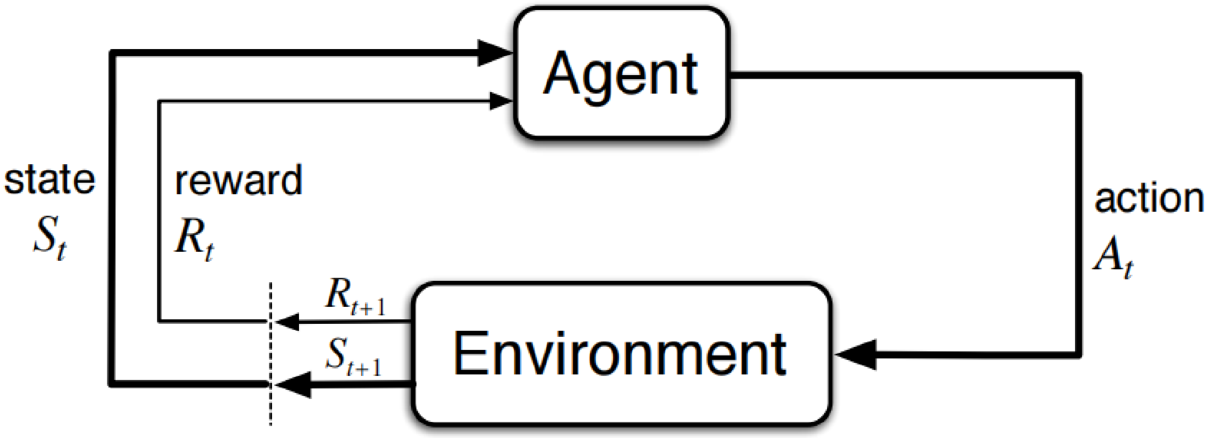
# 强化学习1-马尔可夫决策过程和贝尔曼方程

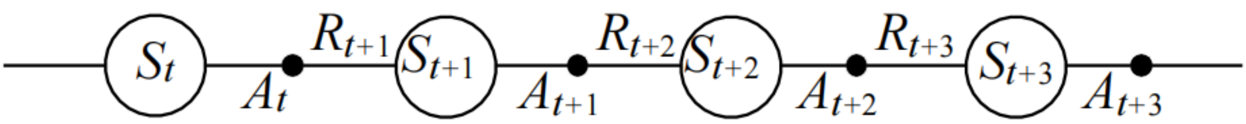
强化学习问题通常可以建模为一个MDP过程，在这个框架下，强化学习算法通过与环境的交互不断学习和改进策略。贝尔曼方程用于计算和更新状态的值函数，通过求解贝尔曼方程可以获取强化学习的最优策略。

## 马尔可夫决策过程MDP

马尔可夫决策过程MDP以这个模型为基础，其中最重要的是Action，State和Reward。




MDP可以由变量序列表示，也被称为轨迹（trajectory）。




$$\ldots S_t, A_t, R_{t+1}, S_{t+1}, A_{t+1}, R_{t+2}, S_{t+2}, \ldots$$

在MDP中，在一个状态a采取某个动作s会跳转到下一状态s'，并获得奖励r，其概率可以表示为 $p(s',r|s,a)$。如果 $p(s',r|s,a)$ 随环境变化，则称为non-stationary，反之则为stationary。给出一些常见的概率：

状态-行动期望奖励：

$$r(s,a) \doteq \mathbb{E}[R_t \mid S_{t-1}=s, A_{t-1}=a] = \sum_{r\in\mathbb{R}} r \sum_{s'\in\mathcal{S}} p(s',r|s,a),$$

状态-行动-下一状态期望奖励：

$$
\begin{aligned}
r(s,a,s') &\doteq \mathbb{E}[R_t \mid S_{t-1}=s, A_{t-1}=a, S_t=s'] \\
&= \sum_{r\in\mathbb{R}} r\cdot p(r|s,a,s'). \\
&= \sum_{r\in\mathbb{R}} r\frac{p(s',r|s,a)}{p(s'|s,a)}.(\text{贝叶斯公式})
\end{aligned}
$$

其中 $p(s'|s,a) = P(S_t = s'|S_{t-1} = s, A_{t-1} = a) = \sum_{r\in\mathbb{R}} p(s',r|s,a)$

## 基本概念

- 策略 $\pi$ ：在状态s采取动作a的概率，即：$\pi(a|s) = p(A_t = a|S_t = s)$ 。现实任务一般都是随机性策略，即在一个状态下按照概率采取不同动作，而非只能采取唯一动作的确定性策略。
- 奖励r：这里考虑的是应景获得的奖励，不考虑衰减的奖励 $G_t = \sum_{k=0}^{t} R_t$ ，考虑衰减的奖励 $G_t = \sum_{k=0}^{t} \gamma^k R_t$
- episode：

Episodic Task：有终止点的任务. Continuing Task: 没有终止点的任务。

对于Continuing Task, 目标函数是unbound的，所以添加一个discounting factor $\gamma$ , $\gamma$ 越小表示越急功近利。以下讨论的也都是continuing task。

$$
E(G_t) = \sum_{k=0}^{n} \gamma^k E(R_{t+k+1}) \le \sum_{k=0}^{n} \gamma^k E(R_{max}) \le \frac{1}{1-\gamma} E(R_{max})
$$

$$
E(G_t) = R_{t+1} + \gamma E(G_{t+1})
$$

## Bellman equation

期望回报(Expected Return)是在一个策略下给定所有可能轨迹的回报的期望值，强化学习的目的就是通过优化策略来使得期望回报最大化，其训练目标：$maxE(G_t) = E(R_{t+1} + \gamma R_{t+2} + ...)$

### State value:

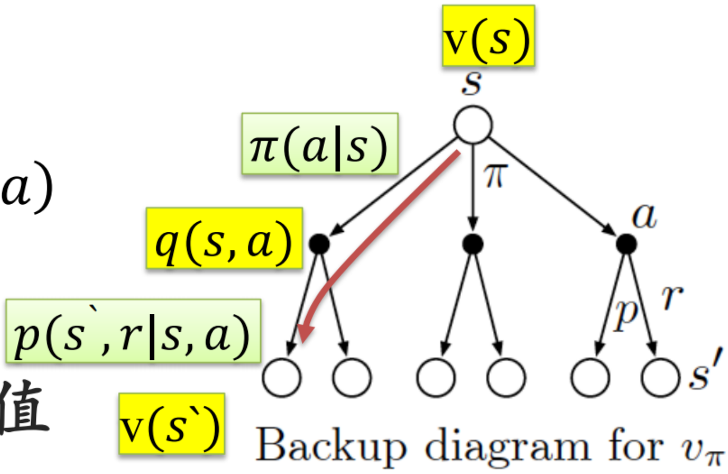
在一个状态下的value值，也就是一个状态的期望回报。

$$
\begin{aligned}
v_\pi(S) = E_\pi[G_t|S_t = s] = E(R_{t+1} + \gamma G_{t+1}|S_t = s) \\
= \sum_{a\in A} \pi(a|s) \cdot \sum_{s'} \sum_{r} p(s',r|s,a)[r + \gamma E(G_{t+1}|S_{t+1} = s')] \\
= \sum_{a\in A} \pi(a|s) \cdot \sum_{s'} \sum_{r} p(s',r|s,a)[r + \gamma v_\pi(s')]
\end{aligned}
$$

这里的 $s'$ 和 $r$ 就体现了状态和奖励的不确定性（采取相同动作后奖励不一定一致，到达的状态不一定一致）。

### ◻ 状态值 -- 动作值

$$v_\pi(s) = \sum_a \pi(a|s) q_\pi(s,a)$$




### ◻ 状态值 – 下一状态值

$$v_\pi(s) = \sum_a \pi(a|s) q_\pi(s,a)$$

$$= \sum_a \pi(a|s) \sum_{s`,r} p(s`,r|s,a)(r + \gamma v_\pi(s`))$$

**State-value function的矩阵形式：**

首先将State-value公式拆分成两项之和的形式：

$$v_\pi(s) = r_\pi(s) + \gamma \sum_{s'} p_\pi(s'|s) v_\pi(s')$$

$$r_\pi(s) = \sum_a \pi(a|s) \sum_r p(r|s,a) r, p_\pi(s'|s) = \sum_a \pi(a|s) p(s'|s,a)$$

假设状态为 $s_i(i = 1, \ldots, n)$ , Bellman方程为:

$$v_\pi(s_i) = r_\pi(s_i) + \gamma \sum_{s_j} p_\pi(s_j|s_i) v_\pi(s_j)$$

写成矩阵形式就是：

$$v_p i = r_\pi + \gamma P_\pi v_\pi$$

If there are four states, $v_\pi = r_\pi + \gamma P_\pi v_\pi$ can be written out as

$$\underbrace{\begin{bmatrix} v_\pi(s_1) \\ v_\pi(s_2) \\ v_\pi(s_3) \\ v_\pi(s_4) \end{bmatrix}}_{v_\pi} = \underbrace{\begin{bmatrix} r_\pi(s_1) \\ r_\pi(s_2) \\ r_\pi(s_3) \\ r_\pi(s_4) \end{bmatrix}}_{r_\pi} + \gamma \underbrace{\begin{bmatrix} p_\pi(s_1|s_1) & p_\pi(s_2|s_1) & p_\pi(s_3|s_1) & p_\pi(s_4|s_1) \\ p_\pi(s_1|s_2) & p_\pi(s_2|s_2) & p_\pi(s_3|s_2) & p_\pi(s_4|s_2) \\ p_\pi(s_1|s_3) & p_\pi(s_2|s_3) & p_\pi(s_3|s_3) & p_\pi(s_4|s_3) \\ p_\pi(s_1|s_4) & p_\pi(s_2|s_4) & p_\pi(s_3|s_4) & p_\pi(s_4|s_4) \end{bmatrix}}_{P_\pi} \underbrace{\begin{bmatrix} v_\pi(s_1) \\ v_\pi(s_2) \\ v_\pi(s_3) \\ v_\pi(s_4) \end{bmatrix}}_{v_\pi}.$$

由矩阵形式可知，每个状态s的state value都与其他状态有关，关系强弱由状态转移矩阵 $P_\pi$ 矩阵。给定一个策略，算出出相应的状态值被称为策略评估，这是强化学习的一个基本问题，可以通过求解此贝尔曼方程得到：

$$v_\pi = (I - \gamma P_\pi)^{-1} r_\pi$$

但是这种矩阵求逆复杂度为 $O(n^3)$，现实中求解贝尔曼方程一般采用policy iteration，即通过有限次的迭代得到一个近似解。

### State-action value:

在一个状态下采取一个动作的value，也就是基于此状态和动作的回报。

$$\begin{aligned} q_\pi(S, A) &= E_\pi[G_t | S_t = s, A_t = a] \\ &= \sum_{s'} \sum_r p(s', r | s, a)[r + \gamma v_\pi(s')] \\ &= \sum_{s'} \sum_r p(s', r | s, a)[r + \gamma \sum_{a' \in A} \pi(a'|s') q_\pi(s', a')] \end{aligned}$$

动作函数提供了一种衡量某种状态下动作好坏的标准，q值越大说明动作a越好。可以用到后面的贪心策略中。

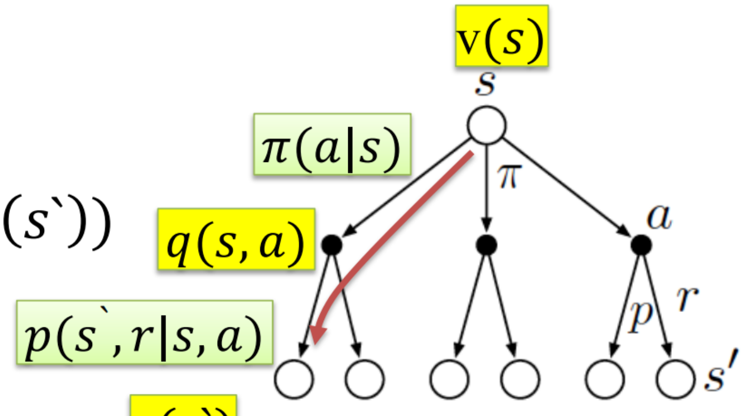
$$\pi(a|s) = \begin{cases} 1 & a = a^*, \\ 0 & a \neq a^*. \end{cases}$$

两者之间的关系：

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

## □ 动作值 － 状态值

$$q_\pi(s,a) = \sum_{s`,r} p(s`,r|s,a)(r+\gamma v_\pi(s`))$$



Backup diagram for $v_\pi$

## □ 动作值 － 下一动作值

$$q_\pi(s,a) = \sum_{s`,r} p(s`,r|s,a)(r+\gamma \sum_{a‘} \pi(a`|s`) q_\pi(s`,a`))$$