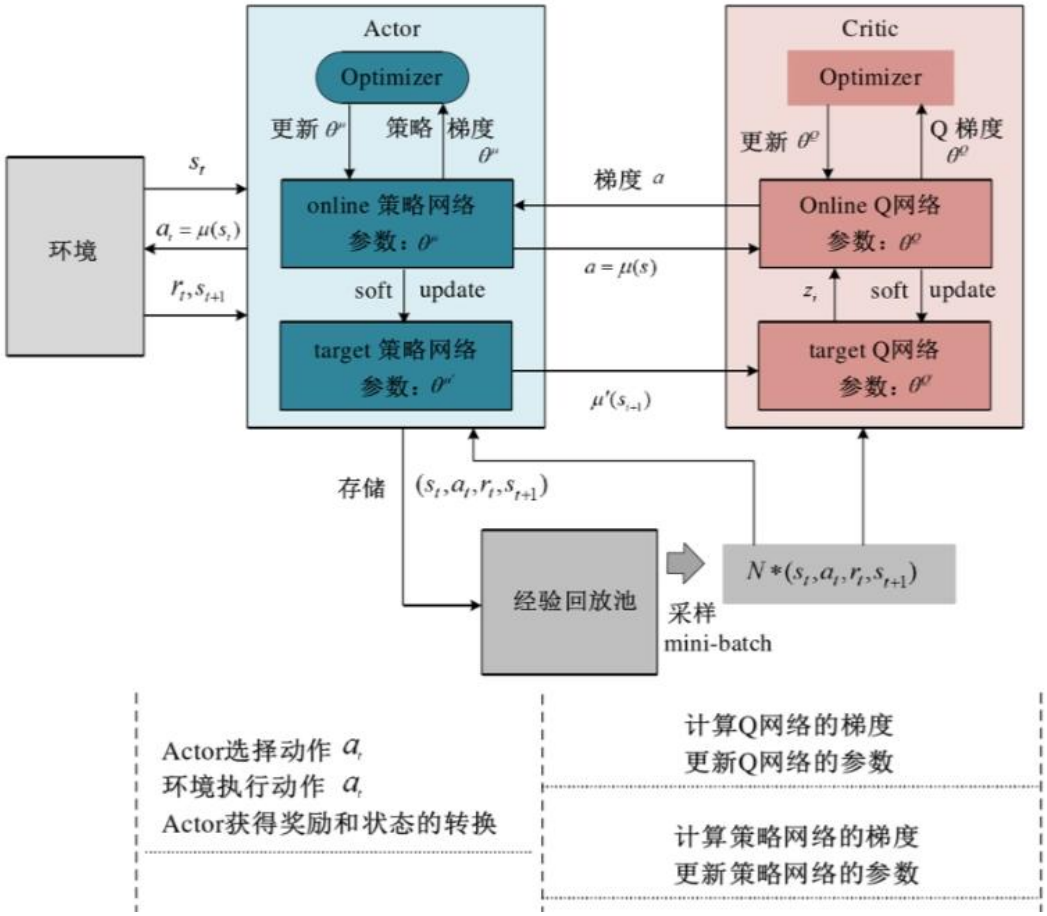
1. 背景

DDPG 算法是对 DQN 算法的一种改进，其采用 Actor-Critic 算法作为基本框架，采用神经网络作为策略网络和 q 值函数的近似，使用随机梯度法训练策略网络和价值网络中的参数。

DDPG 算法架构中使用双重神经网络架构，对于策略函数和价值函数均使用双重神经网络模型架构（Online 网络和 Target 网络），同时引入经验回放机制，Actor 与环境交互产生的经验数据样本存储到经验池中，抽取批量数据样本进行训练，去除样本的相关性和依赖性，使算法更加容易收敛。该算法框架如下所示：



2. 公式推导

DDPG 一共包含 4 个神经网络，用于对 Q 值函数和策略的近似表示。Critic 目标网络用于近似估计下一时刻的(state, action)的 Q 值函数$Q_{w_T}\left(s_{t+1}, \pi_{\theta_T}(a_{t+1}|s_{t+1})\right)$，其中，下一动作值是通过 Actor 目标网络近似估计得到的$\pi_{\theta_T}(a_{t+1}|s_{t+1})$，于是可以得到当前状态下的 Q 值函数的目标值：

$$y_T = r_{t+1} + \gamma Q_{w_T}\left(s_{t+1}, \pi_{\theta_T}(a_{t+1}|s_{t+1})\right)$$

Critic 训练网络输出当前时刻(state, action)的 Q 值函数，用于对当前策略进行评测。为了增强 exploration，DDPG 在行为策略上添加了高斯噪声函数：即使用 Actor 训练网络提供当前状态的策略，再加上一些探索噪声ε，得到当前状态的动作值：

$$a_t = \pi_\theta(a_t|s_t) + \varepsilon$$

则采用 Critic 训练网络输出当前时刻的(state, action)的 Q 值函数为$Q_w(s_t, a_t) = Q_w(s_t, \pi_\theta(a_t|s_t) + \varepsilon)$，Critic 训练网络的损失函数为，通过最小化损失函数更新 Critic 训练网络的参数：

$$loss = \frac{1}{N}\sum_{i=1}^{N}\left(y_{Ti} - Q_{wi}(s_t, a_t)\right)^2$$

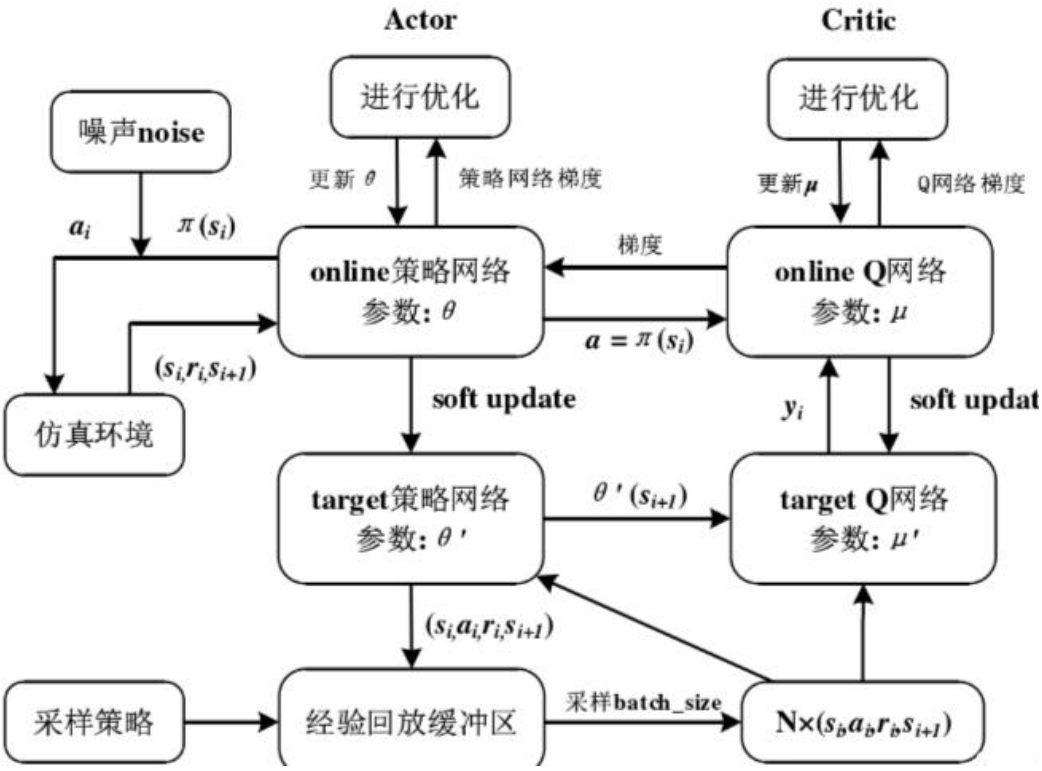
Actor 训练网络在参数更新时的策略梯度为：

$$\nabla_\theta J = \frac{1}{N}\sum_{i=1}^{N}\nabla_w Q_w(s_i, \pi_\theta(s_i))\nabla_\theta \pi_\theta(s_i)$$

对于目标网络参数w_T和θ_T的更新，DDPG 通过软更新机制(每次更新保留一部分原始值)保证参数可以缓慢更新，从而提高学习的稳定性：

$$w_T = \xi w + (1-\xi)w_T$$
$$\theta_T = \xi\,\theta + (1-\xi)\theta_T$$

该算法的流程图如下:



该算法的伪代码为:

DDPG 算法伪代码

θ^Q和θ^μ随机初始化 **Critic** 网络$Q(s,a|\theta^Q)$和 **Actor** 网络$\mu(s|\theta^\mu)$

$\theta^{Q'} \leftarrow \theta^Q, \theta^{\mu'} \leftarrow \theta^\mu$ 初始化目标网络权重参数Q'和μ'

初始化经验回放区 **R**

for episode = 1，M do：

行动探索，随机噪声 **N** 初始化

获得初始观察状态s_1

for t=1，T do：

$$a_t = \mu(s_t|\theta^\mu) + N_t$$

执行动作a_t，得到奖励r_t和环境状态s_{t+1}数据(s_t, a_t, r_t, s_{t+1})存入 **R**。

从 **R** 中随机采样批量数目值 **N** 的多维数组(s_i, a_i, r_i, s_{i+1})。

$$y_i = r_i + \gamma Q'(s_{i+1}, \mu'(s_{i+1}|\theta^{u'})|\theta^{Q'})$$

最小化损失函数 **L** 来更新 **Critic** 网络：

$$Loss = \frac{1}{N}\sum_i (y_i - Q(s_i, a_i|\theta^Q))^2$$

采样策略梯度更新 **Actor** 策略网络：

$$\nabla_{\theta^\mu} J(\theta^\mu) \cong \frac{1}{N}\sum_i \nabla_{\theta^\mu}\mu(s_i|\theta^\mu)\nabla_a Q(s_i, a|\theta^Q)|_{a=\mu(s_i)}$$

更新目标网络：

$$\theta^{Q'} \leftarrow \tau\theta^Q + (1-\tau)\theta^{Q'};$$
$$\theta^{\mu'} \leftarrow \tau\theta^\mu + (1-\tau)\theta^{\mu'}$$

end for

end for