# 如何用蒙特卡罗方法（Monte-Carlo）估算V值？

那怎么估算每个节点的价值呢？

(1)动态规划虽然能解决问题，但对资源的消耗实在太多，你看，你看又是小猴，又是老猴的。而且，还必须了解整个环境的构造。面对一些未知的环境还真不好搞。

(2)蒙特卡罗会让智能体从某个状态S出发，直到最终状态，然后回过头来给每个节点标记这次的价值G。G代表了某次，智能体在这个节点的价值。

本篇将会介绍如何用蒙特卡罗算法求状态V值。

蒙特卡罗是著名的赌城，少数机智赌徒为了赢钱，活生生被逼成了数学家。因此在这里出现了很多有关概率的理论。大家有兴趣可以去看一下。

### 先看故事



话说，齐天大圣孙悟空，护送唐三藏到西方取经。路过一神秘洞穴，突然被怪物的法宝葫芦吸了进去。

葫芦里面原来是另外一个次元的世界，大圣转了两圈，怎么都出不去。气的大圣破口大骂，可是又破坏不了这葫芦。正在踌躇之际，前面来了一个人。

说道：大圣呀，我是来自西域的蒙特卡罗，我被困在这里很久啦。

这个迷宫呀，不但要找到出口，而且要找出从任何一个位置离出口最近的路呀。

但大圣请放心，只要大圣助我一臂之力，我定可以解决这个问题。

大圣，你如此如此，这般这般便可。

大圣听道，连忙点头称是。

于是大圣从身上拔下一条汗毛，轻轻一吹，变成了小猴。于是大圣开始吩咐小猴：

1. 你从这里出去之后，就随便逛，走的时候要记住走的路径和每一步获得的奖励。
2. 如果你到了最终出口，那就沿原路往回走，在每个经过的路口，留下一张字条： 字条上写上你获得的奖励,再加上你上一个路口的标记的路口的价值。这个价值你可以乘以一个折扣，嗯，就乘以0.9吧。
3. 你就这样一直沿路回来可以了。

于是小猴一蹦一跳就出去了，过了一会，回来了。

这时候大圣正在吃香蕉，看到小猴回来，扔了一根香蕉给小猴，说： 你继续出去吧，和上次一样就好啦。

小猴不解道：那如果遇到我上次见到的路口，我是应该怎么做呢？

大圣想了一下说：你不用管，你把这一次的数给记下就可以了。大圣又给了小猴一根香蕉。让小猴继续跑。

小猴就是这样，一直跑一直跑，跑了许多次，蒙特卡罗的赌徒跟大圣说：大圣爷，我看差不多了，你去让小猴子，把所有的数都给平均一下，就是要求的V值了。

### 蒙特卡罗算法

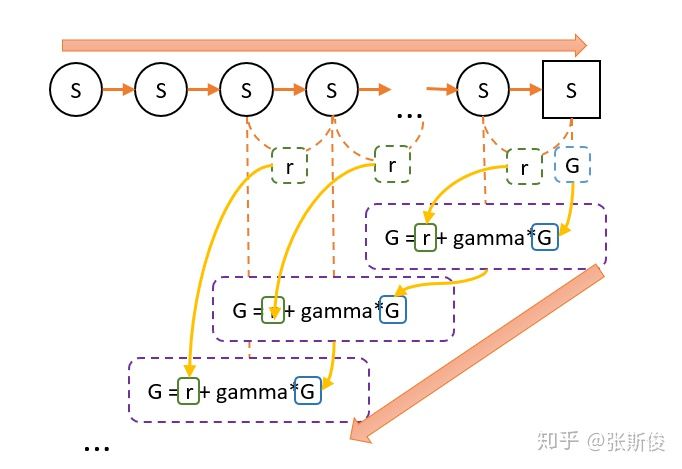
看完上面的故事，应该对蒙特卡罗算法有大概的认识。现在来具体看看蒙特卡罗算法。

1. 把智能体放到环境的任意状态；
2. 从这个状态开始按照策略进行选择动作，并进入新的状态。
3. 重复步骤2，直到最终状态；
4. 从最终状态开始向前回溯：计算每个状态的G值(节点价值)。
5. 重复1-4多次，然后平均每个状态的G值，这就是需要求的V值。

### 原理

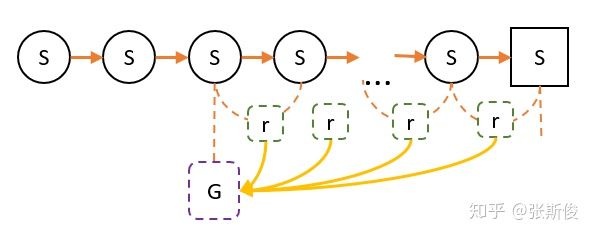
其实从直觉上，很容易理解。

分成两部分，首先要理解每一次算的G值的意义(就是V值到V值的计算)。

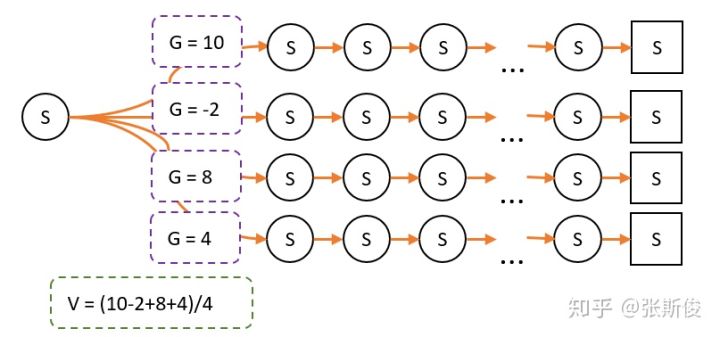


* 1. 根据策略往前走，一直走到最后，期间什么都不用算，只需要记录每一个状态转移时获得的奖励r即可。
* 2. 从终点往后走，一边走一边计算G值。G值等于上一个状态的G值(记作G'),乘以一定的折扣(gamma),再加上r。

所以G值的意义在于，在这一次游戏中，某个状态到最终状态的奖励总和(理解时可以忽略折扣值)



当进行多次试验后，有可能会经过某个状态多次，通过回溯，也会有多个G值。 重复刚才说的，每一个G值，就是每次到最终状态获得的奖励总和。而V值是某个状态下，通过影分身到达最终状态，所有影分身获得的奖励的平均值。



所以理解清楚V值和Q值的意义，理解蒙特卡罗其实很轻松的。

### 再进一步

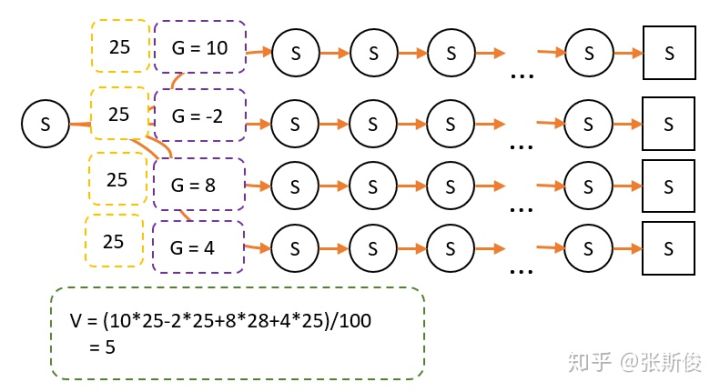
蒙特卡罗算法单独拿出来，在强化学习中效率还是比较低的。所以会结合其他的方式进行应用。这些会在后面具体算法中讲到。

在这一篇中，希望大家能够理解两点：

1. G的意义：在某个路径上，状态S到最终状态的总收获。
2. V和G的关系：V是G的平均数。

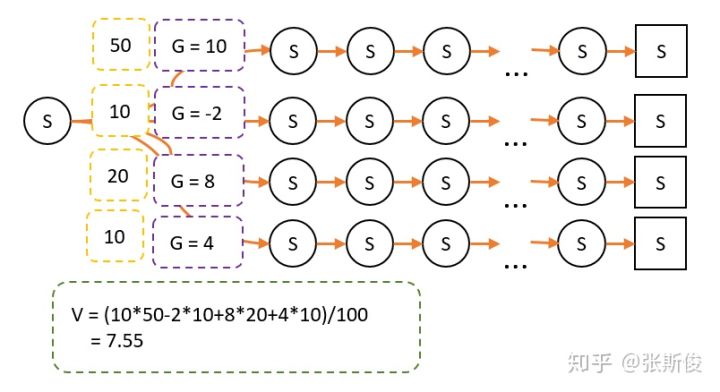
到这里要注意一点：V和策略是相关的，那么在这里怎么体现呢？这个非常重要，因为在PPO算法中，离线策略就与这个有关。这里可以稍微先说一下。

仍以上图为例子，以策略A进行游戏。其中有100次经过S点，经过S点后有4条路径到达最终状态，计算G值和每条路径次数分别如下：



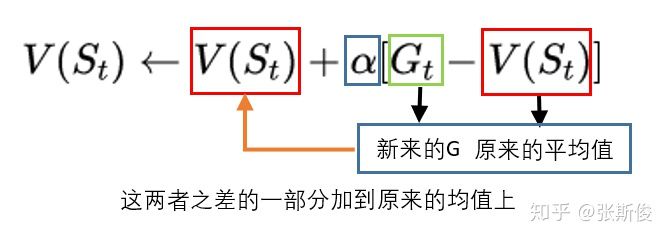
策略A采用平均策略，这时候 V = 5。

现在采用策略B，**由于策略改变，经过某条路径的概率就会产生变化。因此最终试验经过的次数就不一样了。**



最终计算的 V = 7.55。

但为了方便，我们对平均进行一些优化。于是获得用MC估算V值的公式：



### 蒙特卡罗的缺陷

在实际引用中，蒙特卡罗虽然比动态规划消耗要少一点；而且并不需要知道整个环境模型。

但蒙特卡罗有一个比较大的缺点，就是每一次游戏，都需要先从头走到尾，再进行回溯更新。如果最终状态很难达到，那小猴子可能每一次都要转很久很久才能更新一次G值。

在下一篇，将介绍时序差分(TD)算法来解决这个问题。

===========你的支持，就是在下努力的原动力===========

如果专栏对你有用，请点赞并关注在下喔。如果发现有问题，也可以在文章下留言。

你的每一点关注，都是在下的继续努力的动力来源！感激！

[白话强化学习​](https://zhuanlan.zhihu.com/c_1215667894253830144" \t "/home/admin/文档\\x/_blank)zhuanlan.zhihu.com[](https://zhuanlan.zhihu.com/c_1215667894253830144)