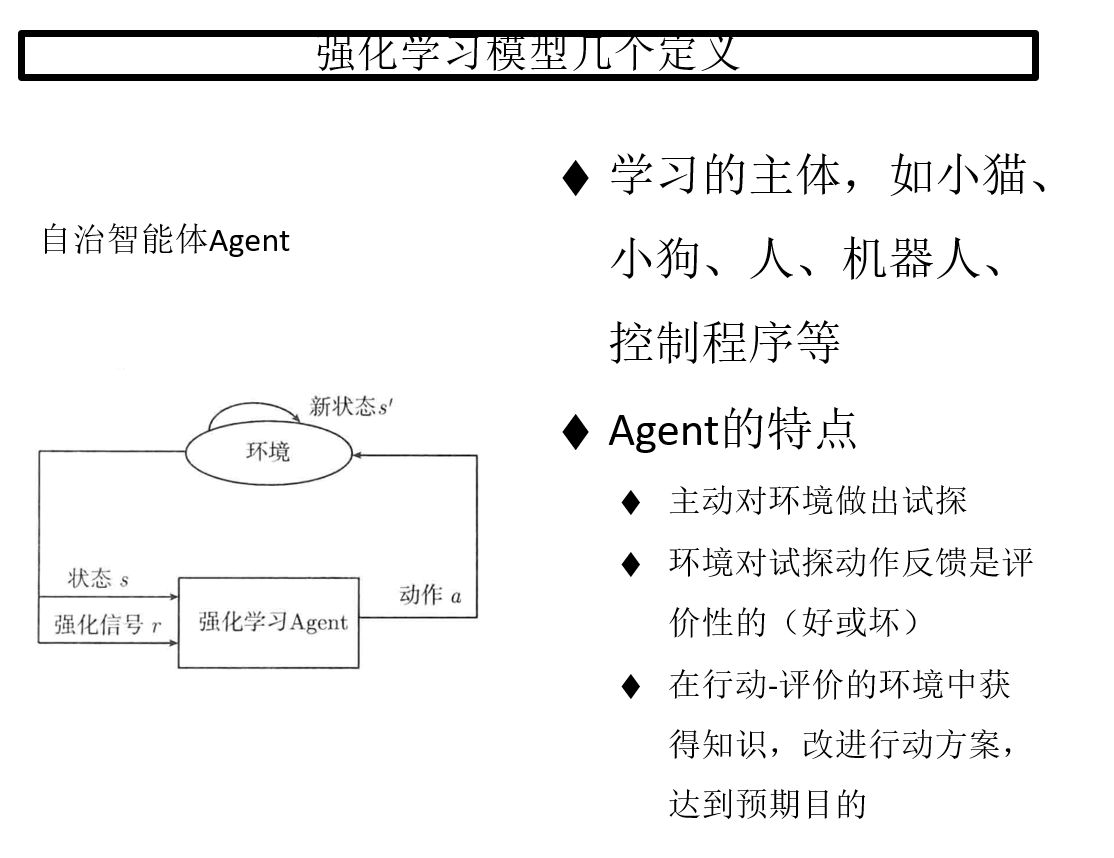
**前置知识**

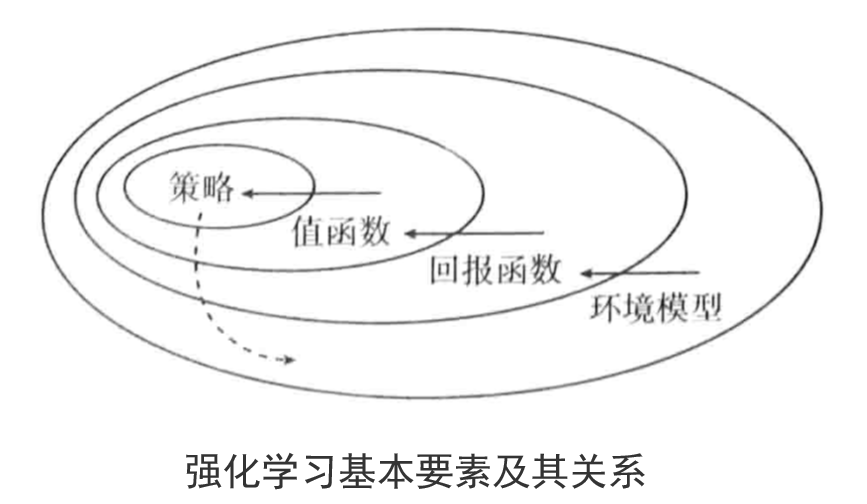
强化学习：只需提供奖励函数

监督学习：需要提供有标签的行动样本



奖励信号R是一个标量信号

Agent的任务是最大化累积奖励信号



**策略**

一个策略就是从环境感知的状态到在这些状态中可采取动作的一个映射，可能是一个查找表，也可能是一个函数

确定性策略：a=π(s)

随机策略：π(a|s) = P[At=a | St =s]

**奖励函数**

是强化学习问题中的目标，它把环境中**感知到的状态**映射为单独的一个奖赏

可以作为改变策略的标准

**效用函数**

从一个状态起agent所能积累的奖励的总和

在决策和评价决策中考虑最多的是效用函数

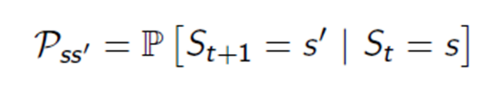
**环境模型**

环境模型模拟了环境的行为，即给定一个状态和动作，模型可以预测必定导致的下一个状态和下一个奖赏（**智能体心中的环境**）

模型一般用于规划，规划是算出来的。但强化学习是试出来的

**马尔可夫奖励过程**

马尔可夫奖励过程（MRP）是一种带有价值的马尔科夫链，由元组（S，P，R, γ ）来表示

* S为有限的状态集
* P为状态转移概率
* R为奖励函数 Rs=E[R(St+1) | St = s]

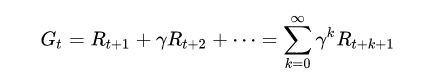
一个状态的奖励取决于下一状态奖励的期望值

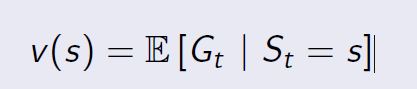
* γ 为折扣因子，γ ∈ [0,1]

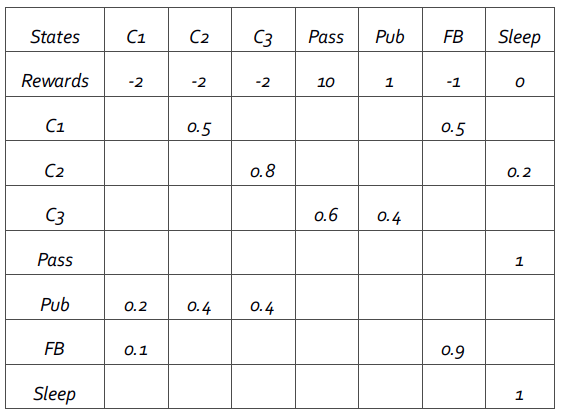
**强化学习的最优策略：**

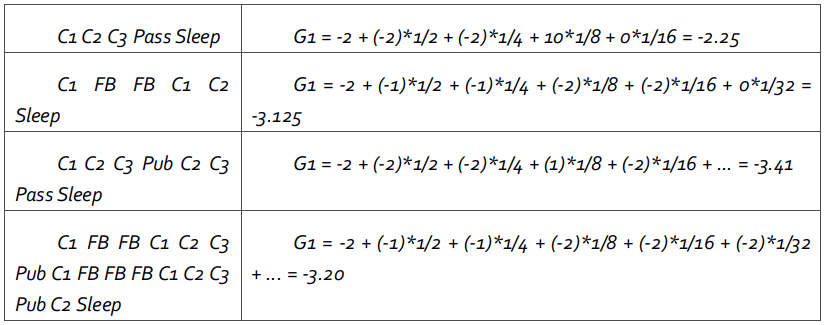
强化学习是找到最优的策略，这里的最优是指得到的总奖励最大

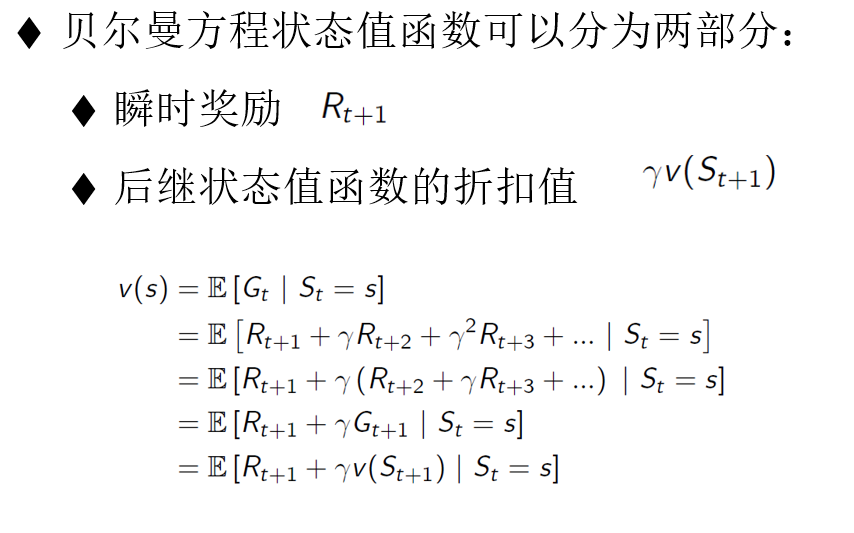
当给定一个策略时，我们就可以计算累积奖励：



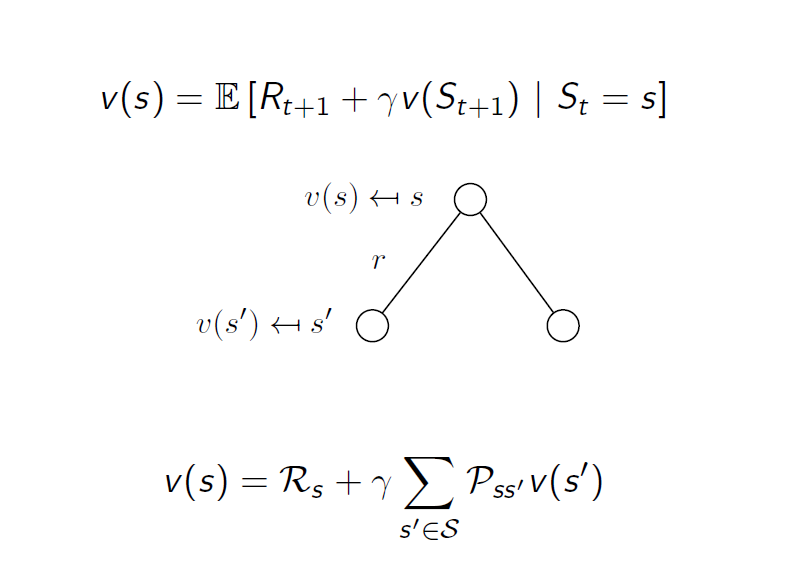
当智能体采用策略π时，累积奖励服从一个分布，累积奖励在状态s处的期望值定义为状态值函数：







相邻状态价值间关系：

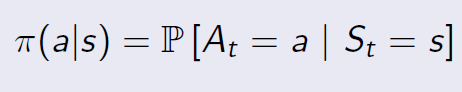


**马尔可夫决策过程**

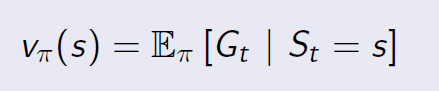
马尔可夫决策过程是一种带有决策作用的马尔科夫奖励过程，由元组（S，A，P，R, γ ）来表示

* S为有限的状态集
* A为有限的动作集
* P为状态转移概率
* R为奖励函数
* γ 为折扣因子，*γ* ∈ [0,1 ]

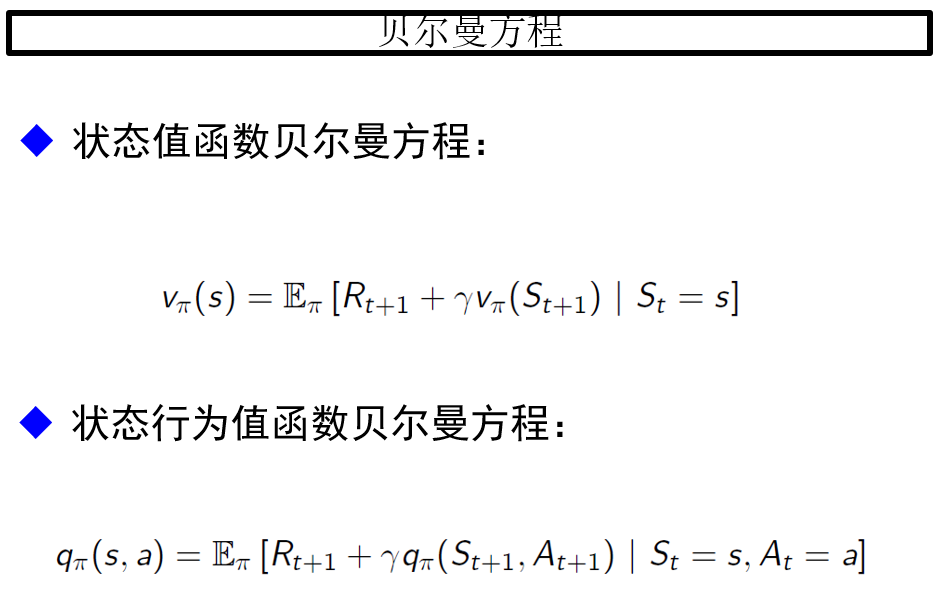
策略

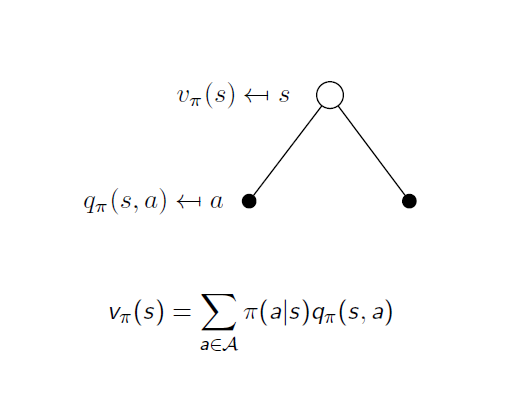
策略是指状态到动作的映射，策略常用符号π表示，它是指给定状态s时，动作集上的一个分布，即

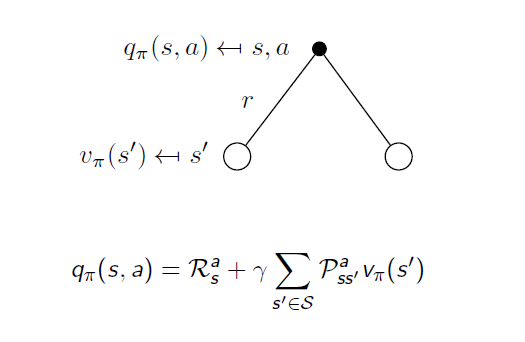
策略决定于当前状态，与历史状态无关；一般是固定的，不随时间变化

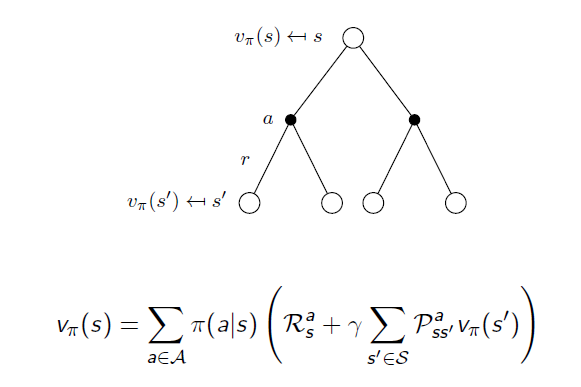
状态值函数：

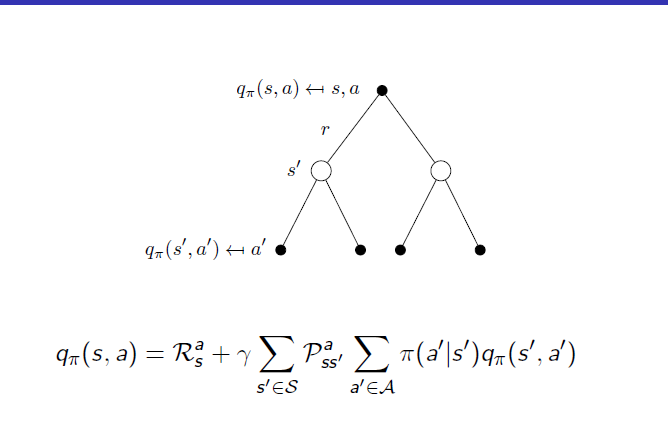
状态行为值函数：

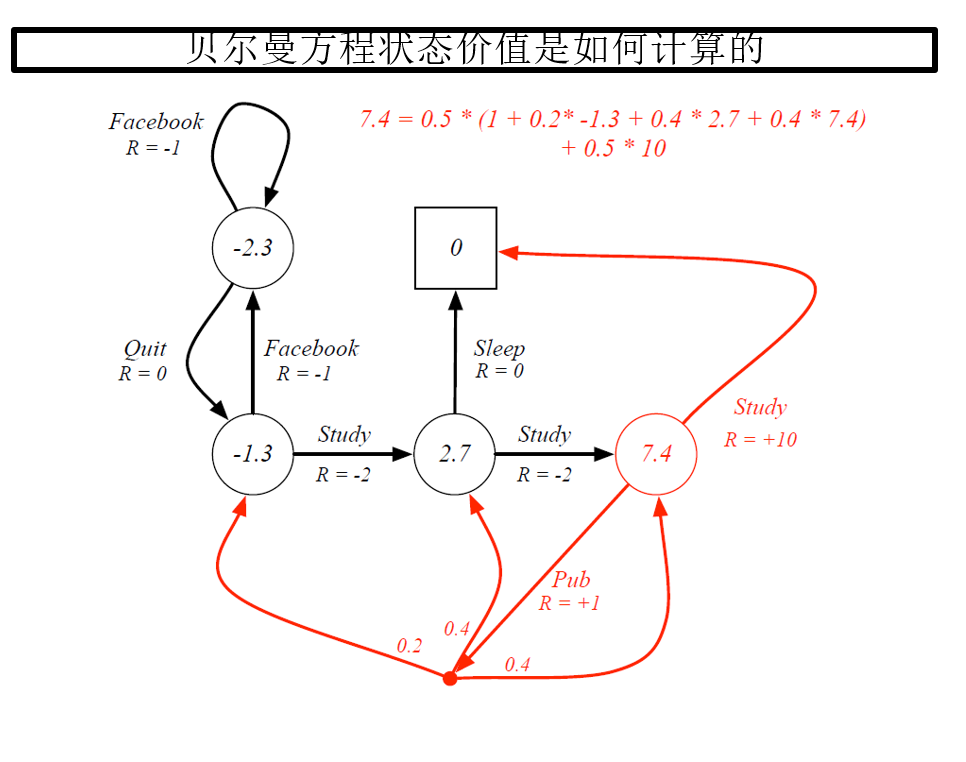
定义式

计算式1

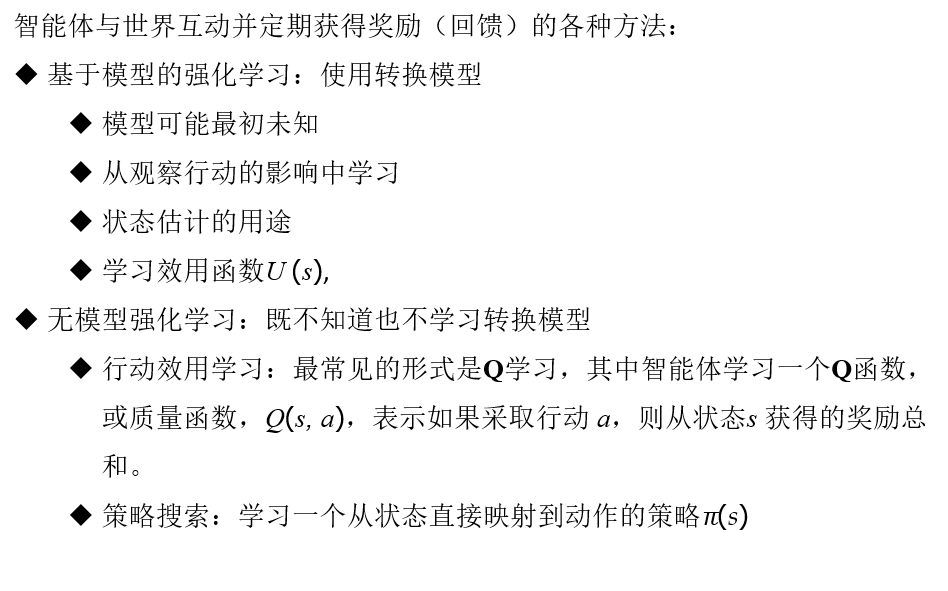
计算式1

计算式2

计算式2

**！！！！**

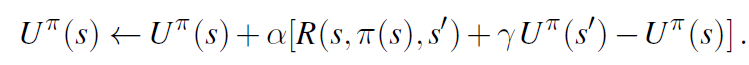
**22.1 从奖励中学习**



**时序差分学习**

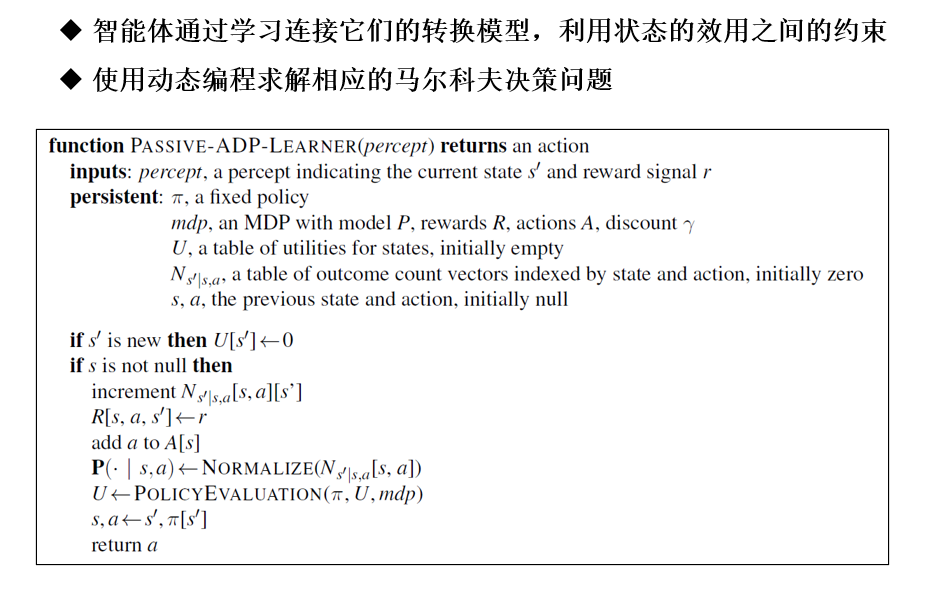
使用**观察到的转换**来调整**观察到的状态的效用值**，以便它们与约束方程相一致

时序差分（TD）方程



* + α是学习率参数
  + 使用连续状态（因此时间连续）之间的效用值差异
  + 调整效用值估计，当效用值估计正确时在本地保持的理想平衡
  + 不需要转换模型来执行更新

**自适应动态编程（ADP）**



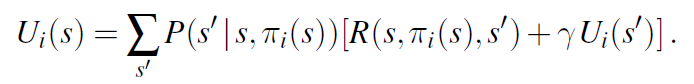
**22.2 被动强化学习**

**直接效用估计**

在每个序列的末尾，算法计算每个状态的观察到的奖励，并更新估计的效用值

将强化学习简化为到一个标准的监督学习问题，其中每个样本都是(state, reward-to-go)对

一个状态的效用取决于后继状态的奖励和预期效用



**22.3 主动强化学习**

区别：被动智能体有一个固定的策略来决定其行为，而主动学习智能体可以做决定

**探索**

* ADP固定策略：从不学习其他状态的效用，从不找到最佳路线
* 也被称为贪婪的智能体：贪婪地采取它认为的行动
* 有时这种方法会得到回报，有时它不会
* 忽略行动不仅仅是提供奖励
  + 行动以所产生状态中的感知形式提供信息
* 不在下一步行动中贪婪，但在无限探索的极限中贪婪**Greedy in Limit of infinite exploration (GLIE)**
  + GLIE 方案必须在每个状态下无限次尝试每个动作
  + 避免错过最佳操作的有限概率

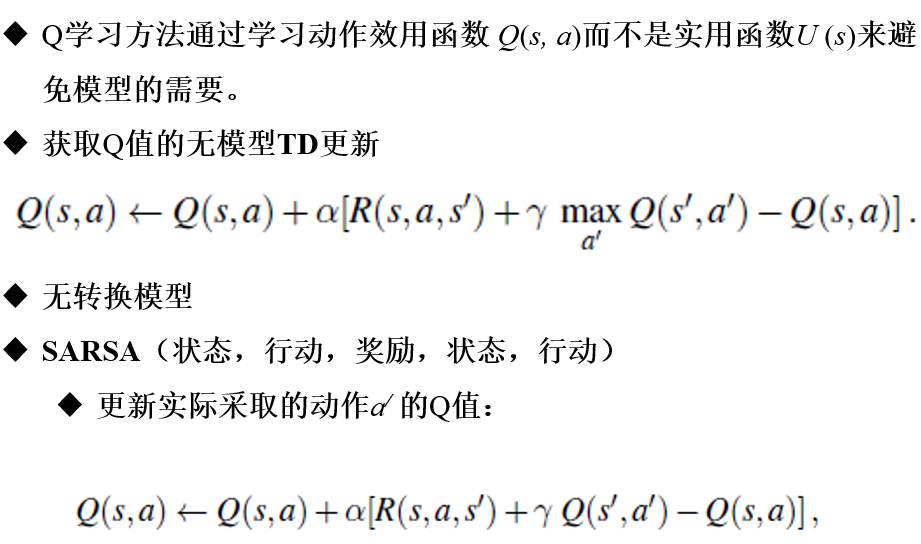
安全探索

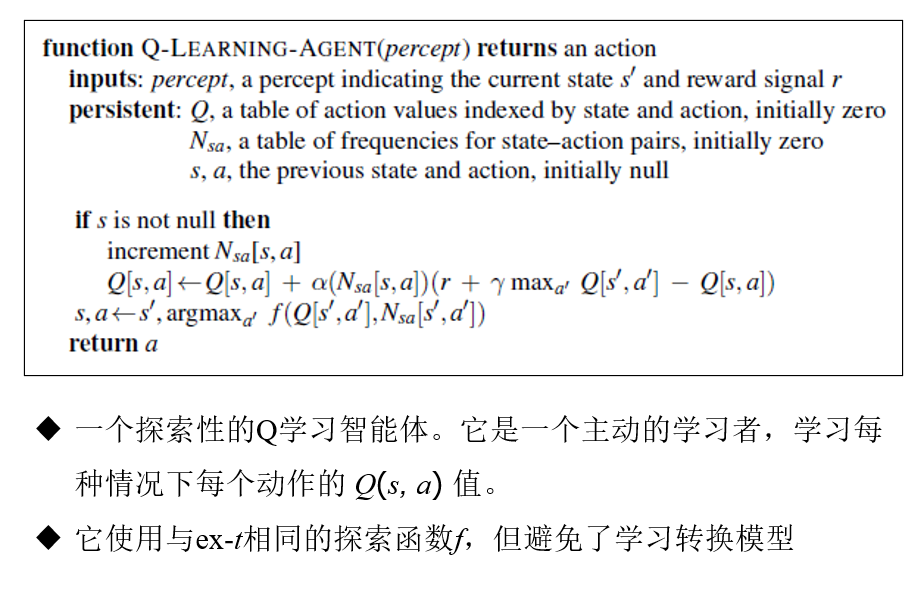
* 许多行为是不可逆转的。
  + 任何后续的行动序列都无法将状态恢复到不可逆转的行动之前的状态。
  + 最坏情况：智能体进入吸收状态（没有行动具有任何效果/奖励）

更好的行动方式：选择一个对具有合理机会成为真实模型的全体模型合理有效的策略

* + 即使该策略对于最大概率模型来说是次佳的。
  + 三种数学方法：
    - 贝叶斯强化学习
    - 探索 POMDP
    - 强大的控制理论

**时序差分Q学习（无模型学习）**

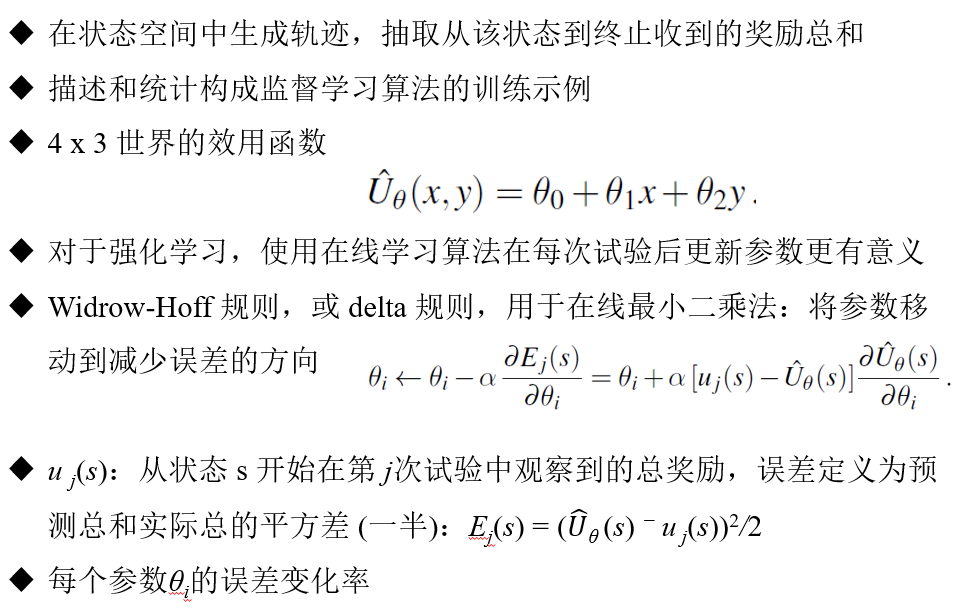




**22.4 强化学习中的泛化**

**强化学习的推广**

**近似直接效用估计**



* 灾难性遗忘
  + 学得太好：忘记了以前的训练
  + 平方特征获得零权重
* 重复实验
* 保留整个学习过程的轨迹和重播轨迹，以确保Q值函数是准确的

**深度强化学习（Deep RL）**

* 超越线性函数近似器
  + 没有接近效用函数或Q函数的良好线性函数
  + 无法发现必要的特征（新域）
* 需要更复杂的非线性函数近似器
* 然而，如果环境与训练数据甚至略有不同，那么深度学习很难获得良好的性能，则训练的系统的行为可能非常不可预测

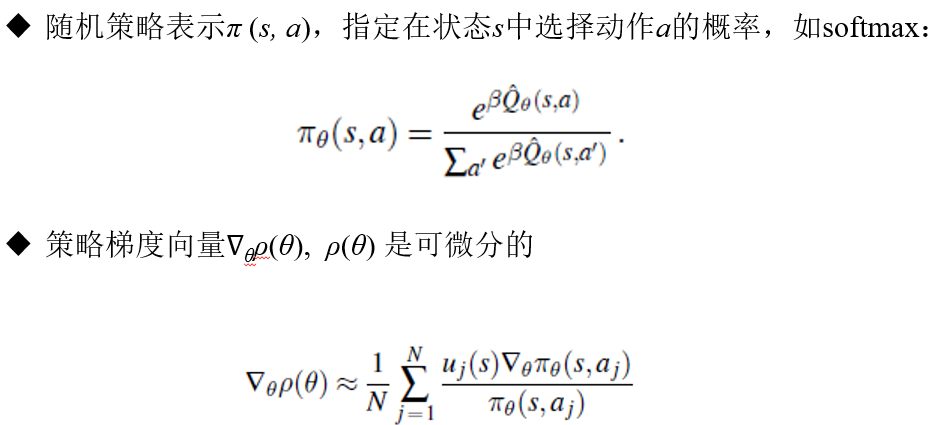
**层次强化学习？**

* 层次强化学习（HRL）与HTN(分层任务网）规划非常相似
* 从一部分程序开始，该程序概述了智能体行为的层次结构
* 为必须通过学习填写的,未指定的选择添加原始值
* 基于联合状态空间的概念，其中每个状态(*s, m*) 由物理状态s和机器状态*m*组成
* 选择状态*σ* =(*s, m*) 是指 *m*的程序计数器在智能体程序中的选择指针

**22.5 策略搜索**

只要继续改变策略，策略的表现有所改善就继续，否则停止

主要感兴趣的是 π 的参数化表示，其参数远远少于状态空间中的状态



**22.6 学徒学习与逆强化学习**

