4 - 免模型预测

有模型与免模型

有模型算法 —— 转移概率即环境是已知的
免模型算法 —— 对于智能体，环境是未知

有模型算法先学习一个环境模型（环境模型即拥有状态转移概率和奖励函数），这样的智能体可以直接计划最优行动策略。优点是不与真实环境交互成本低，但是模型往往不真实以至于难以学习

免模型算法在特定状态下执行特定动作的价值或优化策略，它直接从环境的交互中学习，不需要建立任何预测环境动态的模型。优点是简单直接，缺点则是需要大量的交互。

预测与控制

前面提到的很多经典强化学习算法嗾使免模型的，它们的状态转移概率是未知的，这种情况下会去近似环境的状态价值函数，这其实跟状态转移概率是等价的，我们把这个过程称为预测。

而控制的目标则是找到一个最优策略，该策略可以最大化期望的回报

蒙特卡洛估计

蒙特卡洛估计方法在强化学习中是免模型预测价值函数的方式之一，本质是一种统计模拟方法。

蒙特卡洛基于这样的想法：比如我们有一袋豆子，把豆子均匀地在一定范围内朝这个图形上撒，撒到足够多的数量时数一下这个图形中有多少颗豆子，这个豆子的数目就是图形的面积。当豆子越小撒的越多的时候，结果就越精确。此时我们借助计算机程序可以生成大量均匀分布坐标点，然后统计出图形内的点数，通过它们占总点数的比例和坐标点生成范围的面积就可以求出图形面积。

蒙特卡洛方法的思路是我们可以采样大量的轨迹，对于每个轨迹计算对应状态的回报然后取平均近似，称之为经验平均回报（ empirical mean return ）。蒙特卡洛方法有一定的局限性，即只适用于有终止状态的马尔可夫决策过程。

蒙特卡洛方法主要分成两种算法，一种是首次访问蒙特卡洛（，first-visit Monte Carlo，FVMC）方法，另外一种是每次访问蒙特卡洛（，every-visit Monte Carlo，EVMC）方法。
FVMC 方法主要包含两个步骤，首先是产生一个回合的完整轨迹，然后遍历轨迹计算每个状态的回报。而在 EVMC 方法中不会忽略统一状态的多个回报。

首次访问蒙特卡洛算法

初始化价值函数 $V(s)$，一个空的回报列表 $Returns(s_t)$
for 回合数 $= 1, M$ **do**
  根据策略 π 采样一回合轨迹 $\tau = \{s_0, a_0, r_1, \cdots, s_{T-1}, a_{T-1}, r_T, \}$
  初始化回报 $G \leftarrow 0$
  for 时步 $t = T-1, , T-2, \cdots, 0$ **do**
    $G \leftarrow \gamma G + R_{t+1}$
    repeat
      将 G 添加到 $Returns(s_t)$
      $V(S_t) \leftarrow \text{average}(\text{Returns}(S_t))$
    until s_t 第二次出现，即与历史某个状态 $s_0, \cdots, s_{t-1}$ 相同
  end for
end for

图 4-3 首次访问蒙特卡洛算法伪代码

蒙塔卡罗方法中的更新公式：$$\huge V(s_t)\leftarrow V(s_t)+\alpha[G_t-V(s_t)]$$ 其中 α 表示学习率，$G_t-V(s_t)$ 为目标值与估计值之间的误差。

此外，FVMC 是一种基于回合的增量式方法，具有无偏性和收敛快的优点，但是在状态空间较大的情况下，依然需要训练很多个回合才能达到稳定的结果。
而 EVMC 则是更为精确的预测方法，但是计算的成本相对也更高。

时序差分估计

时序差分估计方法是一种基于经验的动态规划方法，它结合了蒙特卡洛和动态规划的思想。最简单的时序差分可以表示为：

$$V(s_t) \leftarrow V(s_t) + \alpha[r_{t+1} + \gamma V(s_{t+1}) - V(s_t)]$$

这种算法一般称为单步时序差分。其中 $r_{t+1} + \gamma V(s_{t+1}) - V(s_t)$ 被定义为时序差分误差。

$$\begin{cases} V(s_t) \leftarrow V(s_t) + \alpha[r_{t+1} - V(s_t)] & \text{对于终止状态} \\ V(s_t) \leftarrow V(s_t) + \alpha[r_{t+1} + \gamma V(s_{t+1}) - V(s_t)] \end{cases}$$

时序差分和蒙特卡洛的比较

时序差分	蒙特卡洛
可以在线学习（即边走边学）	等游戏结束时才可以学习
可以从不完整序列上进行学习	只能从完整的序列上进行学习
可以在连续的环境下（没有终止）进行学习	只能在有终止的情况下学习。
利用了马尔可夫性质，在马尔可夫环境下有更高的学习效率	没有假设环境具有马尔可夫性质，利用采样的价值来估计某个状态的价值，在不是马尔可夫的环境下更加有效

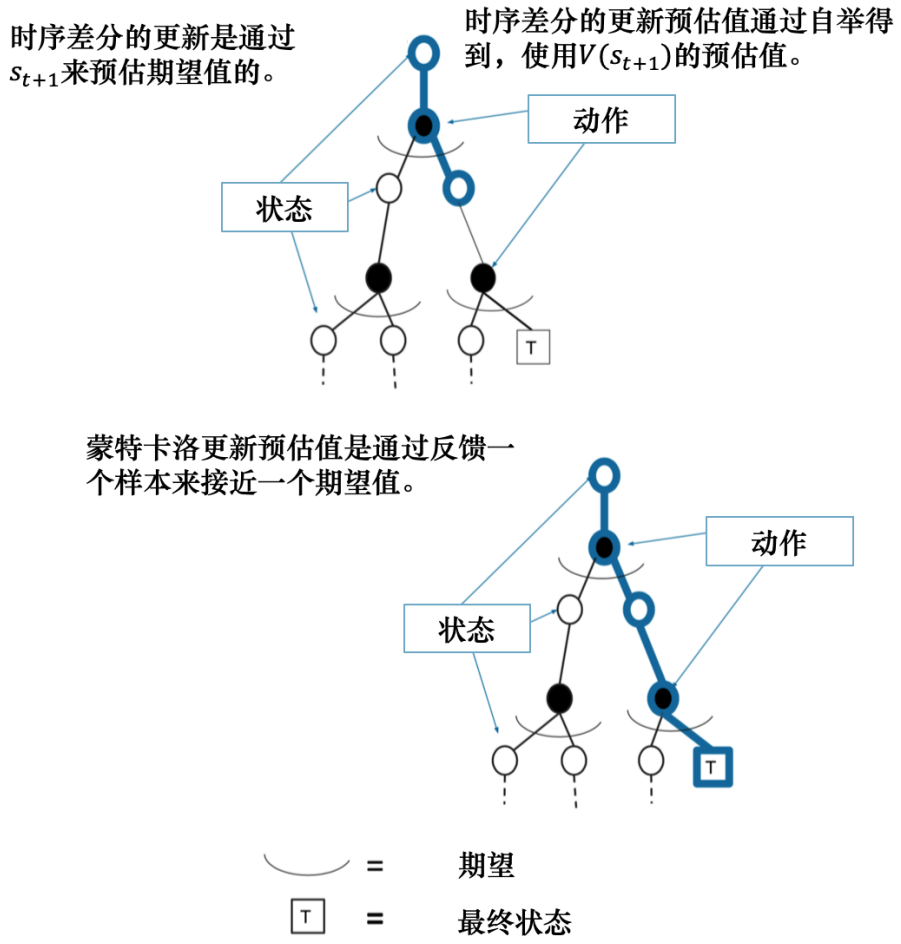

图 4.4 时序差分方法和蒙特卡洛方法的差异

n步时序差分

在使用时序差分方法是，当 n 趋近于无穷大时，就变成了蒙特卡洛方法。这个 n 我们常用 λ 来表示，即 $TD(\lambda)$ 方法。

以下是一些常见用于挑选合适 λ 的方法：

1. 网格搜索（Grid Search）：在一个给定的λ值内选择一个表现最好
2. 随机搜索（Random Search）在一个给定的λ值范围内随机选择一个表现最好
3. 自适应选择：训练中自适应改变 λ 值
4. 交叉验证：通过子集的表现得到一个 λ 的平均值
5. 经验取值：根据经验设置初始值再进一步进行微调

练习题

有模型与免模型算法的区别？举一些相关的算法？

有模型算法 —— 转移概率即环境是已知的（蒙塔卡罗方法，动态规划）
免模型算法 —— 对于智能体，环境是未知（Q-Learning）

举例说明预测与控制的区别与联系。

- 联系：同样是给出下一个状态
- 区别：预测会给出所有可能性，控制时会给出最优解

蒙特卡洛方法和时序差分方法的优劣势。

时序差分	蒙特卡洛
可以在线学习（即边走边学）	等游戏结束时才可以学习
可以从不完整序列上进行学习	只能从完整的序列上进行学习
可以在连续的环境下（没有终止）进行学习	只能在有终止的情况下学习。
利用了马尔可夫性质，在马尔可夫环境下有更高的学习效率	没有假设环境具有马尔可夫性质，利用采样的价值来估计某个状态的价值，在不是马尔可夫的环境下更加有效