SAC Policy Gradient

原文中提到了两种解法：likelihood-ratio gradient estimator和reparameterization，接下来我们简单推导一下这两个解法。
我们优化的函数是：

$$J(\phi) = E_{a\sim\pi_\phi}[log\pi_\phi(a \mid s) - Q_\theta(s,a)]$$

Gradient estimator

$$\begin{aligned}
\nabla_\phi J(\phi) =& \nabla_\phi[\sum_a \pi_\phi(a \mid s)(log\pi_\phi(a \mid s) - Q_\theta(s,a))] \\
=& \sum_a [\nabla_\phi \pi_\phi(a \mid s)(log\pi_\phi(a \mid s) - Q_\theta(s,a)) + \pi_\phi(a \mid s)\nabla_\phi log\pi_\phi(a \mid s)] \\
=& E_{a\sim\pi_\phi}[\nabla_\phi log\pi_\phi(a \mid s)(log\pi_\phi(a \mid s) - Q_\theta(s,a) + 1)]
\end{aligned}$$

注意：这里是把a当作常数处理，然后直接求出目标函数的导数，由于是一个期望形式，我们可以直接用采样的方式来近似估计这个梯度值。

Reparameterization

令$\epsilon \sim N(0,1), \mu = \pi_{\phi_\mu}(s), \delta = \pi_{\phi_\delta}(s)$,可以定义：

$$\begin{aligned}
a^1_\phi =& \delta * \epsilon + \mu \\
a^2_\phi =& tanh(a^1_\phi) \\
a^3_\phi =& scale * a^2_\phi + bias \\
J(\phi) =& E_{\epsilon\sim N}[log\pi_\phi(a^3_\phi \mid s) - Q_\theta(s, a^3_\phi)]
\end{aligned}$$

注意：如果不把a当作常数处理，而是把他转化为一个关于ϕ的函数，这个函数是一般是通过一个正态分布输入噪声构造的。这里我们做的事情实际上是让$J(\phi)$可以通过采样来近似，这就需要让分布不依赖于ϕ，所以我们用了一个$\epsilon \sim N(0,1)$。