**质量学习**

这里之所以有这里i是用来求期望值的，公式通过大数定律得来。

**预期值**

我们期望达到的值，相当于基线（baseline,可能是以平均值为原点），这是另一个网络模型。就是我们期望质量Q得到一个什么样的结果。好比你期望考60分，结果考了59分。上面的期望是对模型的整体期望，而这里的期望是对模型结果的期望

**优势函数**

网络模型其实就是一个拥有巨量维度的函数

**近似优势函数的函数**

由于训练多个网络模型，互相之间随着双方改变而改变，训练难度过大，所以找一个近似的函数，只训练一个网络

**实际公式**

**损失函数**

**策略损失**

期望值想要越来越大的话，就需要对应某个动作执行的概率越来越大，所以说是梯度上升。

这里期望值加了负号，所以就变成了越小，就需要梯度下降了。

**V（预期值）网络损失值**

预期与实际的差异

小v是指预期值的网络模型，西塔θ指的是策略损失值的某个网络模型

**熵**

用来描述混乱程度的值，熵越大表示其他动作的可能性都会增加一些，别太绝对。

**整体损失函数**