# Chapter 5 蒙特卡洛方法

从实际的state actionreward序列中取估计valuefunction获取最优策略。

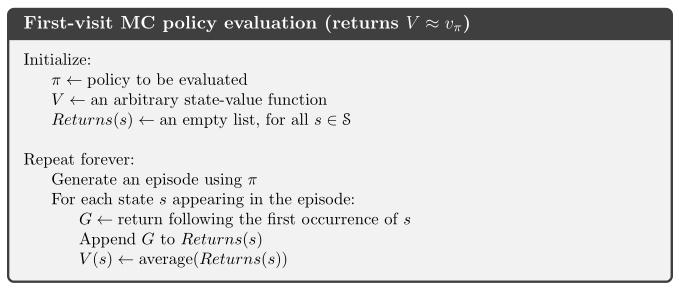
与DP的区别在于：虽然模型还是需要的，但是在每一步不需要考虑不同的transition而只用面向采样到的transition。

蒙特卡洛方法基于episode采用增量的方式，但是不能基于steps。

## 5.1 monte carlo prediction蒙特卡洛预测问题

GPI的第一阶段即：policyevaluation，由于在一个episode中同一个状态可能出现多次，就有了两种evaluation方法：first-visit和every-visit。以常使用的first-visit为例来说明具体过程如下：。

本质上就是求一个平均值，没什么稀奇的。



## 5.2 montecarloestimationofactionvalue

有的时候仅仅有statevalue是不够的，还需要state-actionvalue才能给出有效的policy建议。

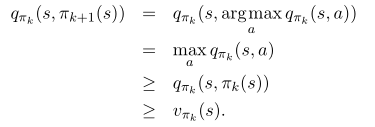
从本质上讲，和上一节的方法没有本质的区别，从维护状态表，到维护状态-action表。存在的问题是:假如policy是一个确定性的策略，那么就只会探索某个state相关的一个action，而其他的action的表现无法计算。

当然有解决的办法，原则就是要保证所有的状态-行为对都以大于0的概率被采样到。比如：exploringstarts、episilon-greedy、off-policy。什么是exploringstarts？就是在一个episode的开端不采用确定的策略，保证所有的action都有可能被采样到。

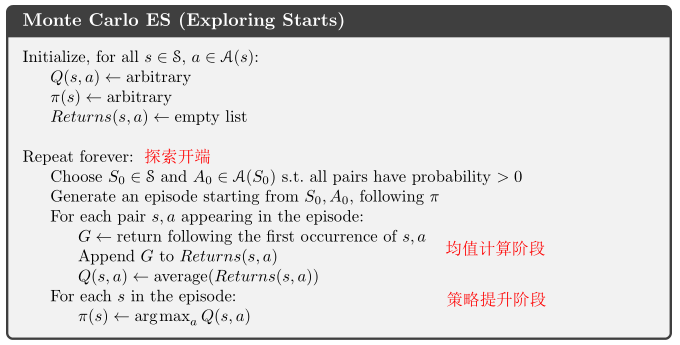
## 5.3 montecarlocontrol蒙特卡洛控制

在本书中，所谓的控制就是GPI第二过程，即策略提升。策略提升的方法就是：基于valuefunction贪心。





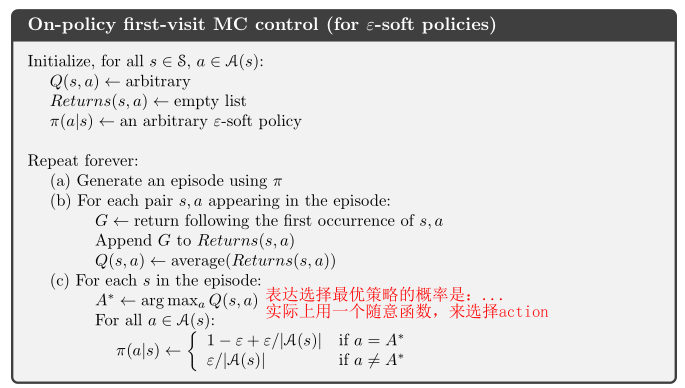
上式说明满足策略提升定理。即构造了一个更好的策略。



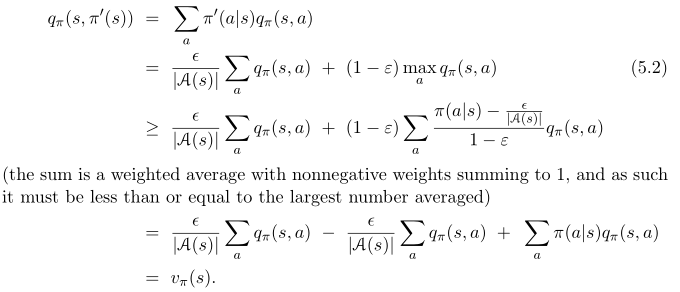
## 5.4Monte Carlo Control without Exploring Starts去除开端探索

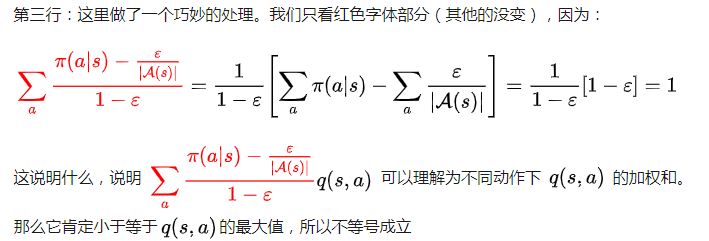
两个方法：on-poliyc方法、off-policy方法。

1、on-ppoliyc方法：即episilon-greedy方法以episilon的概率随机action(包括贪心选择的action)、以1-episilon的概率贪心。



策略提升的证明：就是证明q大于v





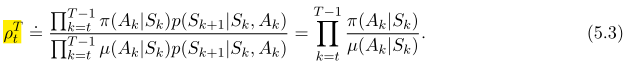
这里参考了(<https://zhuanlan.zhihu.com/p/55731952>)

## 5.5 Off-policy Prediction via Importance Sampling

Off-poliyc的提出是为了解决RL的一个窘境：想要学一个最优的策略，但是又需要表现的不是那么最优从而可以探索所有的可能。上一节的方法是一个折中。一个更好的办法是用两个策略：一个用于产生数据的behavior policy，一个是用来被学习的叫做target policy。从而将数据采样和学习过程分离。

On-policy方式由于简单而被更多的采用，但是off-policy更加强大且通用。在控制问题中，用行为策略保持探索，而目标策略进行贪心，但是现在我们考虑预测问题：目标策略给定而且不变，只是去评估。

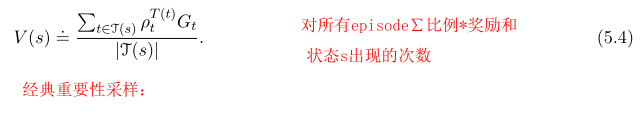
importance sampling:重要性采样 是一种基于一种分布求另外一种分布期望的技术，所谓重要性，就是要为每一个样本分配一个权重。以episode片段为对象，那么定义样本出现率为：

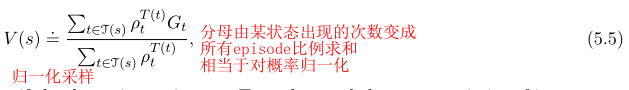


表征实际policy与行为policy生成该片段的概率比值。

|  |  |
| --- | --- |
| J(s) | 状态s出现的时间步的集合 |
| T(t) | 时间t后第一个终态的时间 |
| G(t) | t到T(t)之间的奖励值之和 |
| {Gt}J(S) | 状态s对应的奖励总和 |
| {ptT(t)}J(S) | 样本出现率 |

两种采样方法：(first visit 和every visit均可)





区别在于上式分母为样本数目；下式为权重之和。

以单轨迹分析：

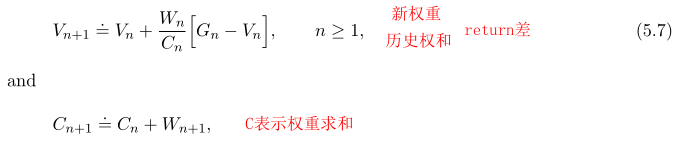
经典：低偏差、高方差

加权：低方差、高偏差(实际更常用)

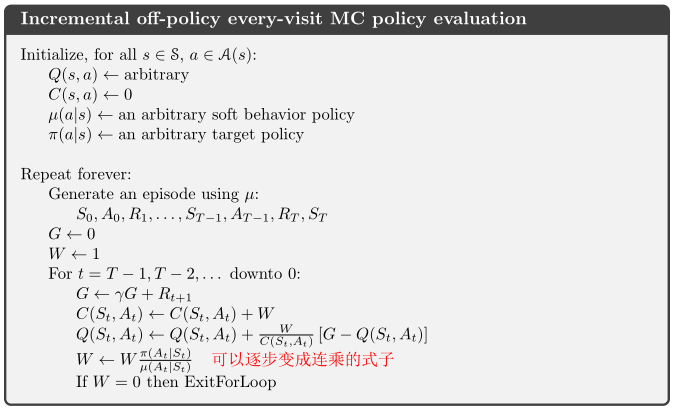
## 5.6 Incremental Implementation 增量方式

采用在线方式或者经典的离线方式都可以之间借鉴第二章的方式来进行增量更新，现在来关注归一化采样的离线增量方法：

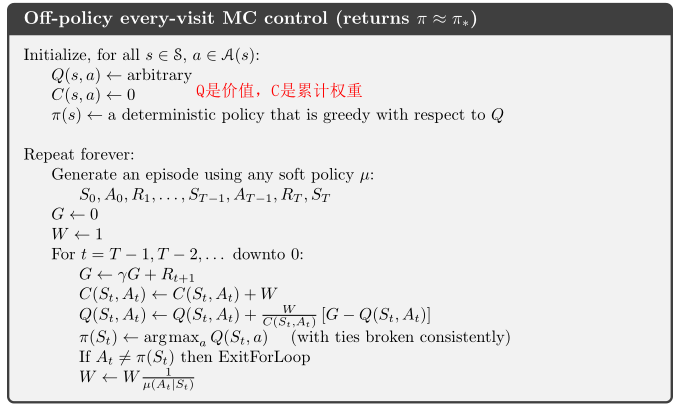




算法如下，之所以要退出是因为：权重为0，表示序列不可能出现在target中。



## 5.7 Off-Policy Monte Carlo Control 蒙特卡洛离线控制



注意算法是先对Q进行更新，然后再预测。假如两个action不一致就会退出。为什么要退出呢？目标策略不会出现这样的(s, a)。既然不会出现，那么内循环下一次迭代计算W的时候就是0。所以有必要在此退出。因为这是monte carlo.

不是TD。

最后一行为什么分子是1？因为是贪心。