不用重参数化技巧的SAC

不能直接求梯度的原因

SAC算法的目标函数：

$$J_\pi(\phi)=E_{s_t\sim D,a_t\sim\pi_\phi}[\log_{\pi_\phi}(a_t|s_t)-\frac{1}{\alpha}Q_\theta(s_t,a_t)] \tag{1}$$

如果按照机器学习通常的做法，就是将期望里的项目（即中括号里的项）对ϕ进行求导，然后对采样的样本取平均值。但这样是不对的，因为采样的过程$a_t\sim\pi_\phi$是和参数ϕ相关的，有待优化的参数，因此前面那种方法求的导不等于原始目标函数对ϕ的梯度。

注：策略梯度方法从来都不是直接能拿一个目标函数去自动求梯度，都是先用公式推出一个策略梯度来

使用REINFORCE的方法对策略梯度进行推导

$$\begin{aligned}
\nabla J_\pi(\phi) &= E_{s_t\sim D}[\nabla_\phi\sum_{a_t}\pi_\phi(a_t|s_t)(\log_{\pi_\phi}(a_t|s_t)-\frac{1}{\alpha}Q_\theta(s_t,a_t))] && (2)\\
&= E_{s_t\sim D}\Big[\sum_{a_t}[\nabla_\phi\pi_\phi(a_t|s_t)\cdot(\log_{\pi_\phi}(a_t|s_t)-\frac{1}{\alpha}Q_\theta(s_t,a_t))+\pi_\phi(a_t|s_t)\cdot\frac{\nabla_\phi\pi_\phi(a_t|s_t)}{\pi_\phi(a_t|s_t)}]\Big] && (3)\\
&= E_{s_t\sim D}\Big[\sum_{a_t}\nabla_\phi\pi_\phi(a_t|s_t)\cdot(\log_{\pi_\phi}(a_t|s_t)-\frac{1}{\alpha}Q_\theta(s_t,a_t)+1)\Big] && (4)\\
&\text{为了利用}a_t\sim\pi_\phi\text{采样，再改写} && (5)\\
&= E_{s_t\sim D}\Big[\sum_{a_t}\pi_\phi(a_t|s_t)\frac{\nabla_\phi\pi_\phi(a_t|s_t)}{\pi_\phi(a_t|s_t)}\cdot(\log_{\pi_\phi}(a_t|s_t)-\frac{1}{\alpha}Q_\theta(s_t,a_t)+1)\Big] && (6)\\
&= E_{s_t\sim D,a_t\sim\pi_\phi}\Big[\nabla\log\pi_\phi(a_t|s_t)\cdot(\log_{\pi_\phi}(a_t|s_t)-\frac{1}{\alpha}Q_\theta(s_t,a_t)+1)\Big] && (7)
\end{aligned}$$

上式中的+1可以去掉，因为$\sum_{a_t}\nabla_\phi\pi_\phi(a_t|s_t)=\nabla_\phi\sum_{a_t}\pi_\phi(a_t|s_t)=\nabla_\phi 1=0$

也可以将+1换成其他和a无关的量，不会影响梯度，相当于带baseline的梯度算法，例如：

$$\nabla J_\pi(\phi)=E_{s_t\sim D,a_t\sim\pi_\phi}\Big[\nabla\log\pi_\phi(a_t|s_t)\cdot(\log_{\pi_\phi}(a_t|s_t)-\frac{1}{\alpha}Q_\theta(s_t,a_t)-\frac{1}{\alpha}V_\theta(s_t))\Big] \tag{8}$$

经实验，重参数化和非重参数化的方法从性能和运算开销上基本一致，但剪不剪去上边的baseline效果差不少。

使用重参数技巧的SAC

使用重参数化技巧的原因

SAC的actor以state为输入，以μ和σ为输出，即根据状态s建立高斯分布，然后随机采样动作$a \sim \mathcal{N}(\mu, \sigma^2)$，因为是随机采样，导致不能计算梯度$\partial a / \partial \mu$和$\partial a / \partial \sigma$，所以采用重参数化技巧，即先采样一个常数$\xi \sim \mathcal{N}(\mu, \sigma^2)$，然后$a = \mu + \xi \cdot \sigma$，则a仍然服从分布$\mathcal{N}(\mu, \sigma^2)$，且对$\mu$和$\sigma$可导。

另外，由于式8中期望的a_t是从高斯策略$\mathcal{N}(\mu_\phi(s_t), \sigma_\phi(s_t))$中采样的，如果这个高斯分布很扁平的话，会导致梯度估计值方差很大，不利于学习。

因此，可以借鉴VAE中的Reparameterization Trick来减少方差，使学习更稳定。即：

$$a_t = \mu_\phi(s_t) + \epsilon \sigma_\phi(s_t) \tag{9}$$