

图1.1 强化学习分类

1、蒙特卡洛方法：

在没有模型时，我们可以采用蒙特卡罗的方法计算该期望，即利用随机样本来估计期望。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。经验是指通过很多次试验，产生很多次的数据，每次试验都是从任意的初始状态开始到终止状态，如图1.2。完整的episode指必须从某一个状态开始，Agent与Environment交互直到终止状态，环境给出终止状态的即时收获为止。例如走迷宫，一次试验（episode）就是从开始进入迷宫到最后走出迷宫的路径。

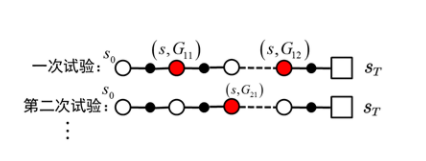
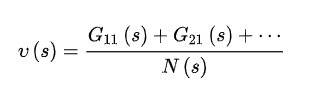
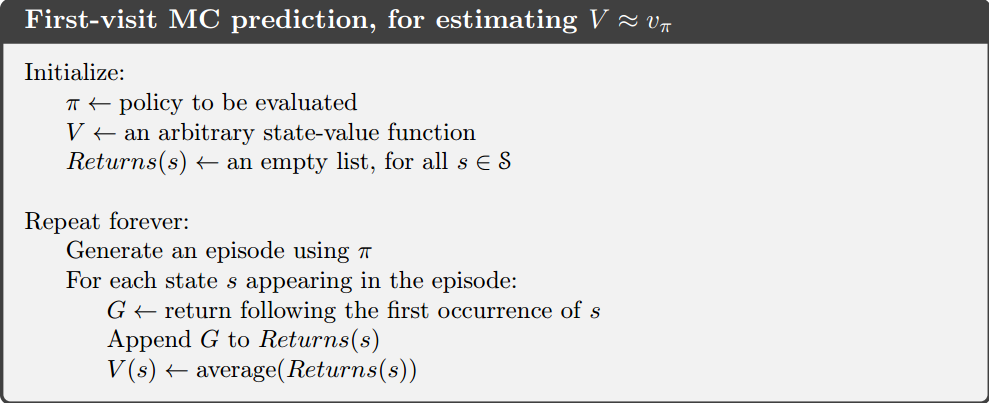


图1.2

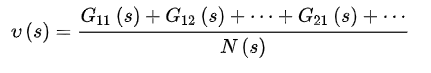
平均是指状态s在每一个样本中的回报的均值。在蒙特卡洛方法中分为first visit和every visit。first visit是指在计算状态s处的值函数时，只利用每次试验中第一次访问到状态s时返回的值，如上图1.2所示，计算s出的均值只利用了，因此计算公式如下：



伪代码如下所示：

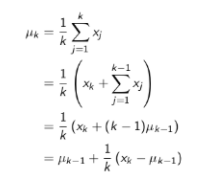


every visit是指在计算状态s处的值函数时，利用所有访问到状态s时的回报值。即：



1. 蒙特卡洛累进评估

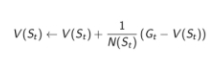
在实际操作时常用的一个实时更新均值的办法，使得在计算平均收获时不需要存储所有既往收获，而是每得到一次收获，就计算其平均收获。公式如下：



对于一系列Episodes中的每一个：

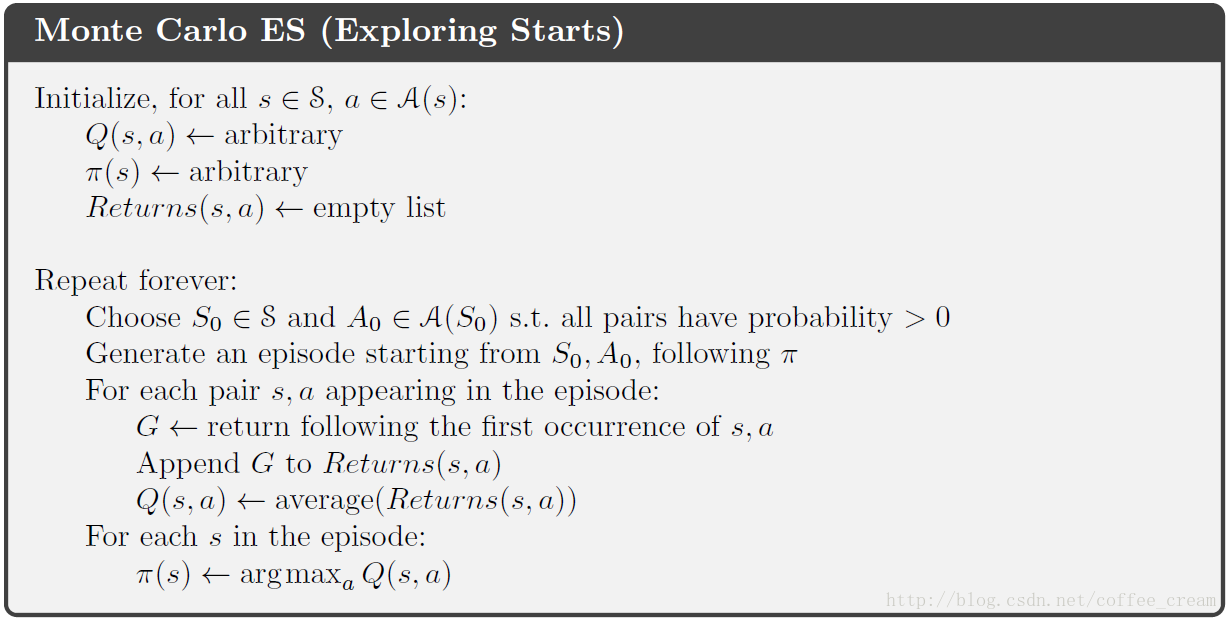


对于Episode里的每一个状态，有一个收获，每碰到一次,使用下式计算状态的平均价值V()：





伪代码如下所示：



3、优缺点

蒙特卡洛强化学习有如下特点：不基于模型本身，直接从经历过的Episode中学习，必须是完整的Episode，使用的思想就是用平均收获值代替价值。理论上Episode越多，结果越准确。如何获得充足的经验是无模型强化学习的核心所在。

同时，蒙特卡洛学习方法的缺点很明显，在每一次测试结束后才开始计算，且需要尽可能多的测试，需要大量的时间，对于确定性的问题需要转化为随机性的问题，误差为概率误差等。

4、示例

二十一点游戏 Blackjack Example

1. https://blog.csdn.net/qq\_37266917/article/details/107350258
2. <https://blog.csdn.net/coffee_cream/article/details/66972281>
3. <https://zhuanlan.zhihu.com/p/25743759>
4. <https://blog.csdn.net/liweibin1994/article/details/79111536/>
5. <https://zhuanlan.zhihu.com/p/28107168>
6. <https://zhuanlan.zhihu.com/p/34395444>
7. Reinforcement learning: an introduction.2017 Draft.