1. 背景

多智能体的情形相比于单智能体更加复杂，因为每个智能体在和环境交互的同时也在和其他智能体进行直接或间接的交互。因此，多智能体强化学习要比单智能体更困难，其难点主要体现在以下几点：

（1） 由于多个智能体在环境中进行实时动态交互，并且每个智能体在不断学习并更新自身策略，因此在每个智能体的视角下，环境是非稳态的，即对于每一个智能体而言，即使在相同的状态下采取相同的动作，得到的状态转移和奖励信号的分布可能在不断改变

（2） 多个智能体的训练可能是多目标的，不同智能体需要最大化自己的利益

（3） 训练评估的复杂度会增加，可能需要大规模分布时训练来提高效率

2. 算法原理

IPPO（Independent PPO）是一种完全去中心化的算法，此类算法被称为独立学习。对于每个智能体使用单智能体算法 PPO 进行训练，因此这个算法叫做独立 PPO 算法，其算法流程如图：

- 对于N个智能体，为每个智能体初始化各自的策略以及价值函数
- **for** 训练轮数 $k = 0, 1, 2 \ldots$ **do**
- 所有智能体在环境中交互分别获得各自的一条轨迹数据
- 对每个智能体，基于当前的价值函数用 GAE 计算优势函数的估计
- 对每个智能体，通过最大化其 PPO-截断的目标来更新其策略
- 对每个智能体，通过均方误差损失函数优化其价值函数
- **end for**