# 如何用时序差分TD估算状态V值？

[](https://www.zhihu.com/people/zhang-si-jun-52)

[张斯俊](https://www.zhihu.com/people/zhang-si-jun-52" \t "/home/admin/文档\\x/_blank)

愿成为一把梯子，助你跨过无数的坑。

### 蒙特卡罗的限制

虽然蒙特卡罗算法能够在不知道环境信息的时候，以采样的方式估算V值。但这种方法也是有一定的限制。

1. 蒙特卡罗算法相对动态规划，会有点不那么准。因为蒙特卡罗每一次的路径都是不一样的。
2. 如果环境的状态空间非常大，或者最终状态只有非常小的概率达到。那么蒙特卡罗算法将会很难处理。

关于蒙特卡罗MC在马可洛夫的运用，可以参考本专栏之前的文章。

[张斯俊：如何用蒙特卡罗方法（Monte-Carlo）估算V值？​](https://zhuanlan.zhihu.com/p/109755443" \t "/home/admin/文档\\x/_blank)

[张斯俊：[番外]蒙特卡罗MC的更新公式怎么来的？​](https://zhuanlan.zhihu.com/p/110118392" \t "/home/admin/文档\\x/_blank)

### 时序差分算法

原文再续，书接上一回。

话说，孙大圣和来自蒙特卡罗的小蒙，一起解决了第一个问题。而让人想不到的是，还有第二关，而且这关的迷宫更大了。

没办法，大圣只能按照之前的方法去计算V值。

但这迷宫实在太大了，小猴子根本找不到出口。终于，疲惫后的猴子实在太累了。走到一半，就回头更新了。

然而，这怎么瞒得过孙大圣的火眼金睛，大圣正想发怒。小蒙却阻止道：等等，让我想想。

几分钟后，小蒙惊喜道：大圣爷，想不到呀，这小猴乱打乱撞，却让我发现了新的算法。

大圣狐疑道：真的？！

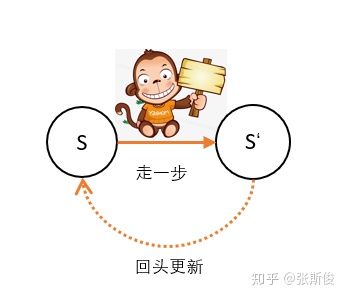
小蒙：真的，姑且把这个算法叫做时序差分算法吧。

大圣：那你说来听听，如果没道理，哼，吃老孙一棒。

小蒙：是的是的

于是小蒙便把这个时序差分算法说出来，为了方便，以后就叫TD算法：

1. 小猴子每走1步，看一下这个路口的V值，还有获得的奖励r；
2. 回到原来的路口，把刚刚看到的V值和奖励r进行运算，估算出V值。



于是小猴子便用这种方式，很快地计算出各个路口的V值。

### TD和MC的比较

TD算法对蒙特卡罗(MC)进行了改进。

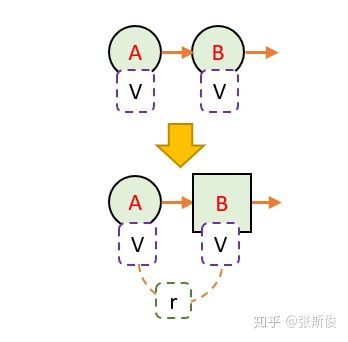
1. 和蒙特卡罗(MC)不同：TD算法只需要走N步。就可以开始回溯更新。
2. 和蒙特卡罗(MC)一样：小猴需要先走N步，每经过一个状态，把奖励记录下来。然后开始回溯。 3. 那么，状态的V值怎么算呢？其实和蒙特卡罗一样，就假设N步之后，就到达了最终状态了。假设“最终状态”之前没有走过，所以这个状态上的纸是空白的。这个时候就当这个状态为0;假设“最终状态”已经走过了，这个状态的V值，就是当前值。然后开始回溯。

### TD原理的直观理解

可以把TD看成是这样一种情况：

从A状态，经过1步，到B状态。什么都不管就当B状态是最终状态了。

但B状态本身就带有一定的价值，也就是V值。其意义就是从B状态到最终状态的总价值期望。(这一点在之前Q值和V值那篇已经说明过，就不在赘述了。)



假设B状态的V值是对的，那么，通过回溯计算，就能知道A状态的更新目标了。

这就有点像从山顶像知道要下山的路有多长。 MC能直接走一趟，看一下到底有多远。 TD则轻巧一点，先走一段路看一下，看一下有没有路牌指示到山脚还有多远。如果有，那么就把刚刚走的那段路加上路牌指示到山脚的距离相加即可。 但又同学可能会问，在一开始，根本没有路牌呀，所以也不知道到底到山脚有多远。 没错，这是对的。但当走很多次的时候，路牌系统就能慢慢建立起来。 例如第一次，只有到了山脚，我才知道山脚前一站离山脚的的真实距离。于是我更新了山脚前一站的路牌。第二次，我在山脚前一站路就能看到路牌，所以我就可以更新山脚前一站的路牌了...一直到山顶，就这样一直建立整座山的路牌系统。

### 更新公式

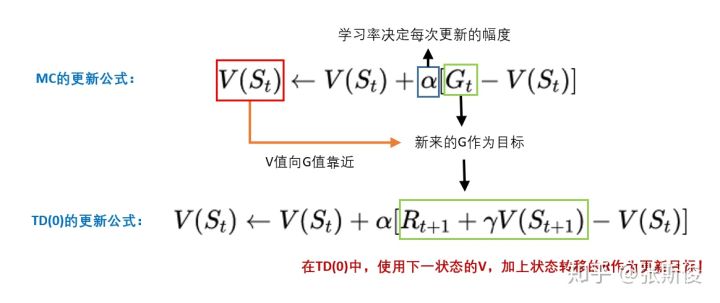
刚刚对TD有个直观的理解：TD并未走完整段路程，而是半路就截断。用半路的路牌，更新当前的路牌。

所以只需要把MC的更新目标，改为TD的更新目标即可。

关于MC的更新目标的理解，可以先看这篇文章。

[张斯俊：[番外]蒙特卡罗MC的更新公式怎么来的？​](https://zhuanlan.zhihu.com/p/110118392" \t "/home/admin/文档\\x/_blank)

在MC，G是更新目标，而在TD，只不过把更新目标从G，改成r+gamma\*V



### 一些额外理解

### 一、整个MC和TD的学习，让我有一个直观的感觉。跟物理的能量转换有一点相似。

在物理中，球从上往下自由落体，是势能转化为动能的过程。随着势能逐渐减少，兑换出来的动能将会越来越多。这是不是有点像V值随着智能体的前进，会慢慢兑现为R值？

又例如:下棋的时候，说的势，其实是V。在开局的时候，必须必对手有更高的势，在中盘对子，就是兑现自己的V，转换为r。

**二、TD更厉害的是，在很多时候，并不需要一直到最后，可以先用后面的估算，然后调整当前状态。**

再走一步。 这种想法后来应用到很多算法上ALPHA ZORO中截断蒙特卡罗搜索树，TD-CFR等。如果状态太多，深度太长，可以用这种方式先截断。然后想办法把后面的V。通过多次就可以更新。

### TD之于Q值估算, 参考: Qlearning---怎样直观理解Qlearning算法.docx

===========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)