# 怎样直观理解Qlearning算法？[理论篇]

[](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。

## Qlearning算法(理论篇)

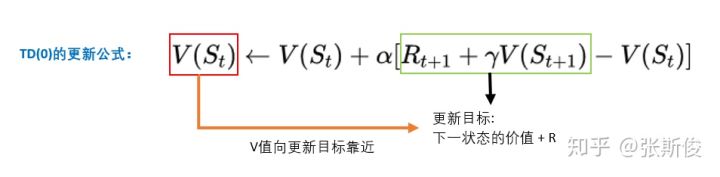
在第二章，将会研究多种RL基本算法，并去实现它。其中包括：Qlearning，DQN及其变种、然后会转到策略算法PG，然后会开始接触AC结构，例如AC、PPO、A3C、DPPO、TD3等较为高级的算法。

从Qlearning开始，如果你对python或者tensorflow不熟悉也不要紧。RLzoo为提供了参考代码，可以通过仔细分析大神的代码，一边学习理论，一边提高自己的代码能力。将会在后续的代码篇进行讲解。

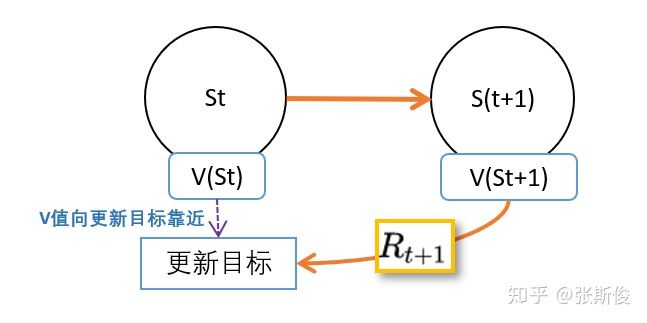
好，下面正式开始Qlearning算法。

### Qlearning

说过如何用TD(0)预估状态价值。



图解：



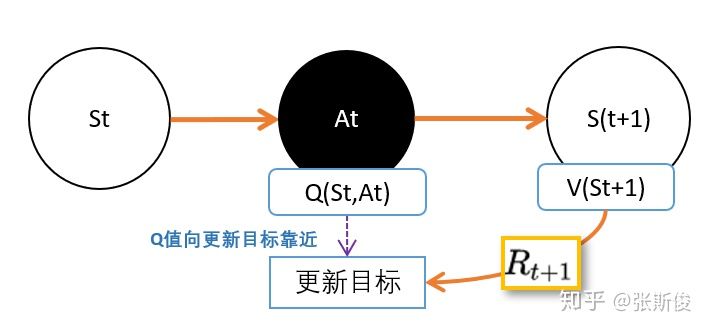
如果想直观了解TD的更新公式，可以看本专栏之前的文章。

[张斯俊：如何用时序差分TD估算状态V值？​](https://zhuanlan.zhihu.com/p/110132710" \t "/home/admin/文档\\x/_blank)

但大家可能会说，TD能够用在V值，那能不能用在计算Q值上呢？答案是肯定的。

### TD之于Q值估算

现在用上TD的思路。 在St，智能体根据策略pi，选择动作At，进入S(t+1)状态，并获得奖励R。 如果你之前对V和Q的理解足够深，那么不难理解下面这张图。

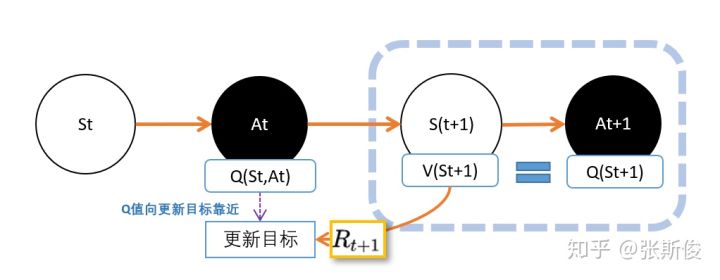


V(St+1)的意义是，在St+1到最终状态获得的奖励期望值。 Q(St,At)的意义是，在Q(St,At)到最终状态获得的奖励期望值。 所以可以把V(St+1)看成是下山途中的一个路牌，这个路牌告诉下山到底还有多远，然后加上R这一段路，就知道Q(St,At)离山脚有多长的路。

### 麻烦来了

但在实际操作的时候，会有一个问题。 在这里要估算两个东西，一个是V值，一个是Q值。

人们想出的办法就是，用下一个动作的Q值，代替V值。



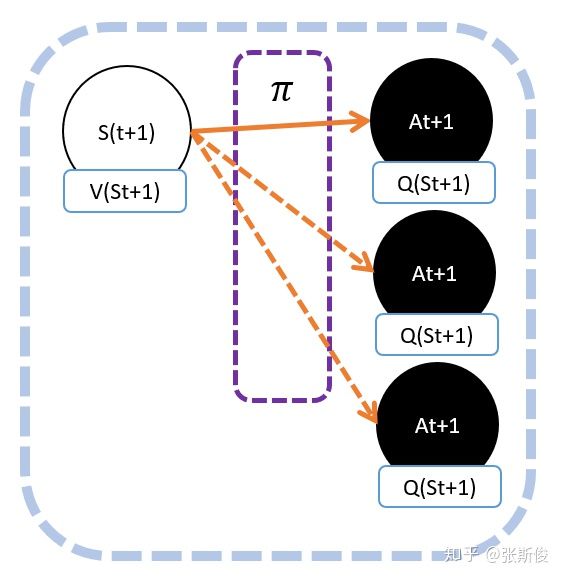
因为从状态St+1到动作At+1之间没有奖励反馈，所以直接用At+1的Q价值，代替St+1的V价值。 这样不就是可以了吗? 给三分钟时间大家想一下，理一下思路。

如果你答案是，对就是这样的。那么，可以看一下本专栏的另外一篇文章。

[张斯俊：如何理解强化学习中的Q值和V值？​](https://zhuanlan.zhihu.com/p/109498587" \t "/home/admin/文档\\x/_blank)

如果你答案是：好像有一点问题，但说不上为什么。那么我可以再说明一下。

关键是：马可洛夫链不是链，是树。



在St+1下，可能有很多动作At+1。不同动作的Q值自然是不同的。 所以Q(St+1,At+1)并不能等价于V(St+1)。

嘿嘿，虽然不相等，但不代表不能用其中一个来代表V(St+1)。

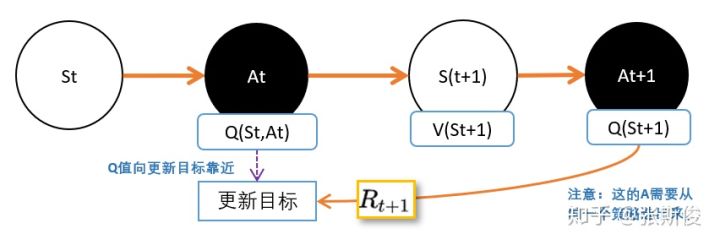
人们认为有个可能的动作产生的Q值能够一定程度代表V(St+1)。

1. 在相同策略下产生的动作At+1。这就是SARSA。
2. 选择能够产生最大Q值的动作At+1。这就是Qlearning。

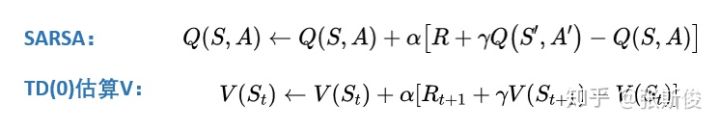
### SARSA

为什么SARSA用相同策略下产生的动作At+1是合理的。答案很简单，它管用。 其实在强化学习，虽然涉及很多数学，但它并不是严谨科学，它更像是工业，只要实际操作管用就行。

现在回到SARSA：



其实SARSA和上一篇说的TD估算V值几乎一模一样，只不过挪了一下，从V改成Q了。



注意，这里的At+1是在同一策略产生的。也就是说,St选At的策略和St+1选At+1是同一个策略。这也是SARSA和Qlearning的唯一区别。

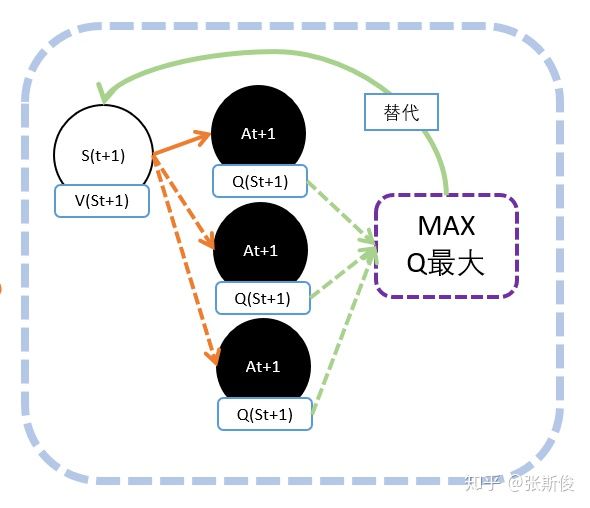
SARSA在迭代的时候，我们首先基于−贪婪法在当前状态选择一个动作，这样系统会转到一个新的状态′, 同时给我们一个即时奖励, 在新的状态′，我们会基于−贪婪法在状态S’′选择一个动作′，但是注意这时候我们并不执行这个动作′，只是用来更新的我们的价值函数.

Q-Learning基于状态S，用e−贪婪法选择到动作A, 然后执行动作A，得到奖励R，并进入状态S’′，此时，如果是SARSA，会继续基于状态′，用e−贪婪法选择A’′,然后来更新价值函数。但是Q-Learning则不同。对于Q-Learning，它基于状态S′, 没有使用e−贪婪法选择A′，而是使用贪婪法选择A′，也就是说，选择使Q(S′,a)最大的作为A′来更新价值函数。

对应到上图中就是在图下方的三个黑圆圈动作中选择一个使(S’′,a)最大的动作作为A’,此时选择的动作只会参与价值函数的更新，不会真正的执行。价值函数更新后，新的执行动作需要基于状态S′，用e−贪婪法重新选择得到。这一点也和SARSA稍有不同。对于SARSA，价值函数更新使用的A′会作为下一阶段开始时候的执行动作。

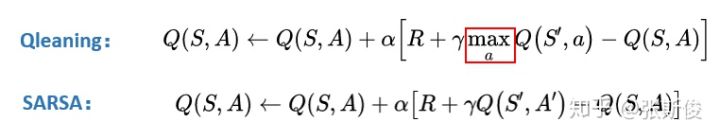
### Qlearning

Qlearning能够产生最大Q值的动作At+1的Q值作为V(St+1)的替代。



道理其实也很简单：因为需要寻着的是能获得**最多奖励**的动作，Q值就代表能够获得今后奖励的期望值。所以只会选择Q值最大的，也只有最大Q值能够代表V值。

所以大家可以对比一下Qleanring和SARSA，两者的差别仅仅在Qlearning中



就出现了鼎鼎大名的Qlearning算法。

### 总结

Qlearning和SARSA是多么鼎鼎大名，但直觉上理解还是很简单的。

现在来总结一下整个思路：

1. Qlearning和SARSA都是基于TD(0)的。不过在之前的介绍中，用TD(0)估算状态的V值。而Qlearning和SARSA估算的是动作的Q值。
2. Qlearning和SARSA的核心原理，是用下一个状态St+1的V值，估算Q值。
3. 既要估算Q值，又要估算V值会显得比较麻烦。所以用下一状态下的某一个动作的Q值，来代表St+1的V值。
4. Qlearning和SARSA唯一的不同，就是用什么动作的Q值替代St+1的V值。

SARSA : 选择的是在St同一个策略产生的动作。

Qlearning: 选择的是能够产生最大的Q值的动作。

相信大家对Qlearning已经有一定直觉上的认识。从这一篇开始，会有实战篇。实战篇采用的zooRL的示例代码。通过阅读代码，将会对如何运用算法有进一步理解。

在实际应用中，如果我们是在模拟环境中训练强化学习模型，推荐使用Q-Learning，如果是在线生产环境中训练模型，则推荐使用SARSA。

===========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！