11 - DDPG算法

DPG方法

深度确定性策略梯度算法（ deep deterministic policy gradient，DDPG），是一种确定性的策略梯度算法。

首先我们知道 DQN 算法的一个主要缺点就是不能用于连续动作空间，这是因为在 DQN 算法中动作是通过贪心策略或者说 argmax 的方式来从 Q 函数间接得到，这里 Q 函数就相当于 DDPG 算法中的 Critic。

所以，DDPG 算法并没有做真正意义上的梯度更新，只是在寻找最大值，本质上还是 DQN 算法的思路。因此 DDPG 算法中 Critic 结构会同时包含状态和动作输入，而不是 Actor-Critic 算法中那样只包含状态.

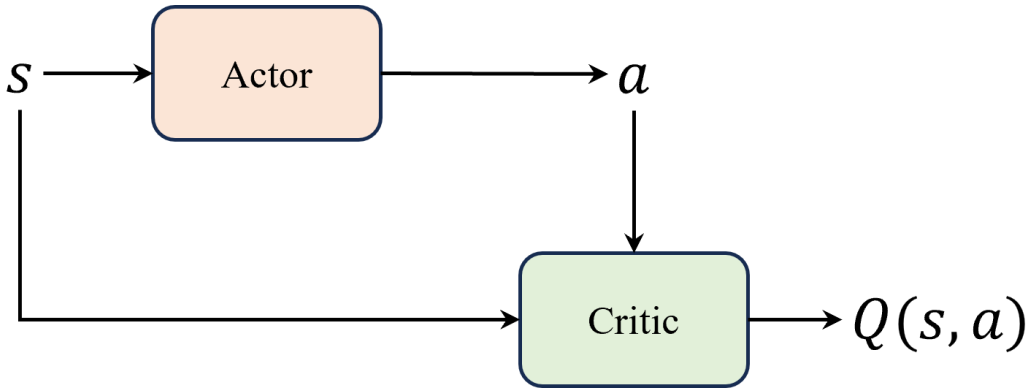

图 11-2 DDPG 网络结构

这里相当于是把 DQN 算法中 ϵ-greedy 策略函数部分换成了 Actor 。注意 Actor 网络 $\mu_\theta(s)$ 与输出概率分布的随机性策略（ stochastic policy ）不同，输出的是一个值，因此也叫做确定性策略（ deterministic policy ）。效仿策略梯度的推导，我们也可以推导出 DPG 算法的目标函数

$$\nabla_\theta J(\theta) \approx \mathbb{E}_{st\sim\rho^\beta}[\nabla_a Q(s_t,a)|_{a=\mu_\theta(s_t)}\nabla_\theta \mu_\theta(s_t)]$$

其中 ρ^β 是策略的初始分布，用于探索状态空间，在实际应用中相当于网络模型的初始参数。

DDPG 算法

OU 噪声作为 DDPG 算法中的一种探索策略，具有平滑、可控、稳定等优点，使得算法能够更好地在连续动作空间中进行训练，探索更广泛的动作空间，并找到更优的策略。它是 DDPG 算法成功应用于连续动作空间问题的重要因素之一。

OU 噪声主要由两个部分组成：随机高斯噪声和回归项

$$dx_t = \theta(\mu - x_t)dt + \sigma dW_t$$

DDPG算法的优缺点

优点	缺点
适用于连续动作空间	高度依赖超参数
高效的梯度优化	只适用于连续动作空间

优点	缺点
经验回放和目标网络	高度敏感的初始条件
	容易陷入局部最优

TD3算法

双Q网络

双 Q 网络的思想其实很简单，就是在 DDPG 算法中的 Critic 网络上再加一层，这样就形成了两个 Critic 网络。其本质和Double DQN的原理是一致的。

延迟更新

延迟更新更像是一种实验技巧，即在训练中 Actor 的更新频率要低于 Critic 的更新频率。

我们就可以在训练中让 Actor 的更新频率低于 Critic 的更新频率，这样一来 Actor 的更新就会比较稳定，不会受到 Critic 的影响，从而提高算法的稳定性和收敛性。

噪声正则

我们也可以给 Critic 引入一个噪声提高其抗干扰性，这样一来就可以在一定程度上提高 Critic 的稳定性，从而进一步提高算法的稳定性和收敛性。