**游戏记录：**

，也可以说是游戏环境本身。

计算时函数得出最后的状态的可能性，

所以要使用乘法

**目标函数如何来进行求解**

公式的含义是： 找到最佳策略，过程是找到最高期望值

Trajectory轨迹，表明一组状态对应行为的一系列数据

**按照大数定律的思想，直接穷举所有的可能性就好了：**

**转成积分**

**计算梯度：**

**推到过程：**

**奖励去均值**

**重要性采样**

**找到的近似替代**

**条件限制**

**近端策略优化2**