# Trust Region Policy Optimization

# 信赖域策略优化

John Schulman: 伯克利的博士，2016毕业，导师为Pieter Abbeel

Pieter Abbeel: 是伯克利的副教授，同时也是OpenAI的研究科学家，是机器人强化学习领域最有影响力的人之一。毕业于斯坦福大学。

## 策略梯度的缺点：

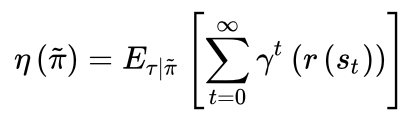
根据策略梯度方法，参数更新方程式为：

../../../../Desktop/屏幕快照%202018-02-26%20上午8.53.19.pn

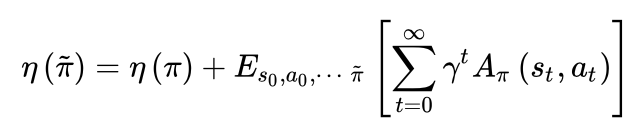
策略梯度算法的硬伤在于更新步长，当步长不合适时，容易导致越学越差，最后崩溃。所以，合适的步长对于强化学习非常关键。

所谓合适的步长实质当策略更新后，回报函数的值不能更差。如何选择步长，或者如何找到新的策略使得新的回报函数的值单调递增或单调不减，这是TRPO要解决的问题。

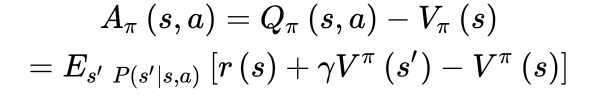
用表示一组状态-行为序列：s0, u0, … , sH, uH，则强化学习的回报函数为：



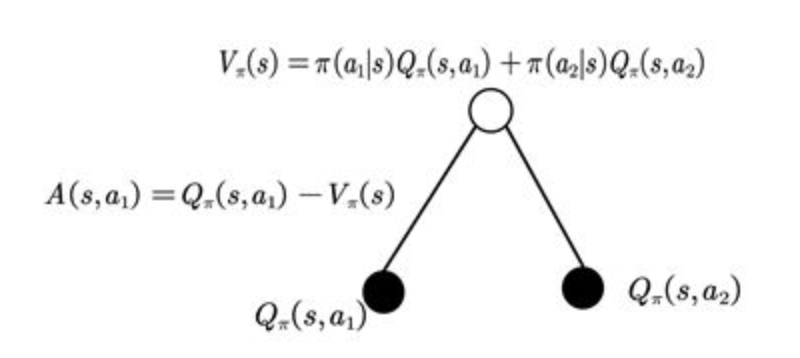
TRPO是找到新的策略，使得回报函数单调不减，想法为将新的策略所对应的回报函数分解成旧的策略所对应的回报函数+其他项。TRPO 的起点就是2002年Sham Kakade提出的这样一个等式：



其中：



称为**优势函数。**

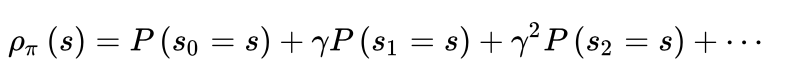


值函数V(s)为在该状态下所有可能动作所对应的动作值函数乘以采取该动作的概率的和，即时改状态下所有动作值函数关于动作概率的平均值。

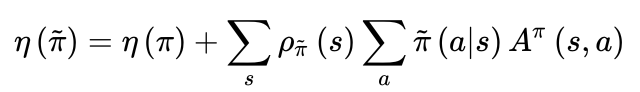
动作值函数Q(s,a)是单个动作所对应的值函数。

Q(s, a) – V(s)能评价当前动作值函数相对于平均值的大小。所以，这里的**优势指的是动作值函数相比于当前状态的值函数的优势。**如果优势函数大于零，则说明该动作比平均动作好，如果优势函数小于零，则说明当前动作还不如平均动作好。

定义：



则：

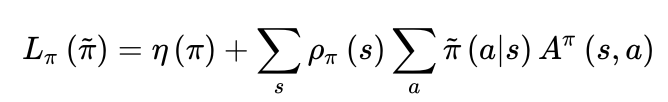


# ../../../../Desktop/屏幕快照%202018-02-26%20上午9.35.37.pn

### TRPO第一个技巧：

引入第一个技巧对状态分布进行处理。忽略状态分布的变化，依然采用旧的策略所对应的状态分布。当新旧参数很接近时，用旧的状态分布代替新的状态分布也是合理的。

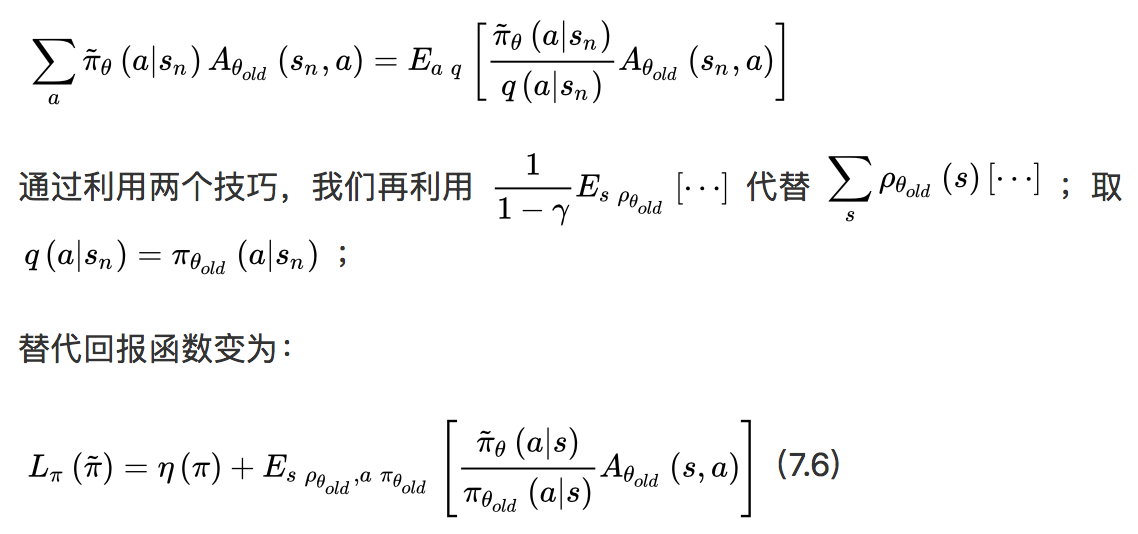
原代价函数变为：



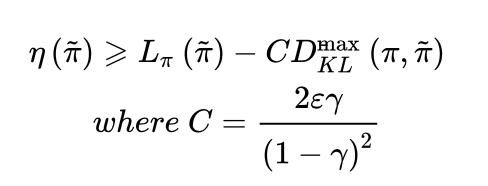
但其中第二项策略部分中的动作a是由新的策略生成，但此时新的策略是带未知参数的，无法用来产生动作。

### TRPO第二个技巧：

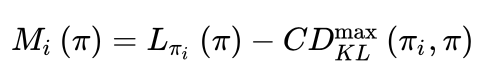
第二个技巧是利用重要性采样对动作分步进行处理。



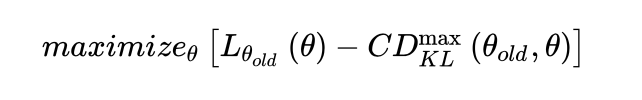
设备如第二个重量级的不等式：



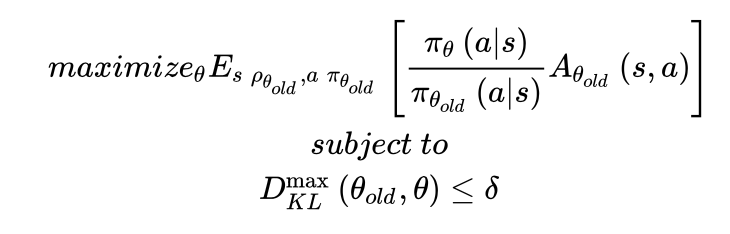
其中DKL() 是两个分布的KL散度。（相对熵，又称为KL散度，信息散度，相对熵是两个概率分布差别的非对称性的度量）该不等式给了的下界，定义为：



则使得Mi最大的新策略就是要更新的策略。该问题形式化为：



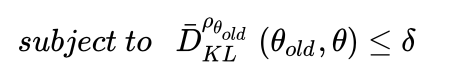
如果利用惩罚因子C则每次迭代步长很小，则问题转化为：



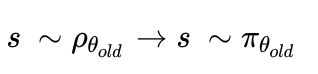
因为有无穷多的状态，因此约束条件DKL有无穷多个，问题不可解。

### TRPO第三个技巧

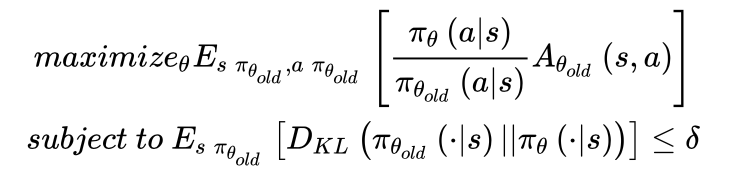
在约束条件中利用平均KL散度代替最大KL散度，即：



### TRPO第四个技巧



最终TRPO问题化简为：



接下来就是利用采样得到数据，然后求样本均值，解决优化问题即可。