置信域算法：目标：Find $\theta^* = \arg\max_\theta J(\theta)$

① 令 $N(\theta_{old})$ 为 θ_{old} 的邻域，即

$$N(\theta_{old}) = \{\theta \mid \|\theta - \theta_{old}\|_2 \le \Delta\}$$

② 如果有一个function $L(\theta|\theta_{old})$ 可在 $N(\theta_{old})$ 中很好地近似 $J(\theta)$，则称 $N(\theta_{old})$ 称为置信域 trust region.

置信域算法重复过程：

① 近似：给定 θ_{old} 构建 $L(\theta|\theta_{old}) \Rightarrow$ 是在 $N(\theta_{old})$ 中 $J(\theta)$ 的近似

② 最大化：在置信域中，寻找 θ_{new}

$$\theta_{new} = \arg\max_{\theta \in N(\theta_{old})} L(\theta|\theta_{old})$$

$$V_\pi(s) = \sum_a \pi(a|s;\theta) Q_\pi(s,a)$$

$$= \sum_a \pi(a|s;\theta_{old}) \frac{\pi(a|s;\theta)}{\pi(a|s;\theta_{old})} Q_\pi(s,a)$$

$$= E_{A\sim\pi(\cdot|s;\theta_{old})}\left[\frac{\pi(A|s;\theta)}{\pi(A|s;\theta_{old})} Q_\pi(s,A)\right]$$

$$J(\theta) = E_S[V_\pi(S)]$$

$$= E_S\left[E_{A\sim\pi(\cdot|s;\theta_{old})}\left[\frac{\pi(A|s;\theta)}{\pi(A|s;\theta_{old})} Q_\pi(s,A)\right]\right]$$

TRPO：sample efficiency 且更加 robust.

但PG算法不够采样高效和鲁棒.

对这个公式进行蒙特卡洛近似

根据 θ_{old} 生成一条轨迹 $s_1, a_1, r_1, s_2, a_2, r_2 \cdots\cdots s_n, a_n, r_n$

用蒙特卡洛近似

$$L(\theta|\theta_{old}) = \frac{1}{n}\sum_{k=1}^{n} \frac{\pi(a_k|s_k;\theta)}{\pi(a_k|s_k;\theta_{old})} \cdot Q_\pi(s_k,a_k)$$

① 近似 即用 $L(\theta|\theta_{old})$ 来近似 $J(\theta)$

但是：不知道 $Q_\pi(s_k, a_k)$

所以要对 Q_π 进行近似。

一个episode的奖励为 $r_1, r_2, \cdots r_n$。

第i时刻的回报为 $u_i = r_i + \gamma r_{i+1} + \gamma^2 r_{i+2} + \cdots + \gamma^{n-i} r_n$

↓

大约为 $Q(s_i, a_i)$ 的值

用蒙特卡洛近似公式

$$\tilde{L}(\theta|\theta_{old}) = \frac{1}{n}\sum_{k=1}^{n} \frac{\pi(a_k|s_k;\theta)}{\pi(a_k|s_k;\theta_{old})} u_i$$

② 最大化：

邻域的不同度量：① $\|\theta - \theta_{old}\| \le \Delta$

② $\frac{1}{n}\sum_{k=1}^{n} KL[\pi(\cdot|s_i;\theta_{old}) \| \pi(\cdot|s_i;\theta)] \le \Delta$ KL散度