# 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：

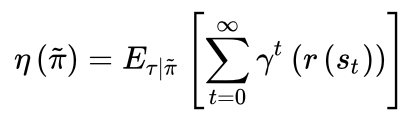
根据策略梯度方法，参数更新方程式为：

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

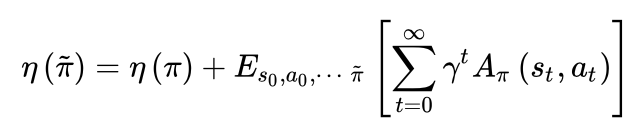
策略梯度算法的硬伤在于更新步长，当步长不合适时，容易导致越学越差，最后崩溃。所以，合适的步长对于强化学习非常关键。

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。

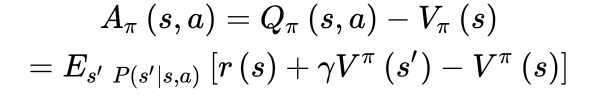
用表示一组状态-行为序列：s0, u0, … , sH, uH，则强化学习的回报函数为：



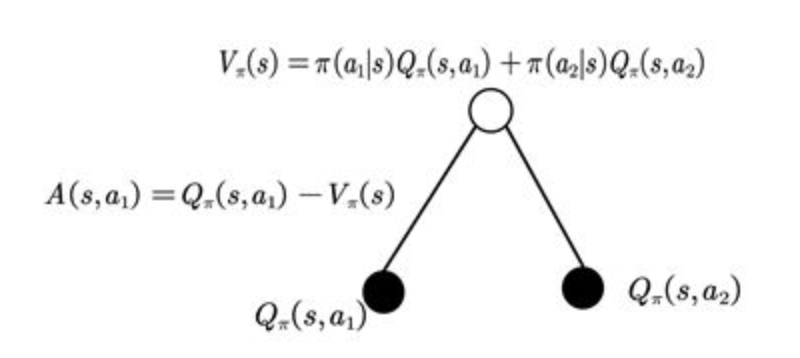
TRPO是找到新的策略，使得回报函数单调不减，想法为将新的策略所对应的回报函数分解成旧的策略所对应的回报函数+其他项。TRPO 的起点就是2002年Sham Kakade提出的这样一个等式：



其中：



称为**优势函数。**

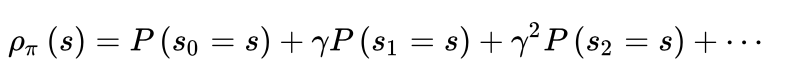


值函数V(s)为在该状态下所有可能动作所对应的动作值函数乘以采取该动作的概率的和，即时改状态下所有动作值函数关于动作概率的平均值。

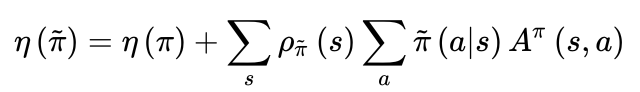
动作值函数Q(s,a)是单个动作所对应的值函数。

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：

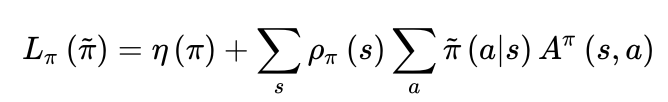


# ../../../../Desktop/屏幕快照%202018-02-26%20上午9.35.37.pn

### 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。

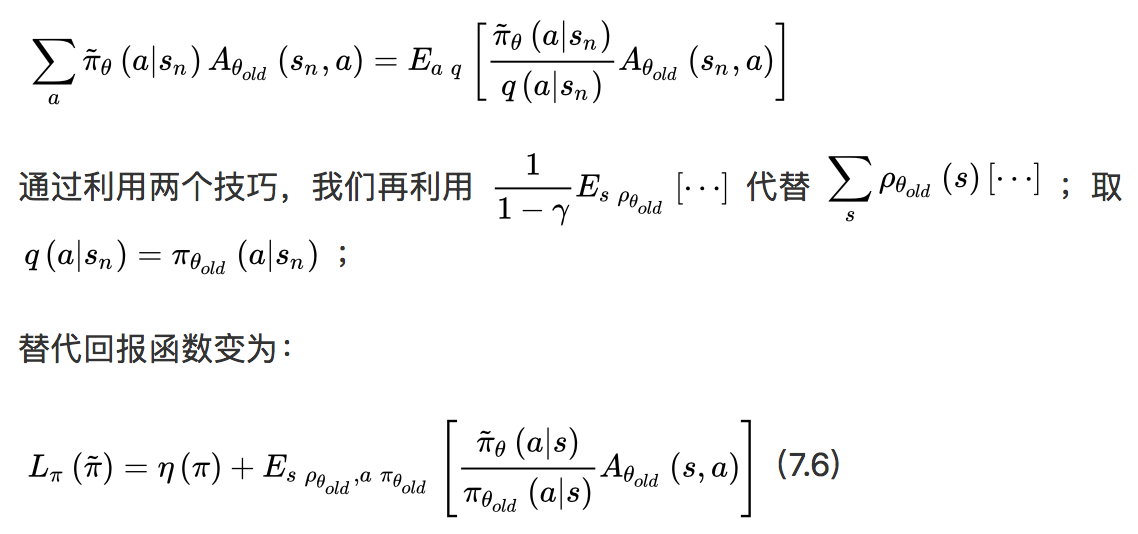
原代价函数变为：



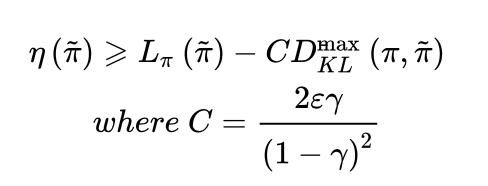
但其中第二项策略部分中的动作a是由新的策略生成，但此时新的策略是带未知参数的，无法用来产生动作。

### TRPO第二个技巧：

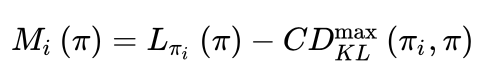
第二个技巧是利用重要性采样对动作分步进行处理。



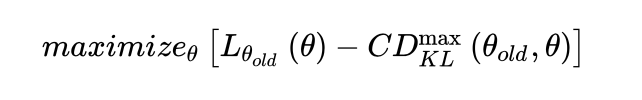
设备如第二个重量级的不等式：



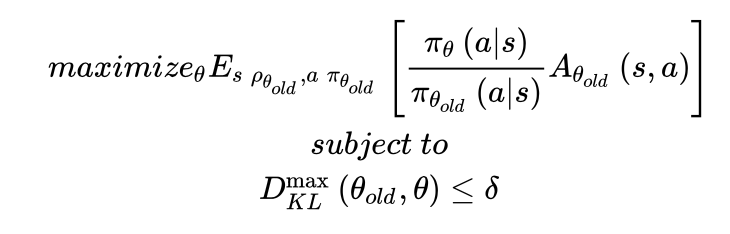
其中DKL() 是两个分布的KL散度。（相对熵，又称为KL散度，信息散度，相对熵是两个概率分布差别的非对称性的度量）该不等式给了的下界，定义为：



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：



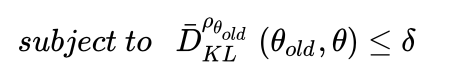
如果利用惩罚因子C则每次迭代步长很小，则问题转化为：



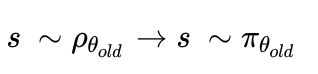
因为有无穷多的状态，因此约束条件DKL有无穷多个，问题不可解。

### TRPO第三个技巧

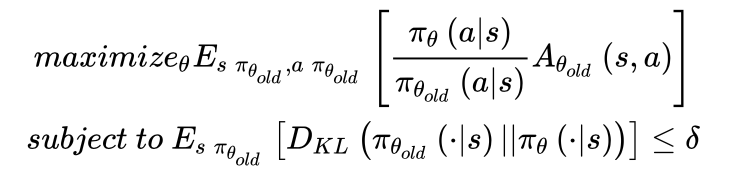
在约束条件中利用平均KL散度代替最大KL散度，即：



### 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。

## 原文：

# Trust Region Policy Optimization

加州大学伯克利分校电子工程与计算机科学系

2017.4.20

**Abstract**：我们描述了一个保证单调改进的优化策略的迭代过程，通过对理论上合理的过程进行几次近似，我们开发了一种称为信赖域策略优化的使用算法。该算法近似于自然策略梯度方法，对于优化大型非线性策略（例如神经网络）非常有效。我们的实验展示了它在各种任务上的鲁棒性：学习模拟机器人游泳，跳跃和行走步态；并使用屏幕图像作为输入来玩Atrai游戏。虽然它的近似值偏离了理论，但TRPO倾向于提供单调的改进，对超参数的调整很少。

1. **Introduction**

大多数策略优化算法可以分为三大类：1）策略迭代方法，在估计当前政策的价值函数和改进政策之间交替进行（2005）；2）策略梯度方法，使用从样本轨迹中获得的期望受益（总回报）梯度的估计值（2008）；3）无衍生优化算法，如交叉熵算法CEM和协方差矩阵自适应法CMA，将回报作为黑箱函数来处理，以便根据策略参数进行优化（2006）。

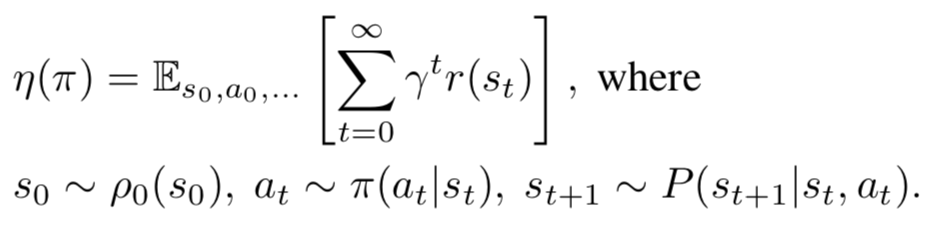
一般的无衍生随机优化法如CEM和CMA在很多问题上是首选，因为它们易于理解和实施且取得了良好的结果。例如，虽然俄罗斯方块是近似动态规划（ADP）的经典基准问题，但随机化方法难以胜任。对于连续控制问题像CMA这样的方法已经成功地学习了控制策略，以满足具有低维度参数化的手工策略类的运动等具有挑战性的任务。由于基于梯度的优化算法比无梯度方法具有更好的样本复杂性保证，所以ADP和基于梯度的方法一致击败无梯度随机搜索的能力并不令人满意。连续的基于梯度的优化在学习具有大量参数的监督学习任务的函数逼近方面非常成功，并且将其成功扩展到强化学习可以有效地训练复杂而强大的策略。

在本文中，我们首先证明，最小化某个特定的目标函数可以保证不平凡的步长大小的政策改进。然后，我们对理论上对齐的算法进行一系列的近似，得到一个实用的算法，我们称之为信赖域策略优化（TRPO）。我们描述了这种算法的两种变体：首先，单路径方法，可以在无模型设置中应用; 其次，vine法需要系统恢复到特定状态，这通常只能在模拟中实现。这些算法是可扩展的，可以优化带有数以万计参数的非线性策略，这对以前的无模型策略搜索提出了重大挑战。 在我们的实验中，我们展示了相同的TRPO方法可以学习游泳，跳跃和行走的复杂政策，以及直接从原始图像中玩Atari游戏。

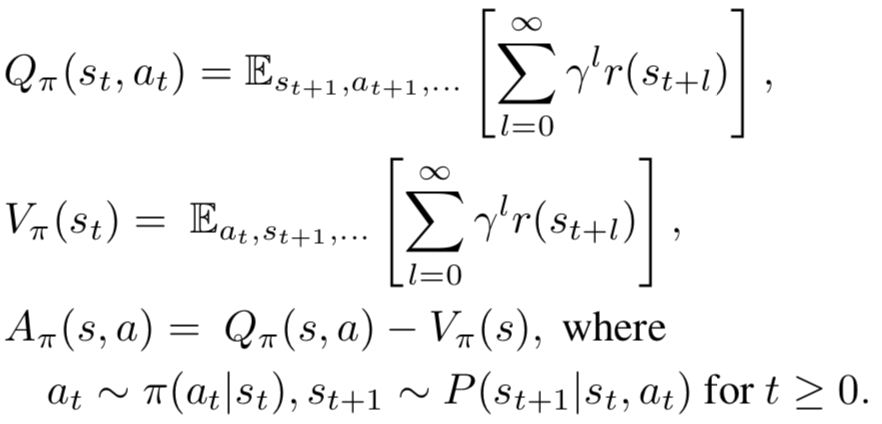
1. **Preliminaries**

考虑由(S, A, P, r, ρ0, γ)定义的MDP过程，其中S是有限状态集合，A是行为的有限行动的集合，P： S×A×S→R为转移概率分布。r：S→R为奖励函数，ρ0：S→R为初始状态s0的分布，γ∈（0,1）为折扣因子。

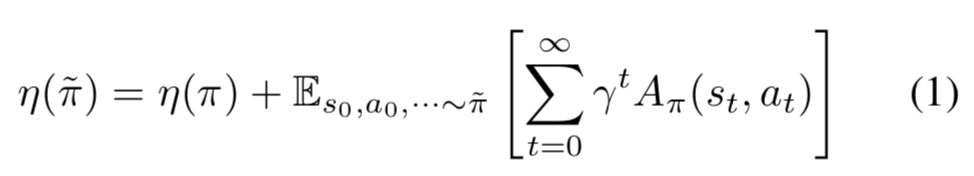
令π表示一个随机策略π：S×A→[0,1]，令η（π）表示其预期折扣奖励：



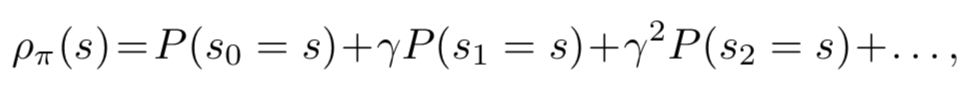
我们将使用状态-行动值函数Qπ，值函数Vπ和优势函数Aπ的以下标准定义：



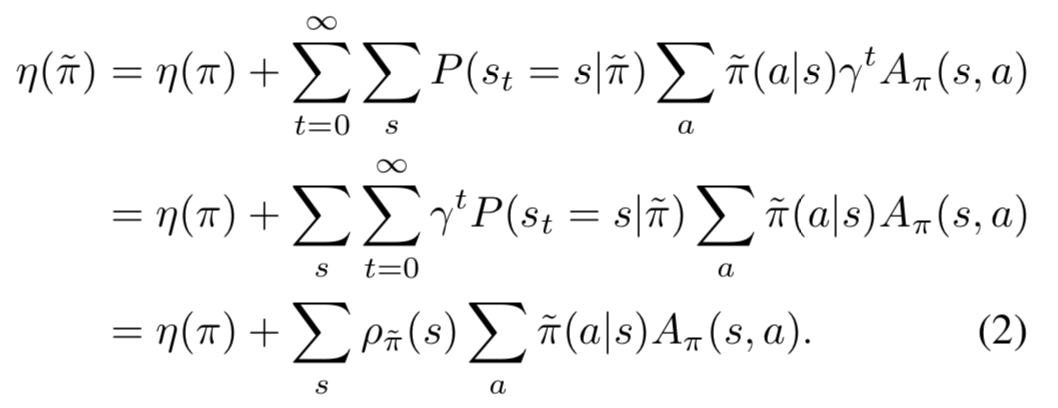
π ̃表示另一个政策的预期回报，其优势超过π，在时间步长上累积:



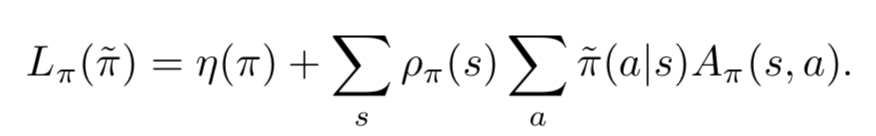
设ρπ为（非标准化）折扣访问频率:



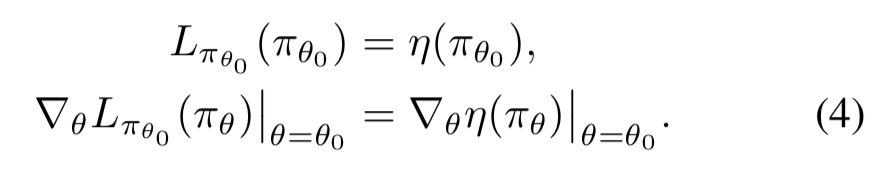
其中s0〜ρ0，动作根据π来选择。 我们可以用方程式（1）重写状态而不是时间步长：



这个方程意味着任何政策更新π→π ̃在每个状态s上具有非负的预期优势，即一个π ̃（a | s）Aπ（s，a）≥0，可以保证提高政策绩效η，或者在所有地方的预期优势都为零的情况下保持不变。这意味着经典结果是，如果至少有一个具有积极优势的状态动作对，那么使用确定性策略pi ̃（s）= argmaxaAπ（s，a）的精确策略迭代执行的更新改进了策略值和非零状态访问概率，否则算法已经收敛到最优策略。然而，在近似的情况下，由于估计和逼近误差，通常将不可避免地存在一些状态，其中预期的优势是负的，ρπ（s）对π的复数依赖性使方程（2）难以直接优化。 相反，我们引入下面的局部近似η：

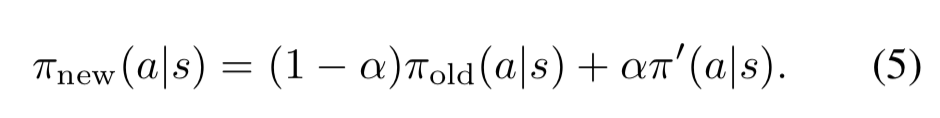


注意到L使用概率ρπ而不是ρπ ̃，忽略了策略变化导致的概率的变化。 然而，如果我们有一个参数化策略πθ，其中πθ（a | s）是参数向量θ的可微函数，则L将η与一阶匹配。也就是说，对于任何参数值θ0:

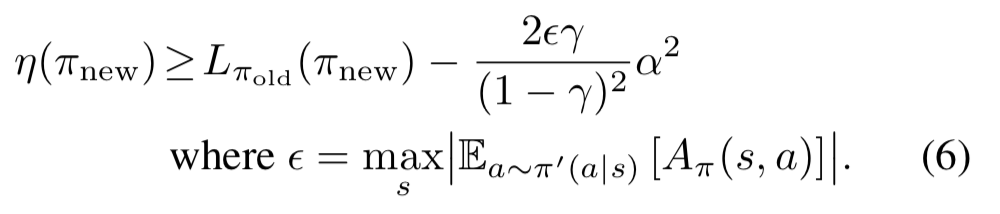


(4)意味着足够小的步长πθ0→π ̃可以改善η，但是对于步长取多大并没有指导。

为了解决这个问题，提出了一中称为保守策略调整的策略更新方法，可以为改进η提供明确的下限。为了定义保守的政策迭代更新，令πold表示当前的策略，π'= argmax π'L πold（π'）。新的策略πnew定义为：



下界为：



但是，请注意，迄今为止，此限制仅适用于由公式（5）生成的混合策略。 这个政策阶级在实践中是笨拙和限制性的，并且希望实际的政策更新方案适用于所有一般的随机政策类别。

1. **Monotonic Improvement Guarantee for General Stochastic Policies**

等式（6）适用于保守的策略迭代，意味着可以改进右侧的策略更新可以保证提高真实性能η。我们的主要理论结果是，方程（6）中的政策改进界限可以扩展到一般的随机政策，而不仅仅是混合政策，用π和π之间的距离度量替换α，常数ε适当。 由于实践中很少使用混合策略，因此这一结果对于将改进保证延伸到实际问题至关重要。 我们使用的特定距离度量是总变差发散度，