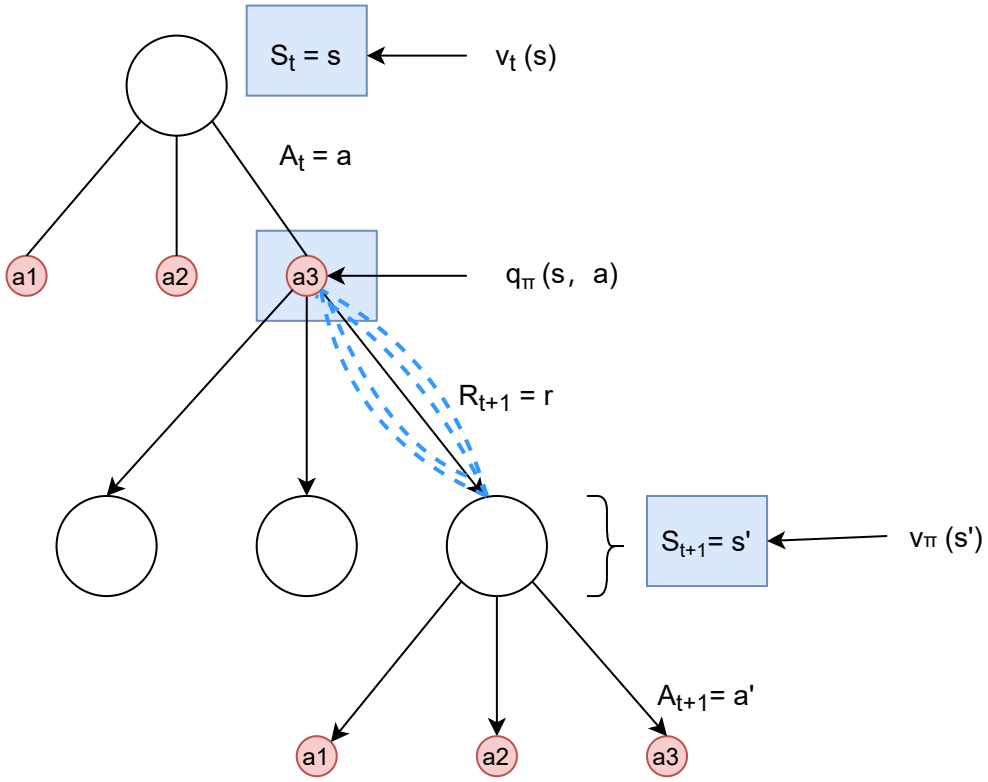
（4）式带入（3）得：

$$v_\pi(s) = \sum_{a \in A} \pi(a \mid s) \sum_{s',r} P(s',r \mid s,a)[r + \gamma v_\pi(s')] \qquad (5)$$

（3）式带入（4）得：

$$q_\pi(s,a)=\sum_{s',r}P(s',r\mid s,a)[r+\gamma\sum_{a'\in A}\pi(a'\mid s')\cdot q_\pi(s',a')] \qquad (6)$$

关于（6）式可以看下图，更容易理解：



（5）式和（6）式 被称为**贝尔曼期望方程。**

- 一个实例：

Summary