使用ppo训练了两个agent，使用不同随机种子，不同网络结构

agent1：2个隐藏层，每层256个

agent2：4个隐藏层，每层128个

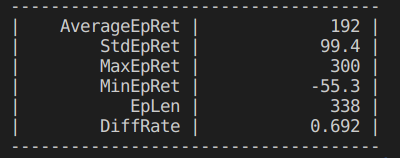
停止训练的时候，20回合测试平均结果都在200左右（收敛状态应该在300左右或更高）

两个agent策略的不一致性测试：

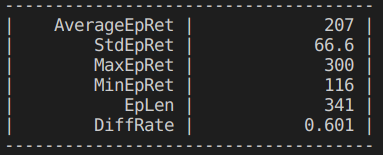
使用一个agent决策，生成轨迹（另一个agent也计算当前状态下的a，但不执行），不一致率大概在0.55-0.65左右

随机指定策略的性能评估：

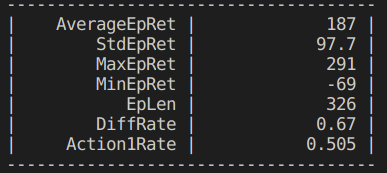
只用agent1



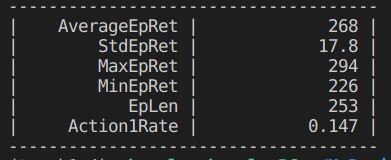
只用agent2



随机选择agent1或2



训练meta网络之后的结果



训练曲线

