10 - Actor-Critic 算法

策略梯度算法的优缺点

优点	缺点
适配连续动作空间	采样效率低
适配随机策略性	高方差
	收敛性差
	难以处理高维离散动作空间

Q Actor-Critic 算法

Actor-Critic 算法目标函数

$$\nabla_\theta J(\theta) \propto \mathbb{E}_{\pi\theta}[Q^\theta(s,a)\nabla_\theta \log \pi_\theta(a|s)]$$

当REINFORCE算法中接入Q函数，我们可以得到目标函数

$$\nabla_\theta J(\theta) \propto \mathbb{E}_{\pi\theta}[Q_\phi(s_t,a_t)\nabla_\theta \log \pi_\theta(a_t|s_t)]$$

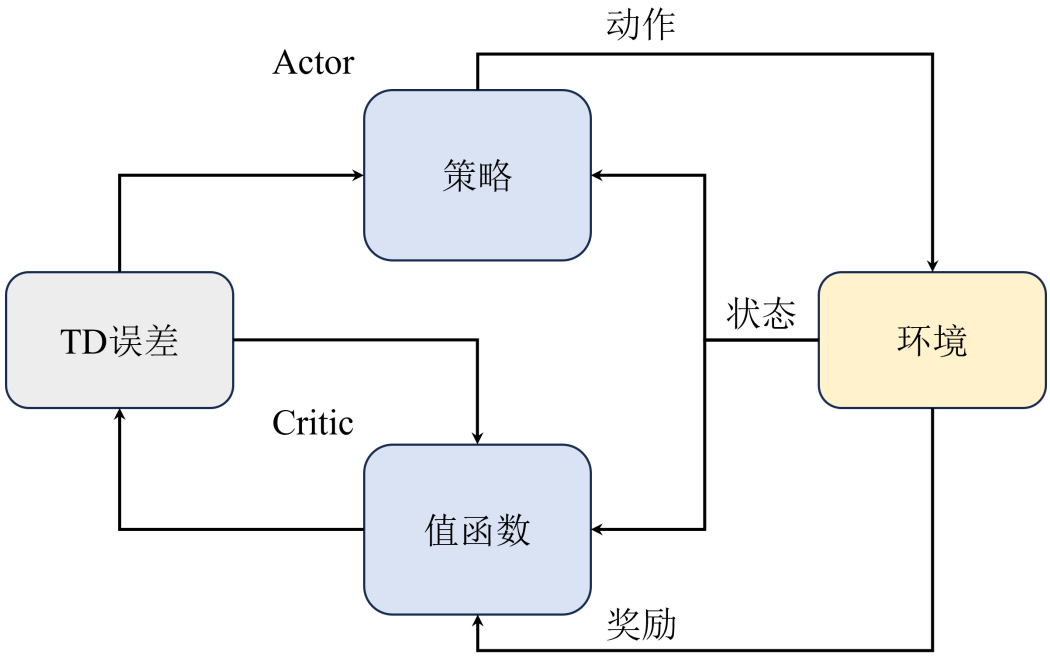

图 10.1 Actor-Critic 算法架构

A2C 与 A3C 算法

为了解决Actor-Critic算法中的高方差问题，我们引入一个优势函数用于表示当前状态-动作对相对于平均水平的优势

$$A^\pi(s_t,a_t) = Q^\pi(s_t,a_t) - V^\pi(s_t)$$

加入了优势函数，目标函数可以表示为

$$\nabla_\theta J(\theta) \propto \mathbb{E}_{\pi\theta}[A^\pi(s_t,a_t)\nabla_\theta \log \pi_\theta(a_t|s_t)]$$

这边是A2C算法，而A3C算法增加了多个进程，每个进程可以独立进行交互以提高训练效率。

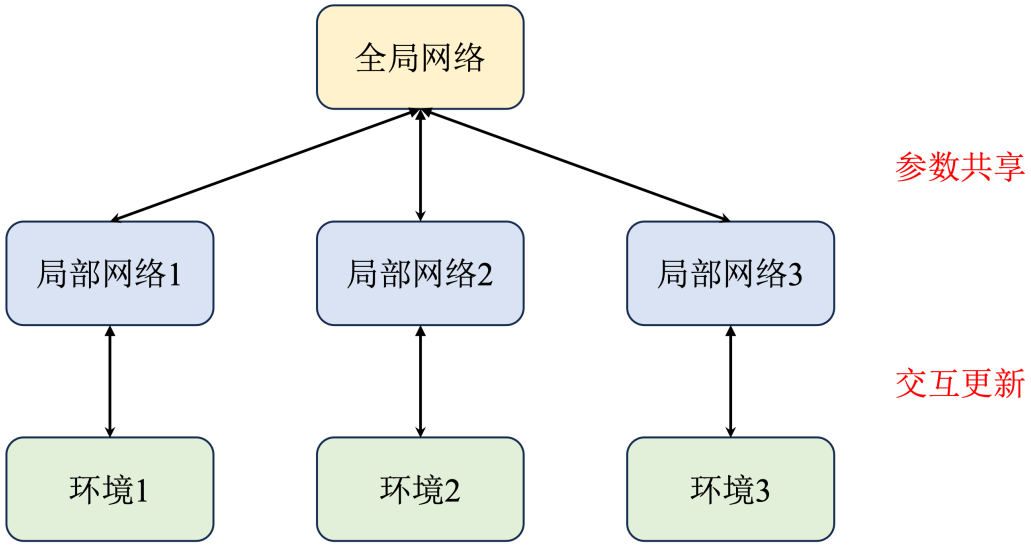

图 10.2 A3C 算法架构

广义优势估计

在A2C算法中引入了优势函数，但是优势函数的本质还是蒙特卡洛法，这可能有时候会产生高方差。前面章节有提到蒙特卡洛和时序差分方法的差异，我们可以发现可两个方法是形成一个互补的关系，因此我们可以融合这两个方法形成一个新的估计方式——广义优势估计（generalized advantage estimation，GAE）

$$
\begin{aligned}
A^{\mathrm{GAE}(\gamma,\lambda)}(s_t,a_t) &= \sum_{t=0}^{\infty}(\gamma\lambda)^l\delta_{t+1} \\
&= \sum_{t=0}^{\infty}(\gamma\lambda)^l(r_{t+l}+\lambda V^{\pi}(s_{t+l+1})-V^{\pi}(s_{t+l}))
\end{aligned}
$$

其中 δ_{t+1} 表示时间步 $t+l$ 时的 TD 误差