# Chapter 5 蒙特卡洛方法

从实际的state action reward 序列中取估计value function 获取最优策略。

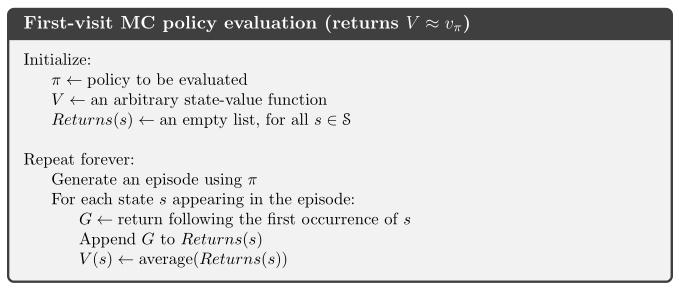
与DP的区别在于：虽然模型还是需要的，但是在每一步不需要考虑不同的transition 而只用面向采样到的transition。

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：policy evaluation，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 monte carlo estimation of action value

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

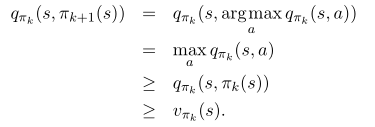
从本质上讲，和上一节的方法没有本质的区别，从维护状态表，到维护状态-action表。存在的问题是:假如policy是一个确定性的策略，那么就只会探索某个state相关的一个action，而其他的action的表现无法计算。

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：exploring starts、episilon-greedy、off-policy。什么是exploring starts？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。

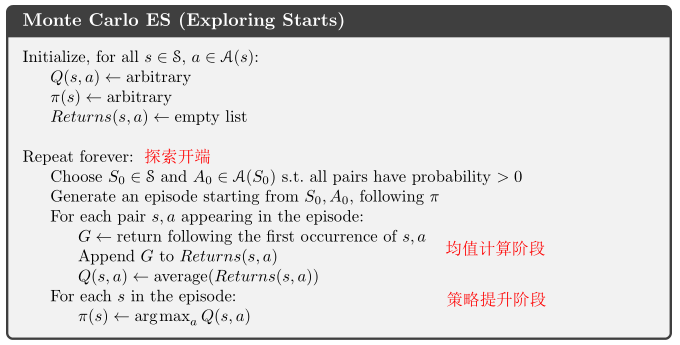
# 5.3 monte carlo control 蒙特卡洛控制

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value function贪心。





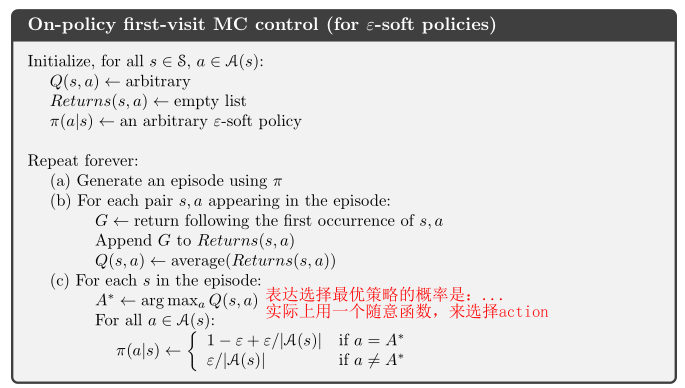
上式说明满足策略提升定理。即构造了一个更好的策略。



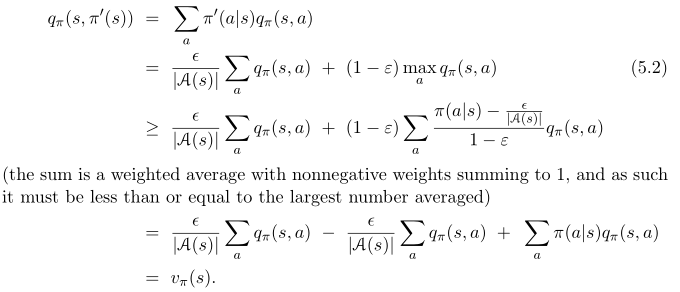
# 5.4 Monte Carlo Control without Exploring Starts 去除开端探索

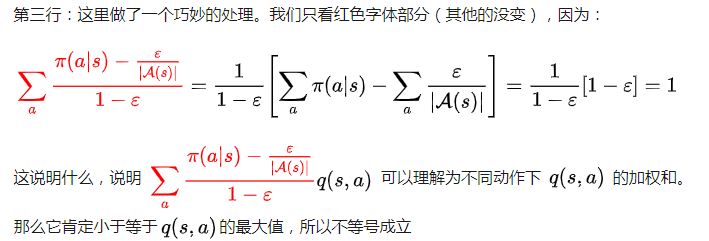
两个方法：on-poliyc方法、off-policy方法。

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的提出只要是为了解决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，一般行为策略保持探索，而目标策略贪心，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。