:retailer 's demand at period
:retailer 's inventory at the end of period
:retailer 's inventory at the start of period
:transship quantity
:order quantity
:intransit order

actor 's observation:

heuristic: 6.296
no transship:6.237
MADRL: 5.925

目前进展:

1. 基本完成考虑transship的多期库存问题的建模
2. 对happo强化学习框架熟悉并予以多次尝试

后续计划:

1. 学习maddpg框架（Multi-agent Deep Deterministic Policy Gradient）。在对happo尝试过程中发现训练效果并不好，猜测是action space较大导致policy network的训练较为困难，而maddpg中的policy network的输出为某一确定action，而非每个action的概率，故可有效应对action space这一问题。（论文：Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments，代码：https://github.com/openai/maddpg）
2. 尝试令各retailer共享policy，即只训练一个policy network以供所有agent使用。当各retailer的参数设定基本一致时，retailer间具有同质性，policy应当也不会有太大变化，可以尝试让各retailer共享一个policy network，即由多智能体强化学习转为单智能体强化学习了。可作为一个baseline，对比多智能体强化学习的效果。
3. 尝试centralized controller作为baseline，通过调整transship的交易费用来试图使得decentral接近central的效果。
4. 灵敏度分析。检验多智能体强化学习在不同设定下的效果：

- Transship的运输费用为fixed cost
- 考虑不同零售商间的transship运输费用不同的情况
- Transship有lead time，如1天
- 不同零售商的需求并非独立，而存在一定的联动

寒假计划先完成前3个