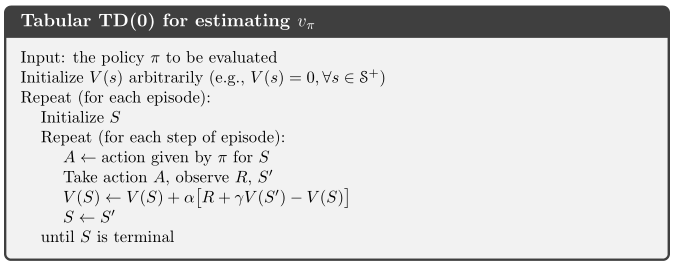
# chapter6 TD Learning

## 6.1 TD prediction

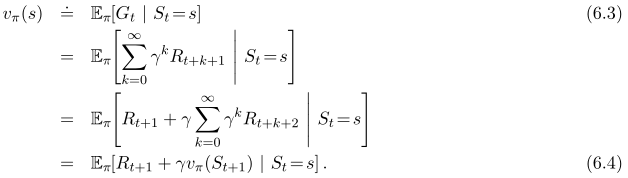
MC与TD都是实践派，区别在于MC在episode级别更新，TD在step级别更新。直觉的想法是：MC偏差小，TD方差小。

****



自举：自举可以视为迭代，从过去的估计中计算当前的估计，TD与DP都有自举的思想。

MC/TD/DP对比：



MC：基于采样累计奖励的G代替理论上的G，收敛期望。

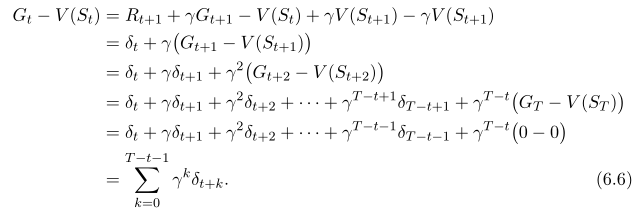
D P：基于已知的转移概率，用对下一时刻的估计(V(St+1))代替理论，收敛期望。

TD：基于采样的单步奖励，使用两个状态的估计值，收敛期望。

在备份图中，MC与TD是sample back up，DP为full back up

TD-error与MC-error：





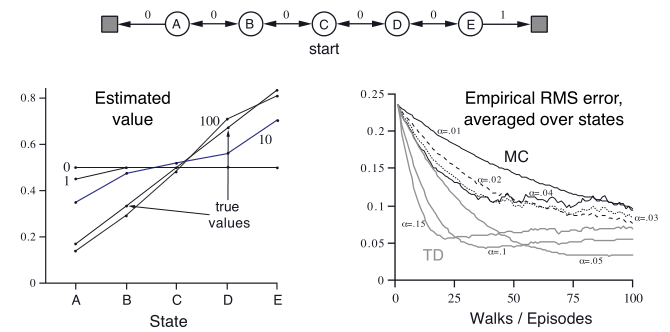
## 6.2 TD 预测的好处

TD：learn a guess from a guess。有效吗？

优势：与DP相比不需要环境建模，与MC比不需要等到episode结束。实现完全的在线、增量方式更新。

结论：TD会收敛到真值。条件：步长参数足够下(倒数足够大)

下一个问题：MC和TD谁快？



游戏描述：每个状态都有0.5的概率向左或者向右。那么状态vlue真值就是到达右侧终态的概率。V(c)=0.5，然后可以根据DP计算其他状态:

A:1/6----E:5/6

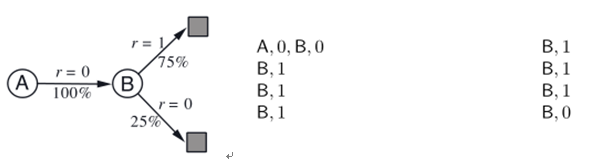
左图：迭代100次以后，TD基本收敛。

右图：TD比MC误差小。

当然这只是一个例子，学术界没有证明。

## 6.3 TD(0)的最优性。

有的时候MC和TD会收敛到不同的值。能够理解其中的原因有助于我们理解这两种方法的区别。模型和数据如下：

 根据数据首先计算V(A):

1、MC： V(A)=0 # 只有一个episode

2、TD： V(A)=V(B)=6/8=3/4 # 批处理方式，否则也是0

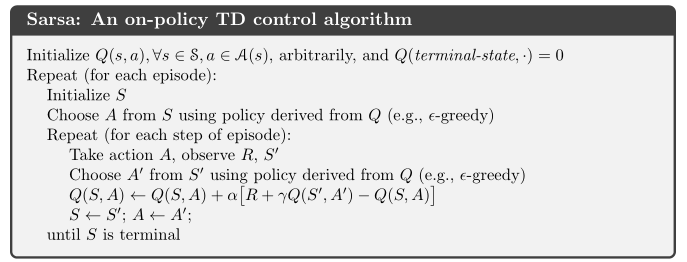
实际上是没有模型的，只是TD方法对可map成这样一个MDP模型。仅仅从数据上来看，MC方法的偏差为0，是最小的，但是假如模型具有MDP属性，我们认为TD方法有较小的误差。

总结：MC期望估计值与训练数据的均方误差最小；TD希望与MDP的最大似然模型保持一致。通常而言，最大似然模型对应生成实验数据概率最大的参数值，基于最大似然模型与MDP，那么我们可以精确的计算出状态的value值。这称之为确定性等价估计。一般而言TD(0)方法收敛于确定性等价估计。确定性等价估计具有某种层面的最优性，但是直接用于解决问题还是复杂度太高，在这种情况下，TD方法可能是近似确定性等价解的唯一方法。

## 6.4 Sarsa：on-policy TD Control-在线TD策略优化



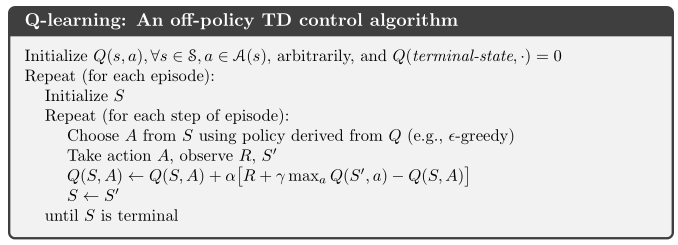
可以看出：一次更新使用了(St , At,Rt+1 ,St+1 ,At+1 )，这也被称作第一个transition，同时也是Sarsa的由来。



假如采用episilon-贪心的方法，通过将epision设置为t的函数，可以逐步将探索比例降低，最终sarsa完全收敛到最优策略。另外，假如St+1是终态的话，Q被设置为0。

## 6.5 Q-learning：off-policy TD control 离线策略优化



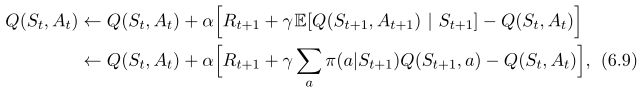


与Sarsa相比：

Sarsa需要一个policy，是一个通常意义上的GPI模型，收敛于Qπ，但是Q-learning直接收敛于Q\*,而且与策略无关。

另外：为什么叫做是离线呢？离线又为什么没有权重比呢？因为和策略无关，所以叫做off-policy，因为单step，所以w=1。

## 6.6 Expected Sarsa-平均Sarsa

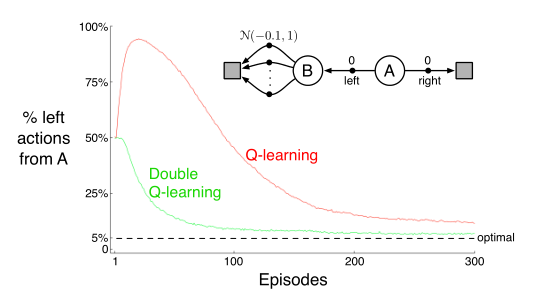


平均Sarsa用复杂度换取更好的性能(Sarsa是单步采样，现在是期望值，偏差更小)，以及步长参数的更广范围。Sarsa既可以在线使用，也可以离线使用，当behavior为一个探索policy，而target policy为贪心的时候，平均Sarsa就转化为Q-learning，所以，平均Sarsa在改善Sarsa的同时还涵盖了Q-learning。

question：哪里有期望值呢？括号前面是state-value function。

## 6.7 Maximization Bias and Double Learning

正向偏差与double learning：



虽然向左的均值为负，小于向右的选择，但是由于向左是一个分布，这就会诱导agent犯错误。随着数据不断更新，模型发现还是向右比较香。所以后面逐渐收敛到5%(Episilon的幅度)。

修正方法：注意到Q-learning的更新方法是：



也就是先基于Q网络对sa估计，然后再求一个最大值。显然Q是有偏的，在此基础上求最大值意义不大，所以可以考虑使用两个网络，一个用来选择action，而用另外的网络来计算其Q值，这样可以消除偏差。这就是double Q-learning。

