基础Actor-Critic方法 (QAC)

在REINFORCE策略梯度法中，我们采用了蒙特卡洛方法，将策略梯度中的action value 用一条轨迹下的累计折扣回报来近似。然而，在策略梯度法中，策略网络的一次更新仅用一次采样的累计折扣回报，这会导致这个折扣回报与真实的action value之间的差距波动比较大，进而使得最终训练所得到的episode reward的抖动很大。这就是我们所说的蒙特卡洛方法的高方差问题。此外，REINFORCE算法只能在一个episode结束后才能进行更新，因此也就要求任务具有有限的步长。总结一下，REINFORCE主要有如下的几个问题：

1. 高方差，训练结果不稳定
2. 只能在回合结束才能进行训练，对一个回合的步长有要求。

本文首先介绍最基础的Actor-critic算法，即QAC算法。我们首先回顾一下上篇文章中我们所推导的REINFORCE的策略梯度：

由于我们说了REINFORCE的算法是基于蒙特卡洛的，因此是用一条轨迹的累计折扣回报近似代替了 。如果我们暂且先不考虑使用某条轨迹的折扣回报做近似，先还原成最初的 ，那么我们就可以写成更一般的策略梯度形式如下：

进一步，由于期望难以准确计算，因此采用SGD，得到：

即通过梯度上升方法，我们就可以得到策略网络参数的更新步骤：

通过上面这个式子，我们就可以来解释Actor-critic和REINCORCE算法的区别了：**如果策略梯度中的 是通过蒙特卡洛方法近似的，那么就是REINFORCE算法；如果 是通过TD-Learning近似的，那么就可以称为Actor-critic算法**。

看到我们要近似action value，我们可以立刻反应到，这不就是Sarsa算法吗！此外，由于目前都普遍采用深度神经网络，因此这里的Sarsa也是**基于值函数近似的Sarsa算法**。

因此，我们发现Actor-critic算法具有两个网络：Actor网络和Critic网络，Actor网络其实也就是策略网络，用来更新 ；Critic网络就是用来预测 的网络。因此，我们将这两个网络的更新步骤统一归纳起来：