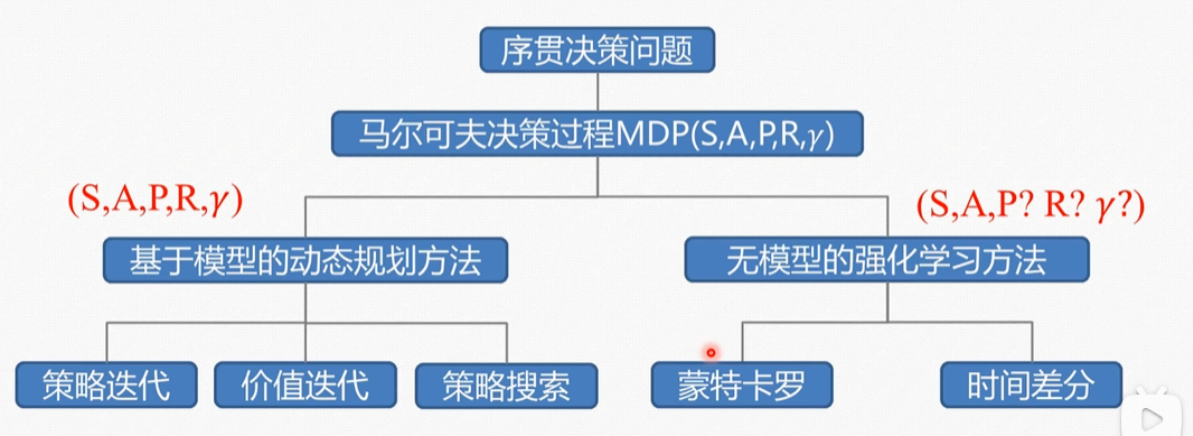
蒙特卡罗算法

就是通过多次采样得到每个动作的价值，强化学习本质上就是在学习这些。

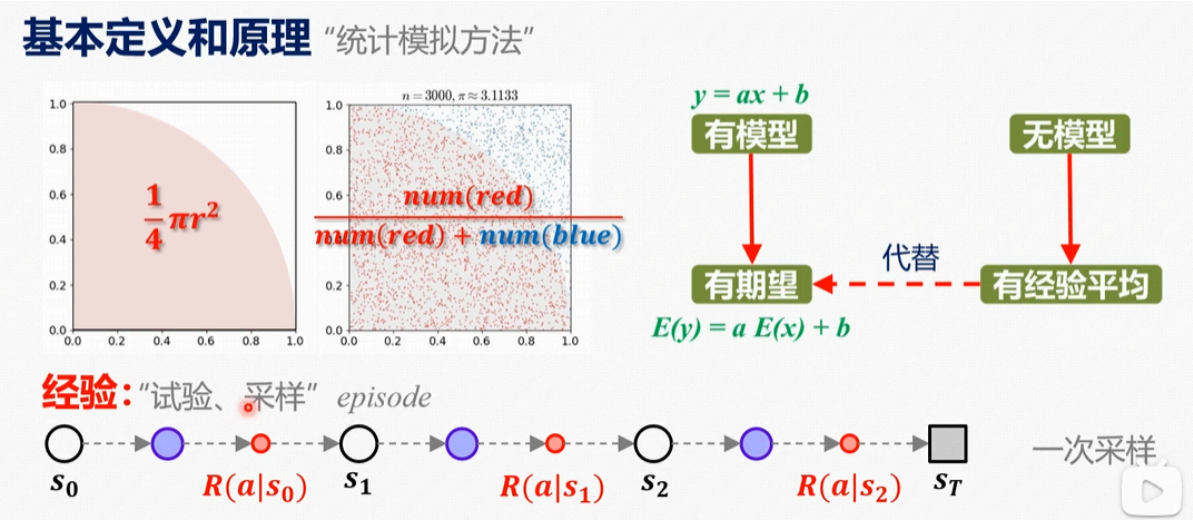


无模型蒙特卡罗强化学习的优势：

（1）能从环境中交互学习，在模拟试验中学习，无环境模型

（2）可以只聚焦于一个子状态空间，例如我们感兴趣的状态；DP方法理论上需要遍历所有状态空间

（3）不需要从其他值的模拟中迭代，不自举，如果马尔可夫属性不够，可以受到更小的影响



我们现在介绍如何用蒙特卡洛方法来估计一个策略在一个马尔可夫决策过程中的状态价值函数。回忆一下，一个状态的价值是它的期望回报，那么一个很直观的想法就是用策略在 MDP 上采样很多条序列，计算从这个状态出发的回报再求其期望就可以了，公式如下：

$$V^{\pi}(s) = \mathbb{E}_{\pi}[G_t|S_t = s] \approx \frac{1}{N}\sum_{i=1}^{N} G_t^{(i)}$$

在一条序列中，可能没有出现过这个状态，可能只出现过一次这个状态，也可能出现过很多次这个状态。我们介绍的蒙特卡洛价值估计方法会在该状态每一次出现时计算它的回报。还有一种选择是一条序列只计算一次回报，也就是这条序列第一次出现该状态时计算后面的累积奖励，而后面再次出现该状态时，该状态就被忽略了。假设我们现在用策略π从状态s开始采样序列，据此来计算状态价值。我们为每一个状态维护一个计数器和总回报，计算状态价值的具体过程如下所示。

(1) 使用策略π采样若干条序列：

$$s_0^{(i)} \xrightarrow{a_0^{(i)}} r_0^{(i)}, s_1^{(i)} \xrightarrow{a_1^{(i)}} r_1^{(i)}, s_2^{(i)} \xrightarrow{a_2^{(i)}} \cdots \xrightarrow{a_{T-1}^{(i)}} r_{T-1}^{(i)}, s_T^{(i)}$$

(2) 对每一条序列中的每一时间步t的状态s进行以下操作：

- 更新状态s的计数器 $N(s) \leftarrow N(s) + 1$；
- 更新状态s的总回报 $M(s) \leftarrow M(s) + G_t$；

(3) 每一个状态的价值被估计为回报的平均值 $V(s) = M(s)/N(s)$。

根据大数定律，当$N(s) \to \infty$，有$V(s) \to V^{\pi}(s)$。计算回报的期望时，除了可以把所有的回报加起来除以次数，还有一种增量更新的方法。对于每个状态s和对应回报G，进行如下计算：

- $N(s) \leftarrow N(s) + 1$
- $V(s) \leftarrow V(s) + \frac{1}{N(s)}(G - V(S))$