【强化学习】蒙特卡洛方法

欧阳露露

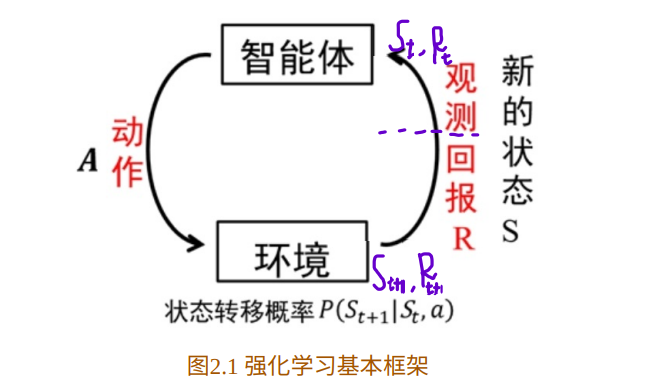
2023-2-27

# 回顾

强化学习算法解决的是序贯决策问题，而一般的序贯决策问题可以利用马尔科夫决策过程的框架来表述，即MDP

什么是序贯决策问题呢？

就是需要连续不断地做出决策，才能实现最终目标的问题。

两个主体：Agent智能体，Environment环境  
一个框架：MDP  


五大元素：  
 为有限的状态集  
 为有限的动作集  
 为状态转移概率  
 为回报函数  
为折扣因子，用来计算累积回报

动作转移概率

策略

累计回报

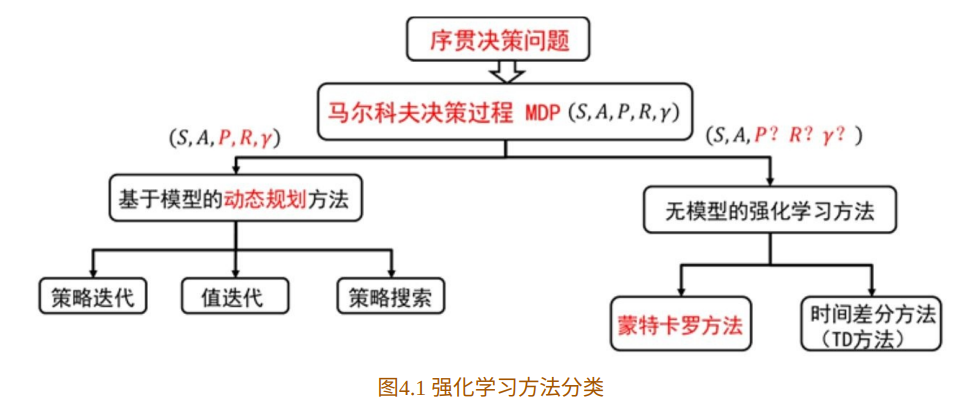
强化学习的一个核心目标——找到一个最优策略使得当前回报最大  
引入价值函数来衡量优劣，价值函数有状态值函数、状态行为值函数  
利用策略迭代找最优策略，策略迭代——先策略评估，再策略改进

状态值函数

状态-行为值函数

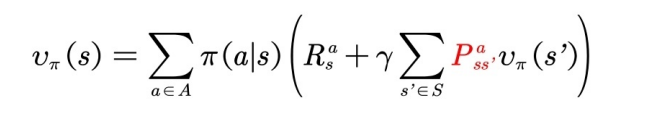
贝尔曼方程

# 基于蒙特卡罗的强化学习方法

  如图4.1所示，无模型的强化学习算法主要包括蒙特卡罗方法和时间差分方法。 

  无模型的强化学习基本思想：策略评估和策略改善。

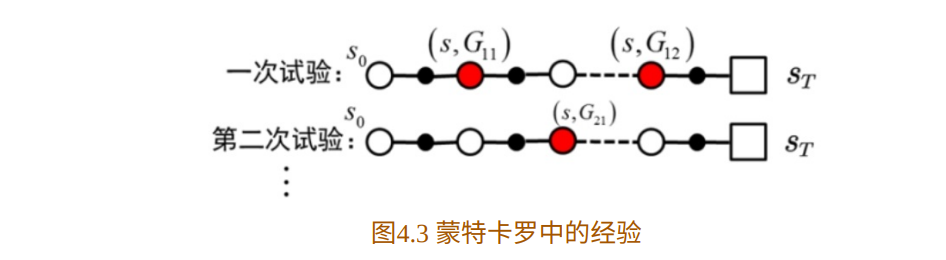
### 策略评估

  在动态规划的方法中，因为模型已知，故其在计算状态处的值函数是利用贝尔曼方程   
  而在无模型强化学习中，模型是未知的。无模型的强化学习算法要想利用策略评估和策略改善的框架，必须采用其他的方法评估当前策略（计算值函数）。 我们回到值函数最原始的定义公式：  
  状态值函数   
  状态-行为值函数   
  状态值函数和行为值函数的计算实际上是计算返回值的期望，动态规划的方法是利用模型计算该期望。在没有模型时，我们可以采用蒙特卡罗的方法计算该期望。

**如何求这个期望呢？**

  蒙特卡罗积分与随机采样方法：  
  蒙特卡罗方法常用来计算函数的积分，如计算下式积分。  
  
  如果的函数形式非常复杂，则(4.13)式无法应用解析的形式计算。这时，我们只能利用数值的方法计算。利用数值的方法计算（4.13）式的积分需要取很多样本点，计算在这些样本点处的值，并对这些值求平均。那么问题来了：如何取这些样本点？如何对样本点处的函数值求平均呢？  
  针对这两个问题，我们可以将（4.13）式等价变换为  
  
  其中为已知的分布。现在就可以回上面的两个问题了。  
  问题一：如何取样本点？  
  答：因为是一个分布，所以可根据该分布进行随机采样，得到采样点。  
  问题二：如何求平均？  
  答：根据分布采样后，在样本点处计算，并对所有样本点处的值求均值：  
  
  以上就是利用蒙特卡罗方法计算积分的原理。

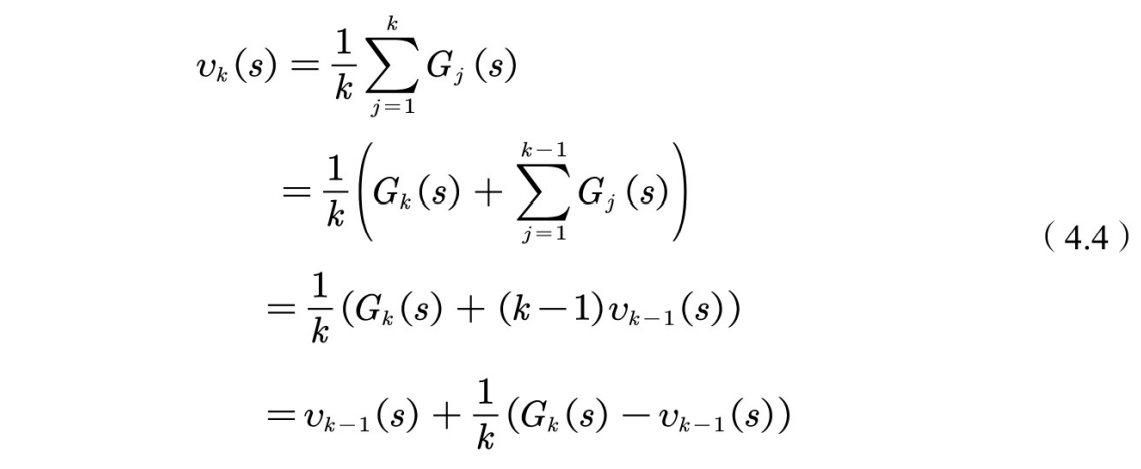
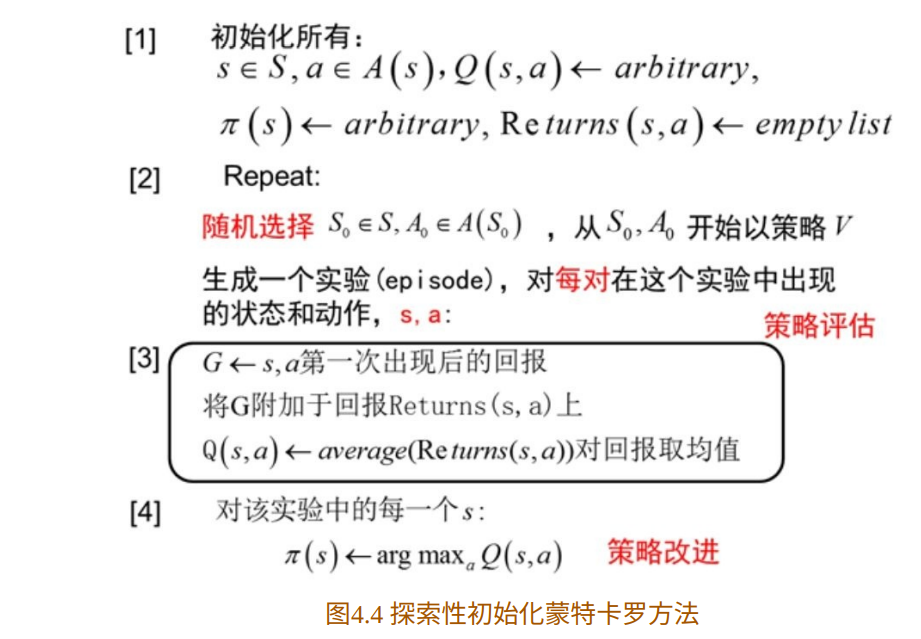
  在没有模型时，我们可以采用蒙特卡罗的方法计算该期望，即利用随机样本估计期望。在计算值函数时，蒙特卡罗方法是利用经验平均代替随机变量的期望。此处，我们要理解两个词：经验和平均。

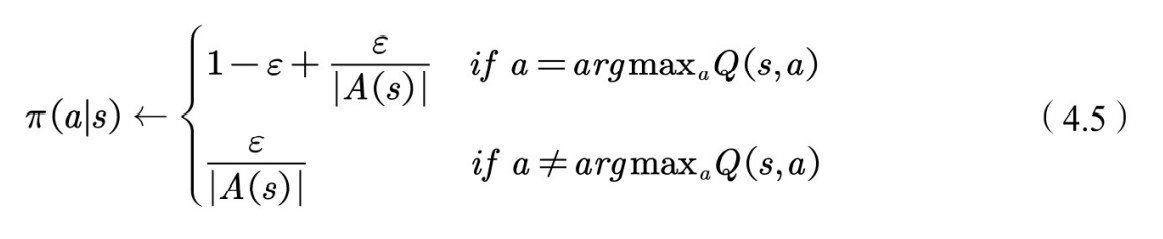
  首先来看下什么是“经验”。  
  当要评估智能体的当前策略时，我们可以利用策略产生很多次试验，每次试验都是从任意的初始状态开始直到终止，比如一次试验（an episode）为 计算一次试验中状态处的折扣回报返回值为，那么“经验”就是指利用该策略做很多次试验，产生很多幕数据（这里的一幕是一次试验的意思），如图4.3所示。  


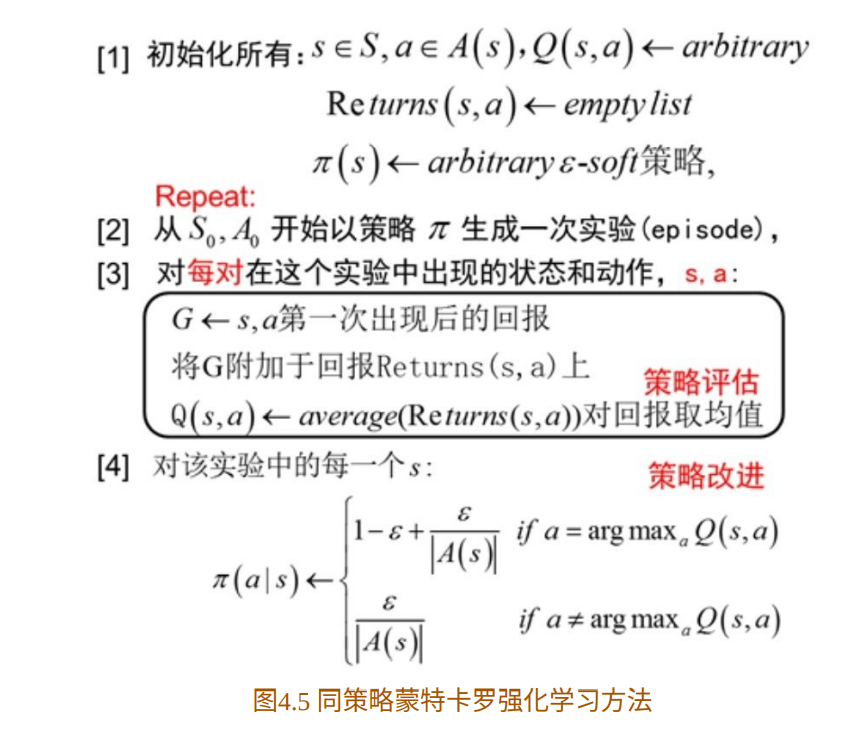
  再来看什么是“平均”。  
  这个概念很简单，平均就是求均值。不过，利用蒙特卡罗方法求状态处的值函数时，又可以分为第一次访问蒙特卡罗方法和每次访问蒙特卡罗方法。 第一次访问蒙特卡罗方法是指在计算状态处的值函数时，只利用每次试验中第一次访问到状态时的返回值。如图4.3中第一次试验所示，计算状态处的均值时只利用，因此第一次访问蒙特卡罗方法的计算公式为  
  
  每次访问蒙特卡罗方法是指在计算状态处的值函数时，利用所有访问到状态时的回报返回值，即  
  
根据大数定律： as   
  由于智能体与环境交互的模型是未知的，蒙特卡罗方法进行策略评估是利用经验平均来估计值函数，而能否得到正确的值函数，则取决于经验——因此，如何获得充足的经验是无模型强化学习的核⼼所在。

**下面就是获取充足经验的方法**

  在动态规划方法中，为了保证值函数的收敛性，算法会逐个扫描状态空间中的状态。无模型的方法充分评估策略值函数的前提是每个状态都能被访问到，因此，在蒙特卡洛方法中必须采用一定的方法保证每个状态都能被访问到，方法之一是探索性初始化。  
  探索性初始化是指每个状态都有一定的几率作为初始状态。在学习基于探索性初始化的蒙特卡罗方法前，我们还需要先了解策略改善方法，以及便于进行迭代计算的平均方法。下面我们分别介绍蒙特卡罗策略改善方法和可递增计算均值的方法。

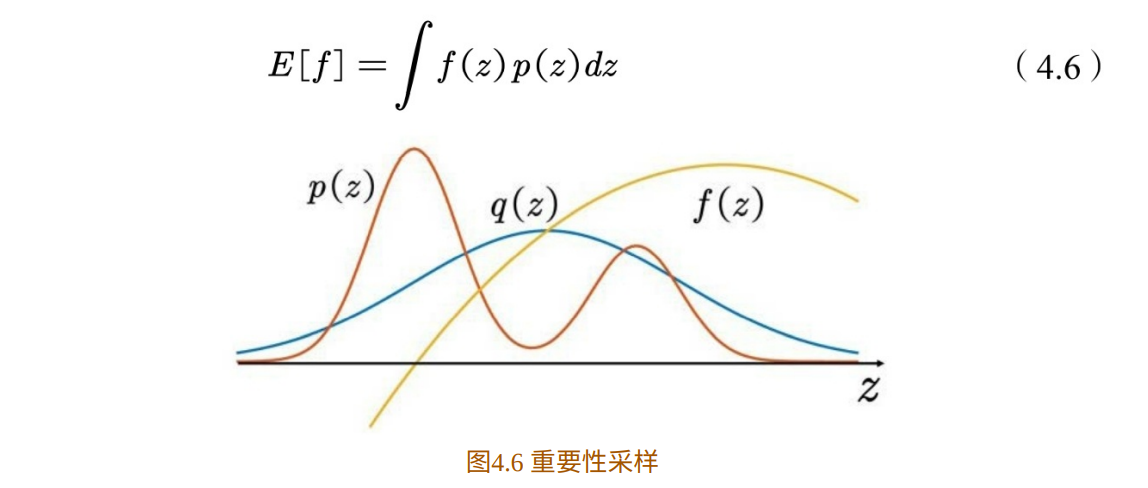
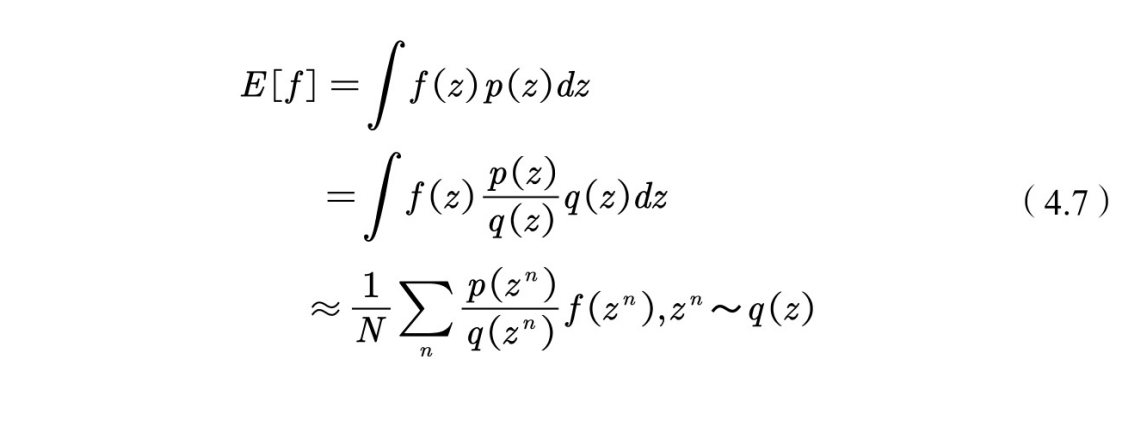
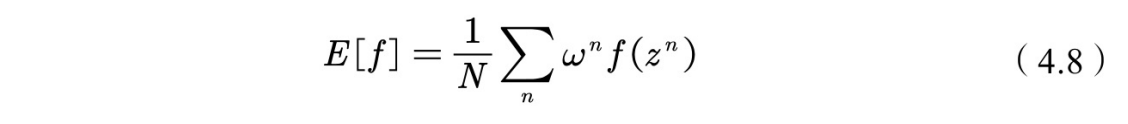
（1）蒙特卡罗策略改善。 蒙特卡罗方法利用经验平均估计策略值函数。估计出值函数后，对于每个状态，它通过最大化动作值函数来进行策略的改善。即  
（2）递增计算均值的方法如（4.4）式所示。  
  
如图4.4所示是探索性初始化蒙特卡罗方法的伪代码  


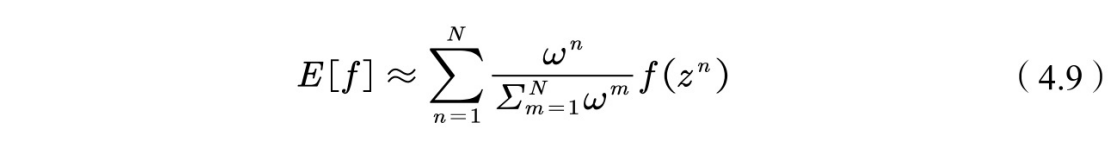
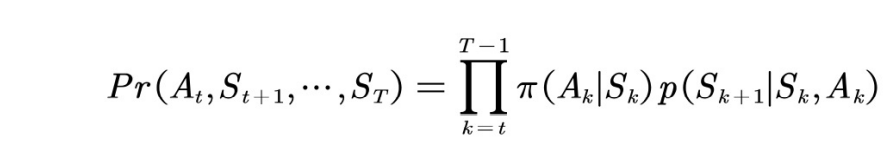
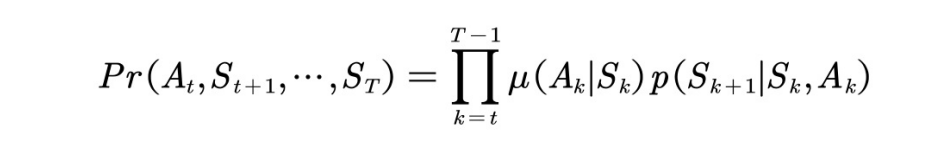
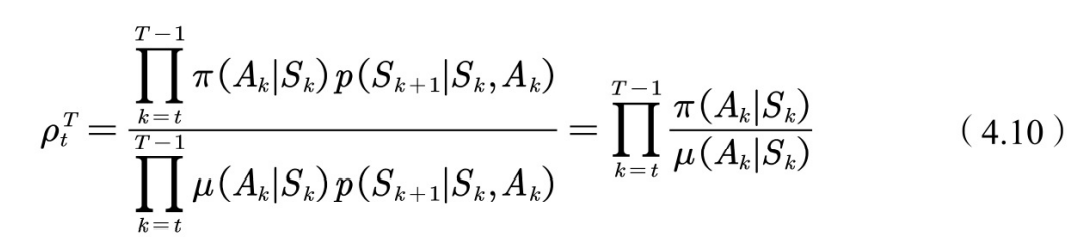
  我们再来讨论探索性初始化。  
  探索性初始化在迭代每一幕时，初始状态是随机分配的，这样可以保证迭代过程中每个状态行为对都能被选中。它蕴含着一个假设：假设所有的动作都被无限频繁选中。对于这个假设，有时很难成立，或无法完全保证。  
  我们会问，如何保证在初始状态不变的同时，又能保证每个状态行为对可以被访问到？  
  答：精⼼设计你的探索策略，以保证每个状态都能被访问到。  
  可是如何精⼼地设计探索策略？符合要求的探索策略应该是什么样的  
  答：策略必须是温和的，即对所有的状态和满⾜：。也就是说，温和的探索策略是指在任意状态下，采用动作集中每个动作的概率都⼤于零。典型的温和策略是-soft策略：  


  根据探索策略（行动策略）和评估的策略是否为同一个策略，蒙特卡罗方法又分为on-policy和off-policy两种方法。  
  若行动策略和评估及改善的策略是同一个策略，我们称为on-policy，可翻译为同策略。  
  若行动策略和评估及改善的策略是不同的策略，我们称为off-policy，可翻译为异策略。  
  接下来我们重点理解这on-policy⽅法和off-policy方法。  
（1）同策略。  
  同策略（on-policy）是指产生数据的策略与评估和要改善的策略是同一个策略。⽐如，要产生数据的策略和评估及要改善的策略都是-soft策略。其伪代码如图4.5所示。  
  
（2）异策略。  
  异策略（off-policy）是指产生数据的策略与评估和改善的策略不是同一个策略。我们用表示用来评估和改善的策略，用表示产生样本数据的策略。

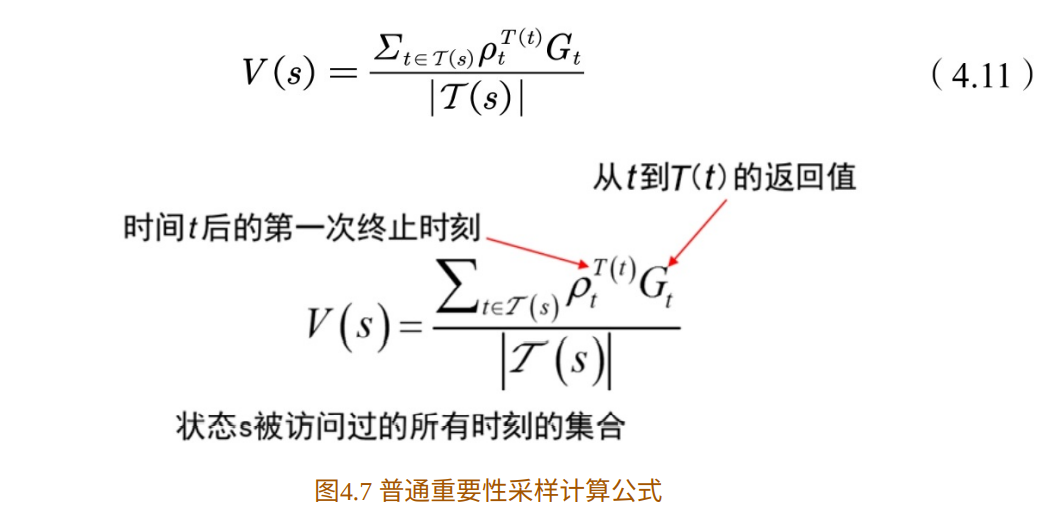
**异策略可以保证充分的探索性。**

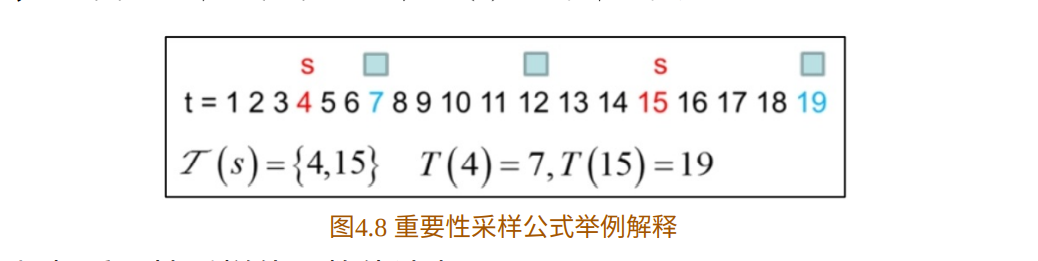
  用于异策略的目标策略和行动策略并非任意选择的，而是必须满足一定的条件。这个条件是覆盖性条件，即行动策略产生的行为覆盖或包含目标策略产生的行为。利用式子表示：满足的任何均满足。

  问题：异策略中，产生数据的策略与评估和改善的策略不是同一个策略，那是否可以利用行为策略产生的数据去评估和改善策略呢？  
  这里就需要利用重要性采样方法。下面，我们介绍重要性采样。  
  我们用图4.6描述重要性采样的原理。重要性采样来源于求期望，如图4.6所示：  
  
  如图4.6所示，当随机变量z的分布非常复杂时，无法利用解析的方法产生用于逼近期望的样本，这时，我们可以选用一个概率分布很简单，很容易产生样本的概率分布，⽐如正态分布。原来的期望可变为  
  
定义重要性权重：，普通的重要性采样求积分如方程(4.7)所示为  


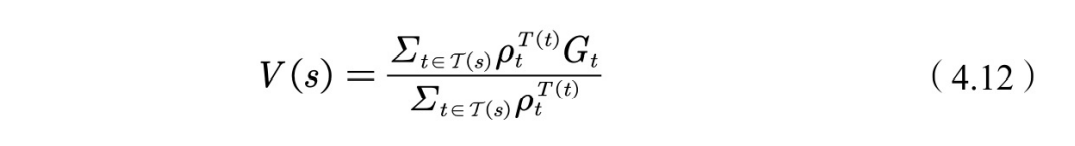
  由式（4.7）可知，基于重要性采样的积分估计为无偏估计，即估计的期望值等于真实的期望。但是，基于重要性采样的积分估计的方差无穷大。这是因为原来的被积函数乘了一个重要性权重，改变了被积函数的形状及分布。尽管被积函数的均值没有发生变化，但方差明显发生改变。  
  在重要性采样中，使用的采样概率分布与原概率分布越接近，方差越⼩。然而，被积函数的概率分布往往很难求得、或很奇怪，因此没有与之相似的简单采样概率分布，如果使用分布差别很大的采样概率对原概率分布进行采样，方差会趋近于无穷大。一种减⼩重要性采样积分方差的方法是采用加权重要性采样：  
  
  在异策略方法中，行动策略即用来产生样本的策略，所产生的轨迹概率分布相当于重要性采样中的，用来评估和改进的策略所对应的轨迹概率分布为 ，因此利用行动策略所产生的累积函数返回值来评 估策略时，需要在累积函数返回值前面乘以重要性权重。 在目标策略下，一次试验的概率为  
  
  在行动策略下，相应的试验的概率为  
  
  因此重要性权重为  


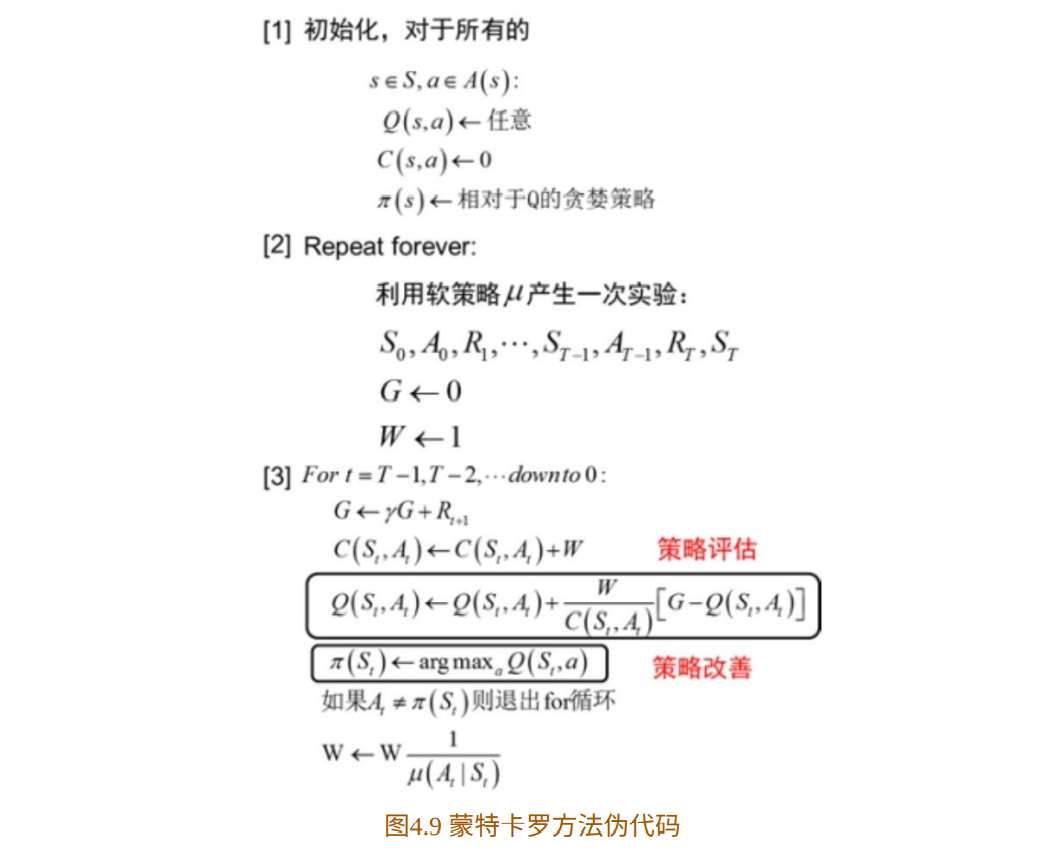
  普通重要性采样的值函数估计如图4.7所示：



  现在举例说明公式（4.11）中各个符号的具体含义。  
  如图4.8所示，是状态访问的时刻，是访问状态相对应的试验的终止状态所对应的时刻。是状态发生的所有时刻集合。在该例中，。  


  加权重要性采样值函数估计为



  最后，我们来看下异策略每次访问蒙特卡罗算法的伪代码，如图4.9所示。  
  
注意：此处的软策略为-soft策略，需要改善的策略为贪婪策略。

  总结一下：本节重点讲解了如何利用MC的方法估计值函数。与基于动态规划的方法相比，基于MC的方法只是在值函数估计上有所不同，在整个框架上则是相同的，即评估当前策略，再利用学到的值函数进行策略改善。本节需要重点理解on-policy 和off-policy的概念，并学会利用重要性采样来评估目标策略的值函数。