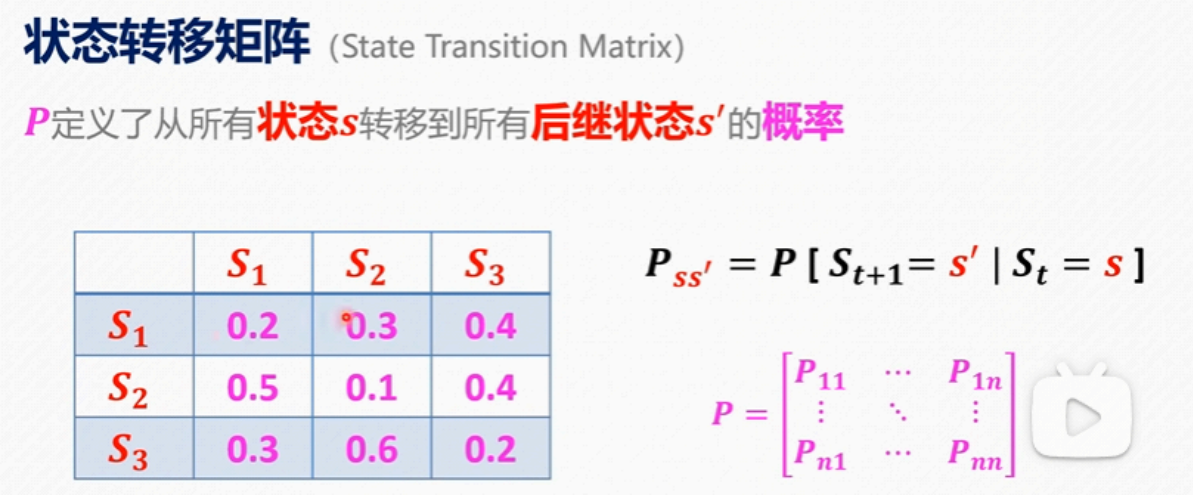
马尔可夫决策过程

强化学习的目标是给定一个马尔可夫决策过程，寻找最优策略

马尔可夫过程

马尔可夫过程是一个二元组<S,P>。S是有限状态集，P是状态转移矩阵。马尔可夫过程是一个无记忆的随机过程，每个状态序列S1、S2具有**马尔可夫性质**，即未来只与现在有关而与过去无关。

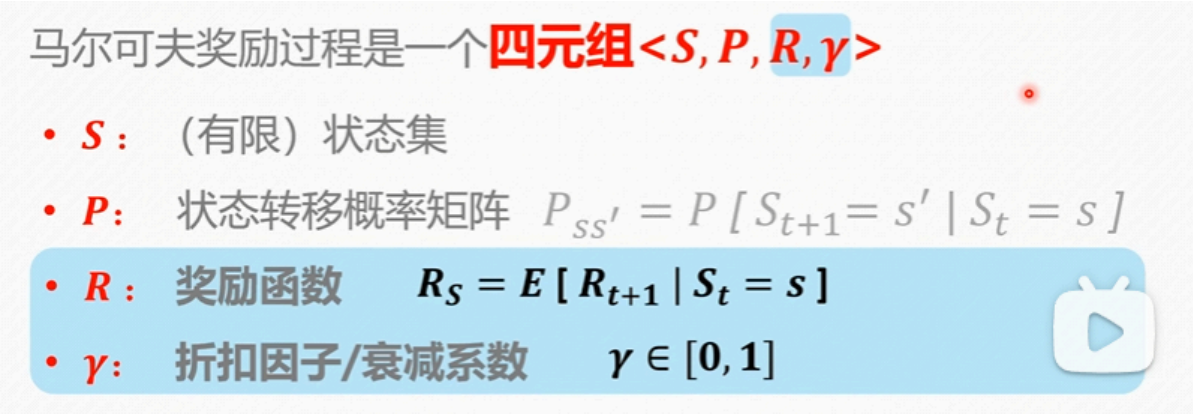
状态转移矩阵



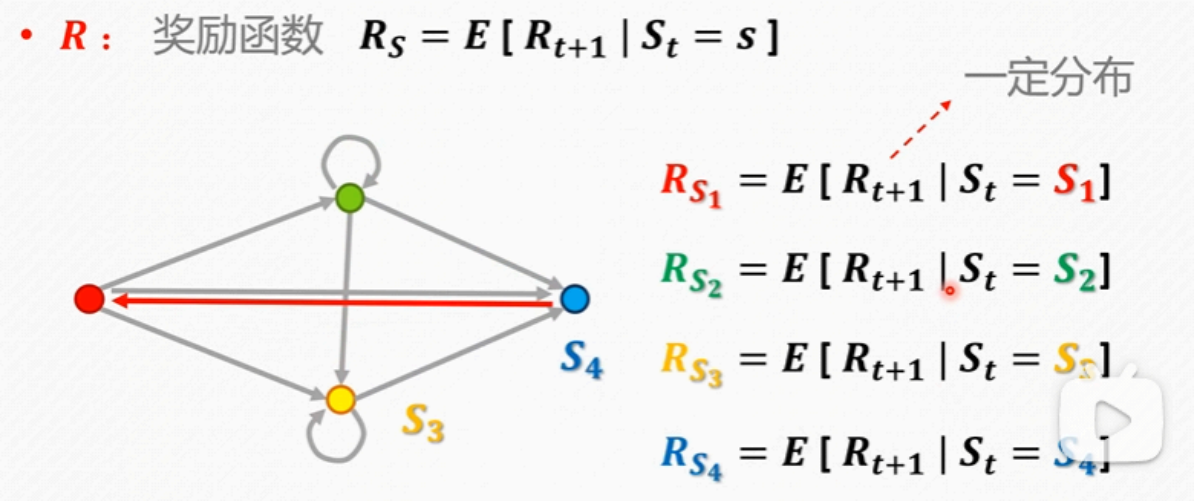
	S_1	S_2	S_3
S_1	0.2	0.3	0.4
S_2	0.5	0.1	0.4
S_3	0.3	0.6	0.2

马尔可夫奖励过程

由四元组<S,P,R,γ>组成

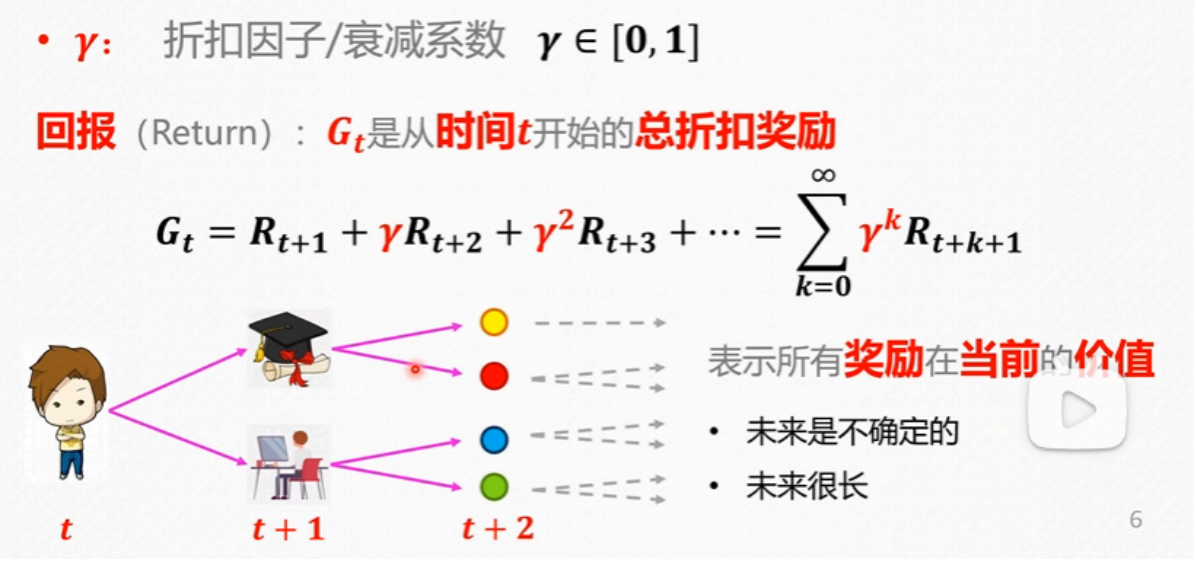


设当前状态为$S_t = s$,由s到S_{t+1}集合的奖励值集合为R_{t+1}，R_S就是集合R_{t+1}的期望。R_S是一个期望，但R_{t+1}不是，他只是代表到达t+1时刻的某一个状态点的奖励。



折扣因子γ

折扣因子γ是用来计算当前状态的总回报的，表示离的越远的时刻的价值越少。**注意，这里的G_t只是某一条路线的折扣奖励，并不是当前状态的最终价值。**



值函数V(S)

V(S)表示状态S的长期价值，即综合了未来所有路线的折扣奖励，算出期望。

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

贝尔曼方程

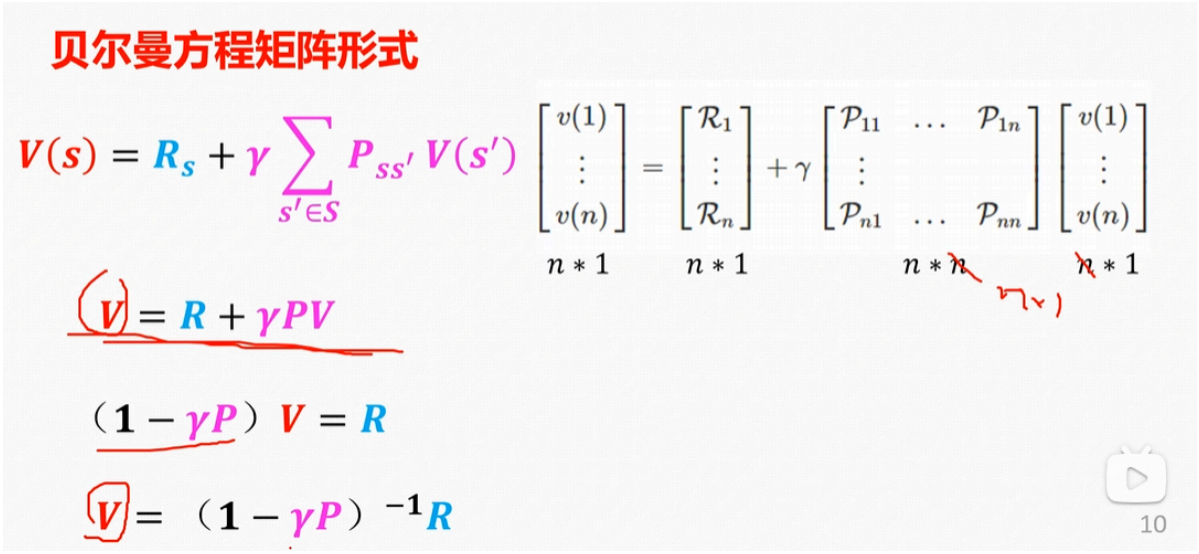
对于G_t可以采用递归的方式表示，即$G_t = R_{t+1} + \gamma G_{t+1}$。因为原式是在s状态的，而要求$E[G(S_{t+1})]$要在s'状态下，所以多了个转移矩阵。

$$V(S) = E[R_{t+1} + \gamma G(s_{t+1}) | S_t = s]$$
$$V(S) = R_s + \gamma \sum_{s' \in S} P_{ss'} V(s')$$

矩阵形式

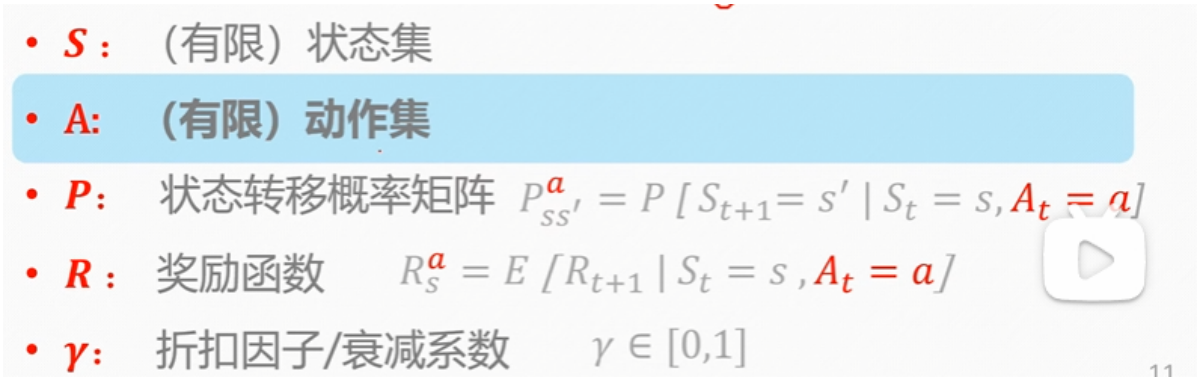
这里需要解释一下为什么两边看成一个V?

- 首先所有的状态本质上都是对自己进行迭代更新，因为状态集是不变的。在标量形式下分开写是为了表达递归关系。
- 贝尔曼方程想表达一个平衡状态下的值函数，即当V达到平衡时（收敛），左右两侧一样



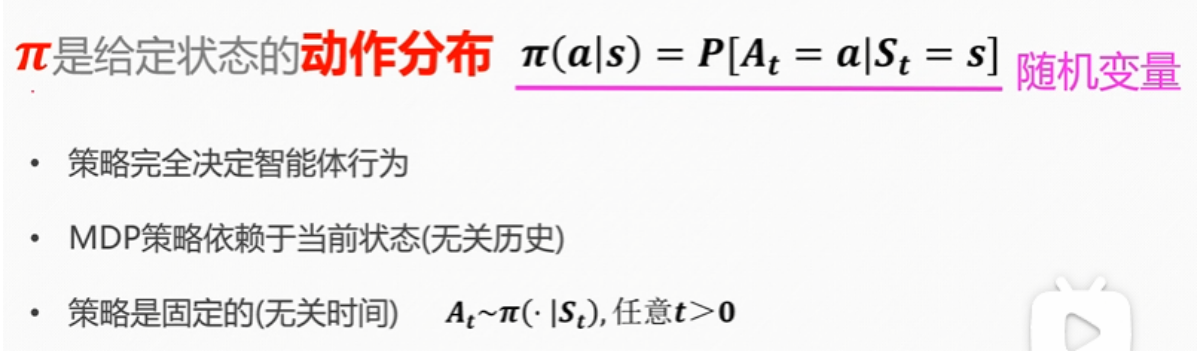
马尔可夫决策过程

由一个五元组<S,A,P,R,γ表示>，其中A是有限动作集。

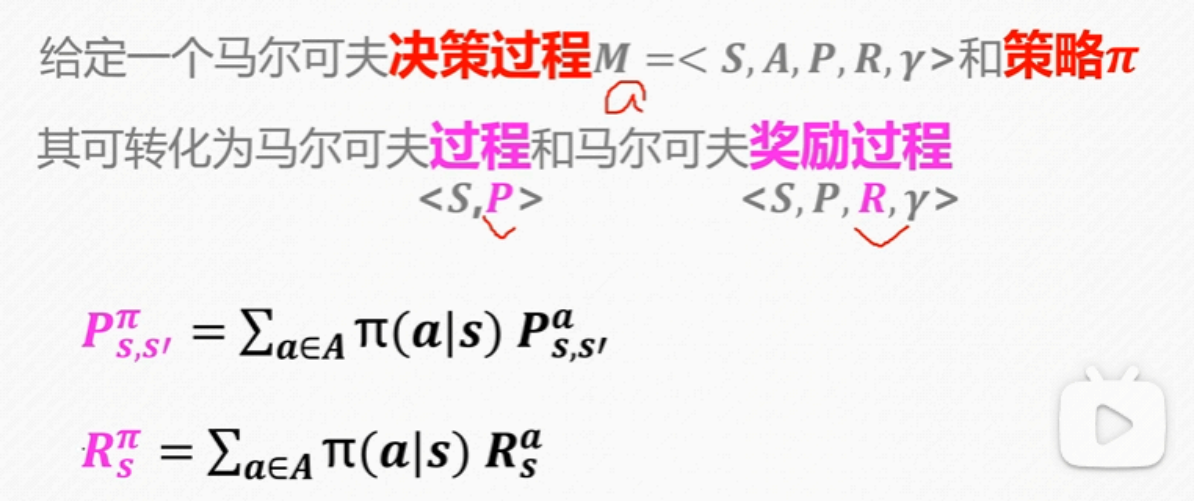


策略

π是一个随机变量，他的分布代表所有动作。

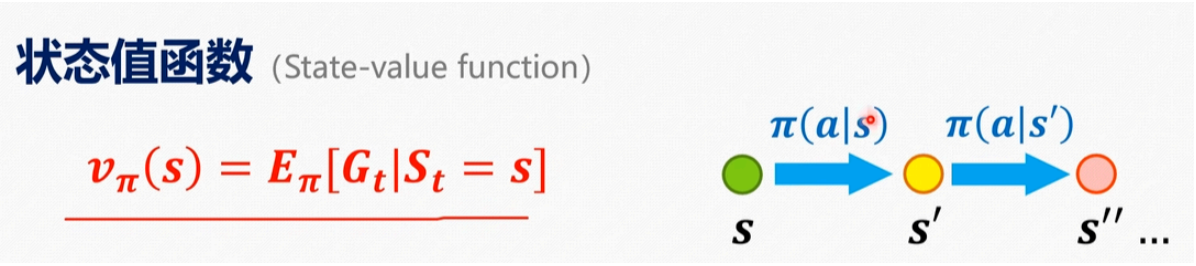


当前S下采用不同a的可能性 * 采用每个a后到不同S'的可能性 = 在策略π下从S到S'的可能性



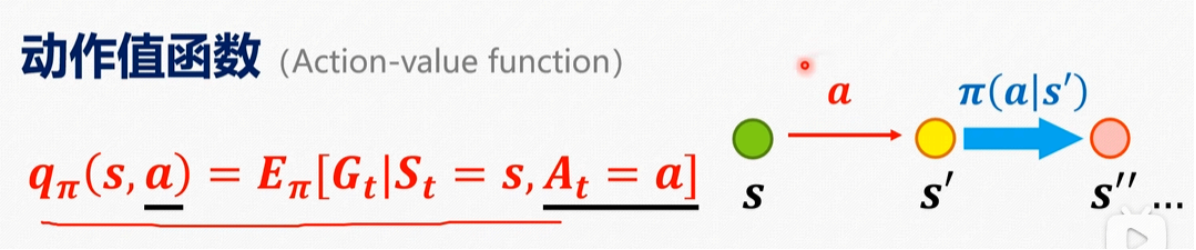
状态值函数

相比马尔可夫奖励过程的值函数，马尔可夫决策过程是智能体自身做出了对应动作后从而引起状态的改变，而前两者则是随波逐流。



动作值函数

在状态S下，先采取策略a后回报的期望。注意，在S下采取状态a并不能到达某个确定的状态，还是具有状态转移矩阵$P^{a}_{S,S'}$。其实就是在执行该动作后的状态值函数。

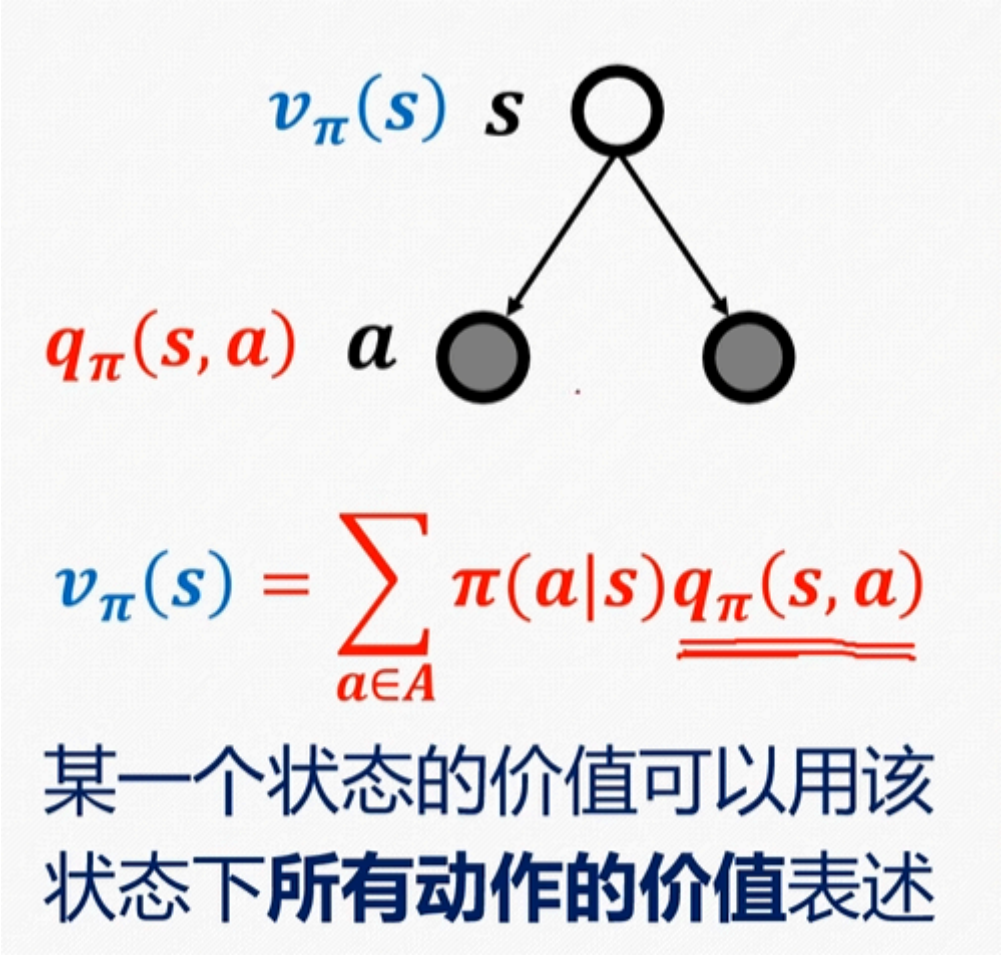


贝尔曼方程

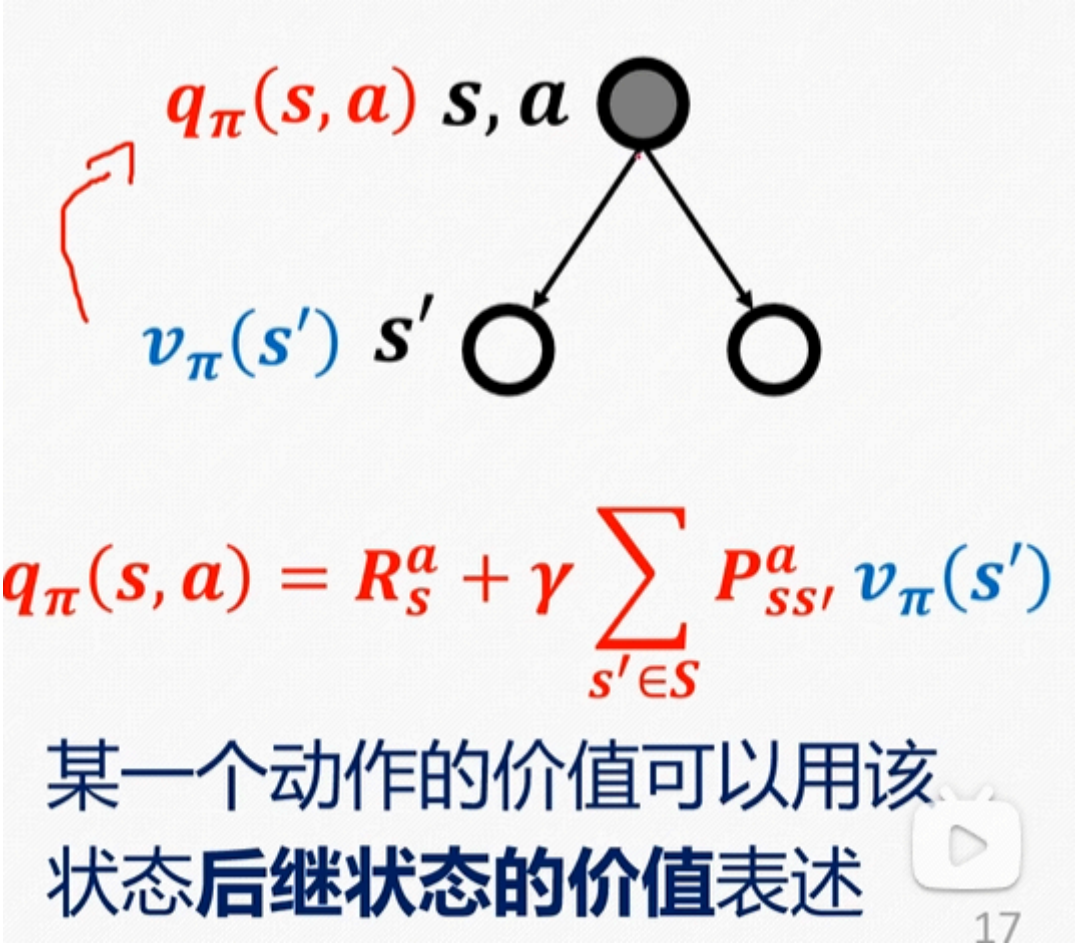
$$
\begin{aligned}
V(s) &= \sum_{a \in A} \pi(a|s)[R^{a}_{s} + \gamma \sum_{s' \in S} P^{a}_{ss'} V(s')] \\
&= \sum_{a \in A} \pi(a|s) q(s, a) \\
q(s, a) &= R^{a}_{s} + \gamma \sum_{s' \in S} P^{a}_{ss'} \sum_{a' \in A} \pi(a'|s) q(s', a') \\
&= R^{a}_{s} + \gamma \sum_{s' \in S} P^{a}_{ss'} v(s')
\end{aligned}
$$

状态值函数和动作值函数的关系

根据定义式可以推出：



根据贝尔曼的递归式子可以推出：



最优策略

最优状态值函数 (Optimal state-value function)

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

最优动作值函数 (Optimal action-value function)

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

最优策略 (Optimal policy)

$$\pi_* \geq any\ \pi$$

注：若$v_{\pi'}(s) \geq v_\pi(s)$，则$\pi' > \pi$

- 存在一个最优策略，使$\pi_* \geq any\ \pi$
- 所有最优策略都能取得最优**状态**值函数
- 所有最优策略都能取得最优**动作**值函数