13 - SAC 算法

最大熵强化学习

优势		劣势
稳定性且可重复性	确定性策略	缺乏探索性
更加灵活	随机性策略	不稳定

从这我们不难看出随机性策略对比于确定性策略存在碾压性的优势。在最大熵学习中，我们在最大化累计奖励的策略基础上增加了一个信息熵的约束，即

$$\pi^*_{\text{MaxEnt}} = \arg\max_{\pi} \sum_t \mathbb{E}_{(s_t,a_t)\sim p_\pi}[\gamma^t(r(s_t,a_t) + \alpha\mathcal{H}(\pi(\cdot|s_t)))]$$

其中 α 只一个超参，政委温度因子，用于平衡积累奖励和策略熵的比重。这里 $\mathcal{H}(\pi(\cdot|s_t))$ 就是策略的信息熵

$$\mathcal{H}(\pi(\cdot|s_t)) = -\sum_{a_t} \pi(a_t|s_t)\log\pi(a_t|s_t)$$

它表示了随机策略 $\pi(\cdot|s_t)$ 对应概率分布的随机程度，策略月随机，熵越大。

SAC 算法

V网络的目标函数定义

$$J_V(\Psi) = \mathbb{E}_{s_t\sim D}[\frac{1}{2}(V_\Psi(s_t) - [Q_\theta(s_t,a_t) - \log\pi_\phi(a_t|s_t)])^2]$$

其梯度为

$$\hat{\nabla}_\Psi J_V(\Psi) = \nabla_\Psi V_\Psi(s_t)(V_\Psi(s_t) - Q_\theta(s_t,a_t) + \log\pi_\phi(a_t|s_t))$$

而，Soft Q函数的目标函数为

$$J_Q(\theta) = \mathbb{E}_{(s_t,a_t)\sim\mathcal{D}}[\frac{1}{2}(Q_\theta(s_t,a_t) - \hat{Q}_\theta(s_t,a_t))^2]$$

其梯度为

$$\hat{\nabla}_\theta J_Q(\theta) = \nabla_\theta Q_\theta(a_t,s_t)(Q_\theta(s_t,a_t) - r(s_t,a_t) - \gamma V_\Psi(s_{t+1}))$$

策略的目标函数为

$$J_\pi(\phi) = \mathbb{E}_{s_t\sim\mathcal{D},\epsilon_t\sim\mathcal{N}}[\log\pi_\phi(f_\phi(\epsilon_t;s_t)|s_t) - Q_\theta(s_t,f_\phi(\epsilon_t;s_t))]$$

其梯度为

$$\hat{\nabla}_\phi J_\pi(\phi) = \nabla_\phi\log\pi_\phi(a_t|s_t) + (\nabla_{a_t}]log\pi_\phi(a_t|s_t) - \nabla_{s_t}Q(s_t,a_t))\nabla_\phi f_\phi(\epsilon_t;s_t)$$