# 如何理解强化学习中的Q值和V值？

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

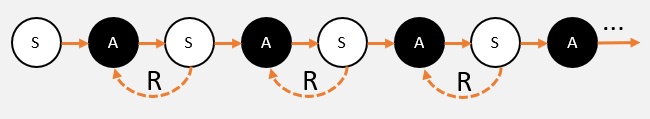
[张斯俊](https://www.zhihu.com/people/zhang-si-jun-52" \t "/home/admin/文档\\x/_blank)

愿成为一把梯子，助你跨过无数的坑。

### 赢在将来

要理解Q值和V值，就需要先理解马尔科夫树，为什么我要把马尔科夫链叫成马尔科夫树呢？我在之前的文章已经说过。如果有需要可以看看。

[张斯俊：1.0 马尔科夫链​](https://zhuanlan.zhihu.com/p/109217883" \t "/home/admin/文档\\x/_blank)

[](https://zhuanlan.zhihu.com/p/109217883)

马尔科夫告诉： 当智能体从一个状态S，选择动作A，会进入另外一个状态S'；同时，也会给智能体奖励R。 奖励既有正，也有负。正代表鼓励智能体在这个状态下继续这么做；负得话代表并不希望智能体这么做。 在强化学习中，会用奖励R作为智能体学习的引导，期望智能体获得尽可能多的奖励。

但更多的时候，并不能单纯通过R来衡量一个动作的好坏。来看下面一个例子：

假设，10天之后进行期末考试，今天有两个选择：

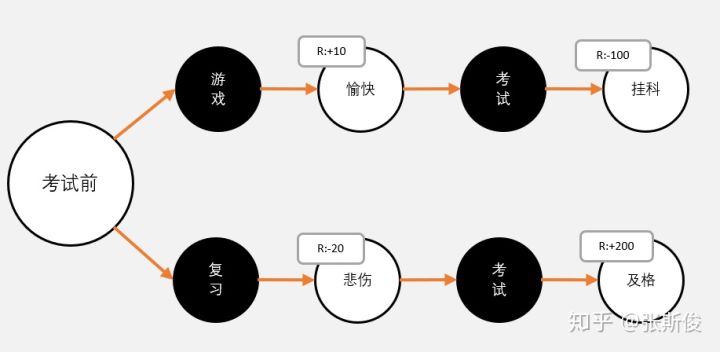
1. 放弃吧，玩游戏！每天可以获得+1心情值；
2. 决心努力一搏，开始学习吧！每天-2心情值。

从这10天看，肯定是选择【1.玩游戏】。因为10天后，虽然考试没过，但至少收获10天的快乐。

但事实上，再看远一点：

因为挂科，接受老师怒吼攻击！心情值马上减5；

父母因为我考得好成绩，给了更多的零用钱。心情值加200点。



所以，假设能预知未来，一定会选择【2.去复习】

因此，必须用长远的眼光来看待问题。要把未来的奖励也计算到当前状态下，再进行决策。

### 更复杂的未来

但在实际情况中，比刚才想象要复杂得多。

之前说过，未来是充满不确定性的，不确定性既包含现在选择的策略，也包含在环境之中。

也就是说，即使现在努力学习，也不能100%保证一定考得好成绩。即使有好成绩，父母也不一定会给更多零用钱。但即使挂科了，老师也不一定大发雷霆。

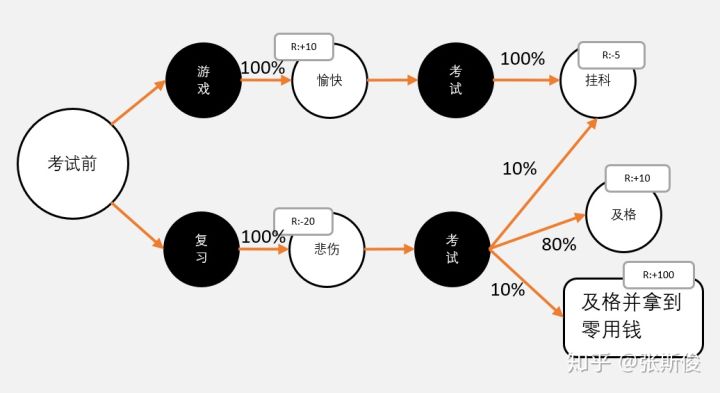
嗯，好吧。那看上去还是应该及时行乐，选择打游戏！

嗯，学渣永远是学渣，而你的学霸朋友（如果有的话），会先让你算一下：

把当前状况再理一下：

1. 10天后考试，玩游戏1天，心情+1；复习1天，心情-2。10天后，玩游戏心情+10，复习心情-20。

1. 不复习，100%挂科，被老师怒吼：-5点心情
2. 复习，10%挂科，同样被老师怒吼：-5点心情；80%不挂科，努力终于有回报：+10点心情；10%不挂科，且得到父母的零用钱 心情暴击+200点。



学霸深情地对学渣说：我知道你对概率和期望不太熟悉，那么现在就向某个吊车尾的一样，用影分身大法吧！

假设，有100个你自己，所有你都在玩游戏，并且所有的你都挂科了。 如果选择不复习。那么100个你一共可以获得多少奖励？

学渣：这我会！ 10天都在玩，那么玩游戏心情就+10，100个我就是1000点心情！ 必定会挂科了-5，100个分身就是-500点。所以最终+500点心情，那么就加500点心情！果然还是打游戏比较划算。

学霸： 你别兴奋呀，那如果复习呢？ 学渣： 10个分身挂科，这些分身复习了，-20心情；复习了还挂科，-5心情；一共-250心情值。 80个分身不挂科，这些分身复习了，-20心情；但不挂科，+10，最终还是只有-10点心情，80个我最终-800。 还有10个分身不挂科，而且获得父母的零用钱，因为复习-20心情，最终+1800点。



哇！比去玩的还多！我必须马上学习！

看完上面的故事，学霸可能会觉得麻烦。那不就是概率求期望么，真的要这么麻烦吗？

但在实际运用中，大多时候并不知道真实概率是多少。以上的概率都是自己估算，没有经过验证的。

在强化学习中，为了获得概率，将会不断地让智能体重复，或者让多个智能体进行试验以获得数据。

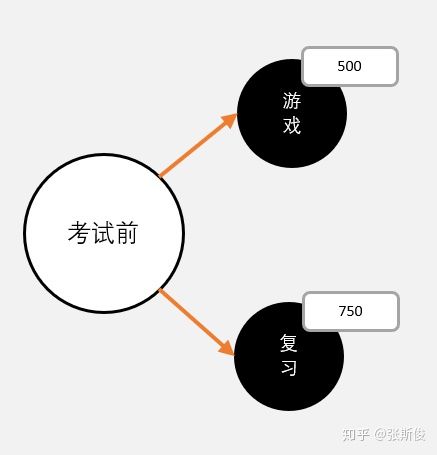
当然，100个影分身是很费查克拉的，在计算的时候也费时间。怎么在最少消耗里预估出合理的概率呢？这些知识将会在后面具体算法为大家讲解。

### Q和V的意义

所以在做决策的时候，需要把眼光放远点，把未来的价值换到当前，才能做出选择。

为了方便，希望可以有一种方法衡量做出每种选择的价值。这样，只要看一下标记，以后的事情也不用理了，选择那个动作价值更大，就选那个动作就可以了。

也就是说，让复习和游戏都有一个标记，这个标记描述了这个动作的价值： 游戏 +500;复习 +750



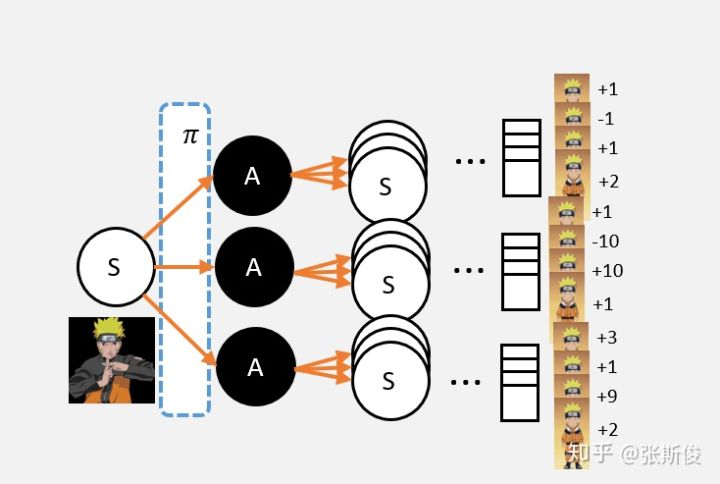
当然，也可以把这个标记标在状态上。为了方便沟通，这样定义 :评估**动作**的价值，称为**Q值**：它代表了智能体选择这个动作后，一直到最终状态**奖励总和**的**期望**；评估**状态**的价值，称为**V值**：它代表了智能体在这个状态下，一直到最终状态的**奖励总和**的**期望**。

价值越高，表示我从**当前状态**到**最终状态**能获得的**平均奖励**将会越高。因为智能体的目标是获取尽可能多的奖励，所以智能体在当前状态，只需要选择价值高的动作就可以了。

对于Q值和V值的定义非常非常重要。对QV的清晰理解，是理解强化学习中几乎所有算法的基础。 为此，不妨再用“影分身”大法再梳理一遍。

### V值的定义

上面的定义理解起来好难，用“影分身”大法，理解起来就容易多了

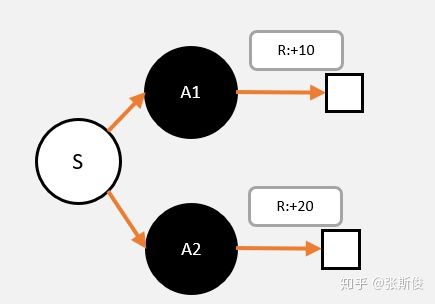


假设现在需要求某状态S的V值，那么可以这样：

1. 从S点出发，并影分身出若干个自己;
2. 每个分身按照当前的**策略** 选择行为;
3. 每个分身一直走到最终状态，并计算一路上获得的所有**奖励总和**;
4. 计算每个影分身获得的**平均值**,这个平均值就是要求的V值。

用大白话总结就是：从某个状态，按照策略 ，走到最终状态很多很多次；最终获得奖励总和的平均值，就是V值。

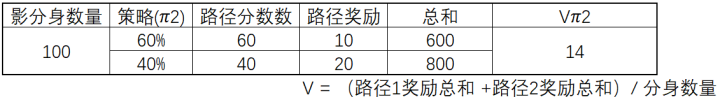
**【敲黑板】** 1. 从V值的计算，可以知道，V值代表了这个状态的今后能获得奖励的期望。从这个状态出发，到达最终状态，平均而言能拿到多少奖励,所以能够轻易地比较出两个状态的价值。 2. V值跟选择的策略有很大的关系。 看这样一个简化的例子，从S出发，只有两种选择，A1，A2；从A1，A2只有一条路径到最终状态，获得总奖励分别为10和20.



现在假设策略 采用平均策略[A1:50%,A2:50%]，根据用影分身(如果是学霸直接求期望)，那么可以求得V值为15



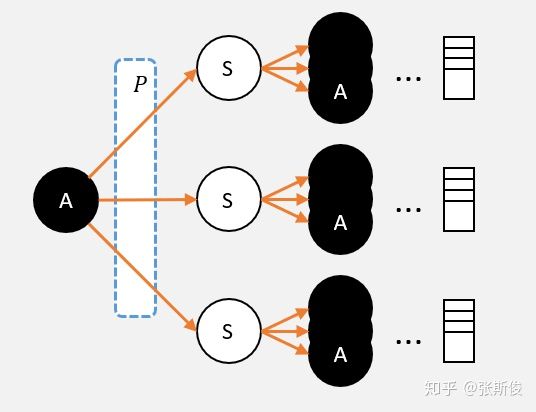
现在改变策略[A1:60%,A2:40%]，那么可以求得V值为14，变少了！



所以大家看到，**V值是会根据不同的策略有所变化的！**

### Q值的定义

如果大家已经了解V值的定义，那么理解Q值也不会有什么困难。Q值和V值的概念是一致的，都是衡量在马尔科夫树上某一个节点的价值。只不过V值衡量的是状态节点的价值，而Q值衡量的是动作节点的价值。



和V值一样，也可以用影分身来理解Q值。

现在需要计算某个动作A(状态为S0)的Q值：

1. 从A节点出发，使用影分身之术；
2. 每个影分身走到最终状态,并记录所获得的奖励；
3. 求取所有影分身获得奖励的平均值，这个平均值就是需要求的Q值。

用大白话总结就是：从动作A，走到最终状态很多很多次；最终获得奖励总和的平均值，就是Q值。

**【敲黑板】** 与V值不同，Q值和策略并没有直接相关，而与环境的状态转移概率相关，而环境的状态转移概率是不变的。

### V值和Q值关系

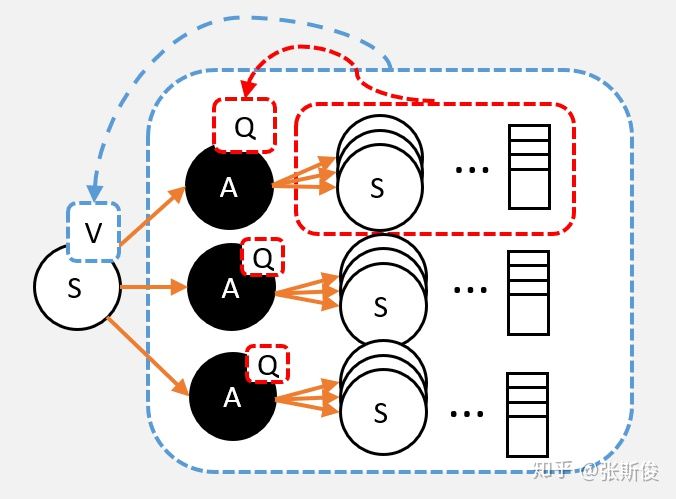
总结一下，从以上的定义，可以知道Q值和V值的意义相通的：

1. 都是马尔科夫树上的节点；
2. 价值评价的方式是一样的：从当前节点出发;一直走到最终节点;所有的奖励的期望值

所以，聪明的同学已经知道，其实Q和V之间是可以相互换算的。

### 从Q到V

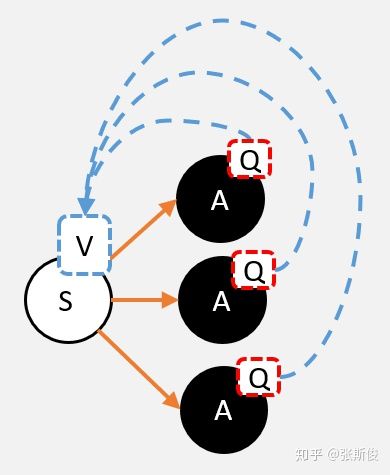
先来看看，怎样用Q值算V值。



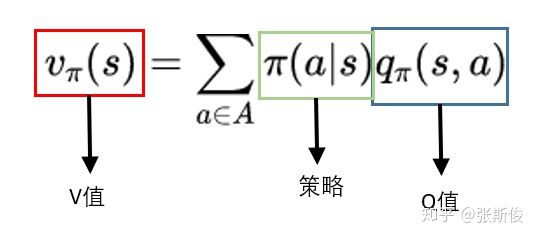
从定义出发，要求的V值，就是从状态S出发，到最终获取的所获得的奖励总和的期望值。也就是蓝色框部分。

S状态下有若干个动作，每个动作的Q值，就是从这个动作之后所获得的奖励总和的期望值。也就是红色框部分。

假设已经计算出每个动作的Q值，那么在计算V值的时候就不需要一直走到最终状态了，只需要走到动作节点，看一下每个动作节点的Q值，根据策略 ，计算Q的期望就是V值了。



更正式的公式如下：

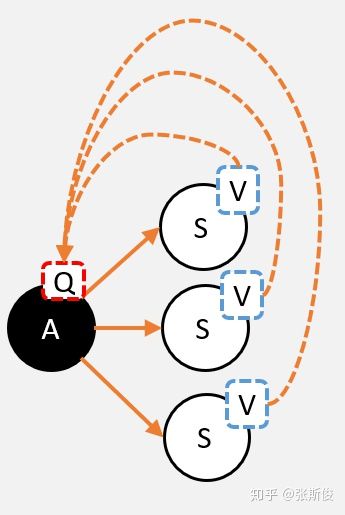


大白话就是：一个状态的V值，就是这个状态下的所有动作的Q值，在策略PI下的期望。

### 从V到Q

现在换个角度，看一下怎样从V换算成Q值。

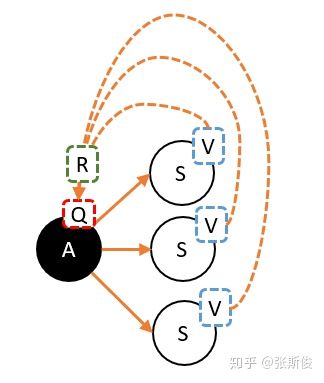
道理还是一样，就是: Q是V的期望！而且这里不需要关注策略，这里是环境的状态转移概率决定的。



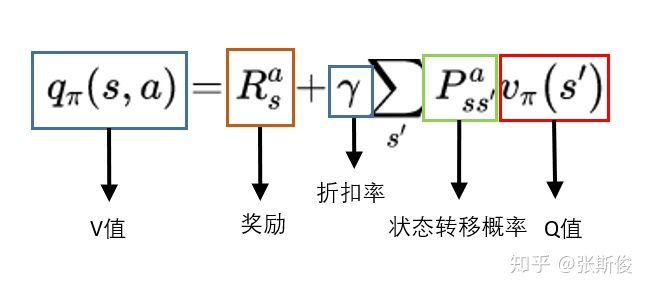
对，但还差点东西。

【敲黑板】

当选择A，并转移到新的状态时，就能获得奖励，必须把这个**奖励也算上！**



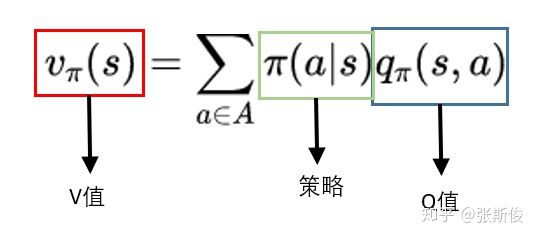
更正式的公式如下：

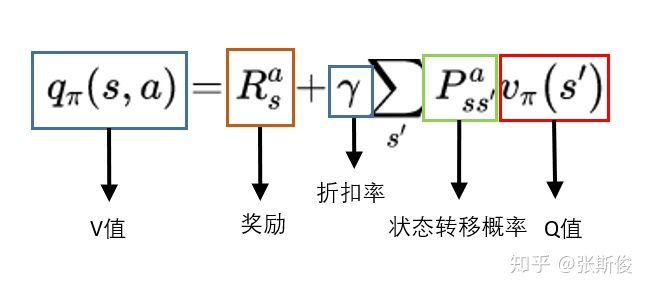


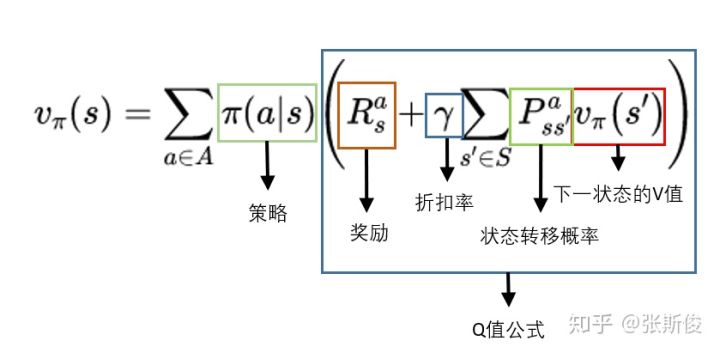
**折扣率** 在强化学习中，有某些参数是人为**主观**制定。这些参数并不能推导，但在实际应用中却能解决问题，所以称这些参数为**超参数**，而折扣率就是一个超参数。 与金融产品说的贴现率是类似的。计算Q值，目的就是把未来很多步奖励，折算到当前节点。但未来n步的奖励的10点奖励，与当前的10点奖励是否完全等价呢？未必。所以人为地给未来的奖励一定的折扣，例如：0.9,0.8，然后在计算到当前的Q值。

现在知道如何从V到Q，从Q到V了。但实际应用中，更多会从V到V。

但其实从V到V也是很简单的。把公式代进去就可以了。







### 总结

1. 比起记住公式，其实更应该注意Q值和V值的意义：他们就像一个路牌一样，告诉从马尔科夫树的一个节点出发，下面所有节点的收获的期望值。也就是假设从这个节点开始，走许多许多次，最终获取的奖励的平均值。
2. V就是子节点的Q的期望！但要注意V值和策略相关。
3. Q就是子节点的V的期望！但要注意，记得把R计算在内。

大家有没有发现，在这一节中，计算某一个节点的Q值和V值，需要许多次试验，取其中的平均值。但实际上，不但需要求一个节点的值，而是求所有节点的值。如果每一个节点都用同样的方法，消耗必然会很大。所以人们发明了许多方式去计算Q值和V值，基于价值计算的算法就是围绕Q和V展开的。这些在以后的文章将会说到。

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！