# 第三章 有限马尔科夫

## 3.4 回报的统一描述

episode task片段式任务，也就是有一个终态，那么：



continuing tasks连续任务：

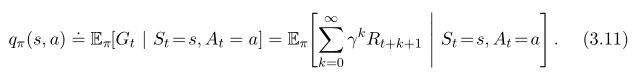
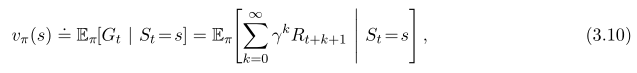


统一描述：



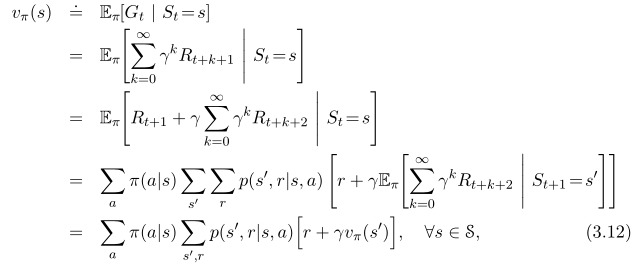
## 3.5 value function

定义：基于某个策略，面向某个状态s，能获得的奖励的期望值：



对于Vπ和qπ可以用经验值进行估计：例如我们可以通过计算实验数据中每个状态的奖励期望，来计算state value function，假如我们独立的统计出不同action后经历的期望值，就可以估计action value function，这种就是后面会学的**蒙特卡洛**方法:。如果state非常多的话，再利用这种方法就不太实际，取而代之的是**基于参数**:的估计方法，会在本书的第二部分介绍。

Value function的重要特点是满足递归关系（贝尔曼等式）:如下：



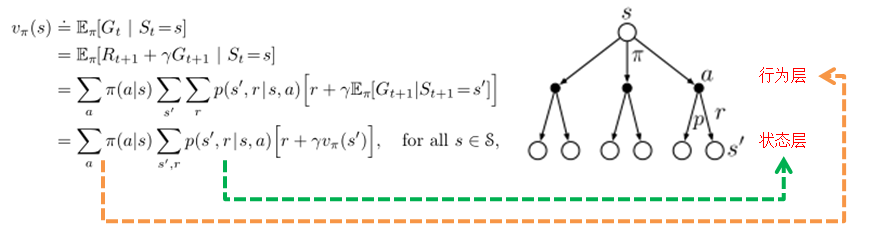
当采取某个action的奖励和后继状态唯一时，可以简化为：

Vπ(s) = ∑p(a|s)\*[r+γVπ(s+1)]

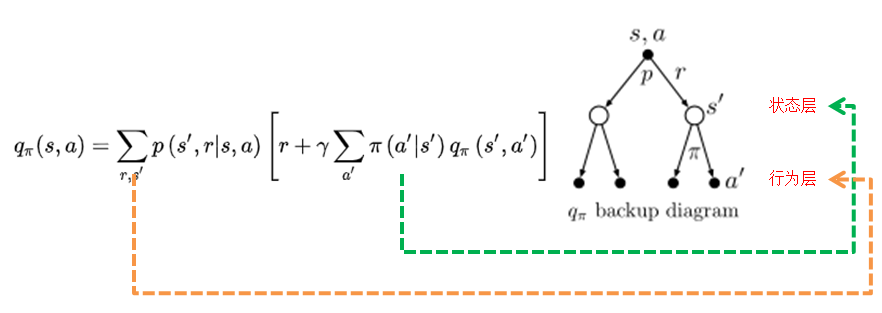
Vπ(s) = E[r+γVπ(s+1)] # 当前状态的价值是后继状态价值的期望。就像一个评价一个老师，可以观察他学生的分数，如果一个班的成绩特别好，我们就认为这个老师的水平就很高。

**附：贝尔曼等式汇总：**

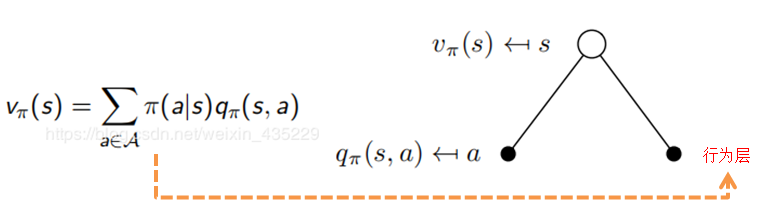
1、v-v



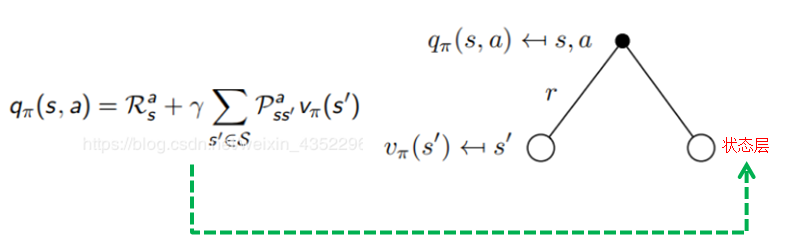
2、q-q



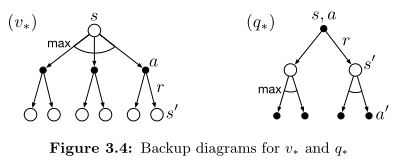
3、v-q



3、q-v

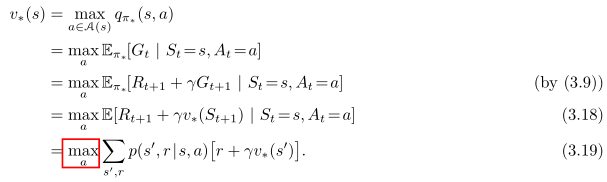


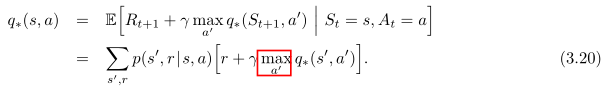
## 3.6 最优策略与最优价值函数











最优小结：

最优最终都落到了最佳的action上了。

价值函数是期望、最优函数是最值。