策略梯度算法

之前的Q_learning和DQN都是基于价值的算法，通过拟合值函数导出一个策略。策略梯度则是基于策略的算法，输入状态直接输出一个目标策略。

策略梯度公式推导

1. **优化目标的梯度**： 使用策略函数 π_θ 表示选择动作 a 的概率，优化目标为：

$$J(\theta) = \mathbb{E}_{\pi_\theta}\left[\sum_{t=0}^{\infty} \gamma^t r_t\right]$$

通过梯度上升法更新参数 θ：

$$\theta \leftarrow \theta + \alpha \nabla_\theta J(\theta)$$

2. **梯度计算：REINFORCE 方法**： 根据策略梯度定理，有：

$$\nabla_\theta J(\theta) = \mathbb{E}_{\pi_\theta}\left[\nabla_\theta \log \pi_\theta(a|s) G_t\right]$$

- G_t 是从时间步 t 开始的累计回报：

$$G_t = \sum_{k=t}^{\infty} \gamma^{k-t} r_k$$

- 核心思想是将目标的期望梯度分解为策略概率的梯度 $\nabla_\theta \log \pi_\theta(a|s)$ 和奖励信号 G_t 的乘积。

3. **梯度更新公式**： 每一步根据采样得到的经验，用以下公式更新策略参数：

$$\theta \leftarrow \theta + \alpha \nabla_\theta \log \pi_\theta(a|s) G_t$$