结论和想法

4个dqn训练1w次收敛到总cost 4000，并不好，但是发现一个有趣的现象。然后完做了之前其他的实验。

发现四个dqn很难收敛的主要原因是第四个agent 生产厂商，容易血崩，全要看他发挥好不好。其他的三个agent dqn已经很好了。然后发现， 如果厂商采取保守的策略，比如simple policy，其他agent用dqn，整体可以表现的很好。厂商用dqn这种需要“智慧”的策略感觉比较难，

可能这就是牛鞭效应，太复杂了很难精确的学习到？

另外更多的有趣的现象和解读就麻烦天辰了。

实验结果

simple policy指，最简单的策略。如果上期亏损，增加10%订单。

第一个实验

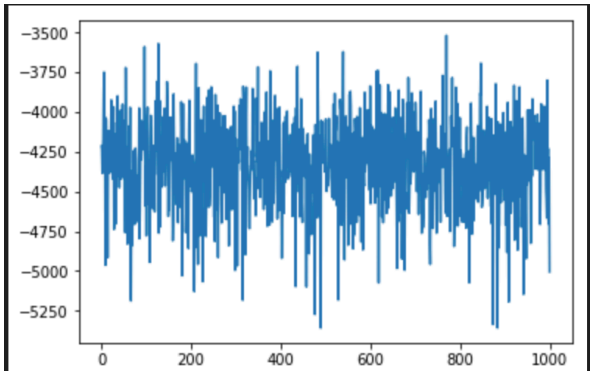
四个 simple policy

数据： 4simple_policy_1000.pickle

保存方式：

```
data = {0:np.array(bg.agents[0].cum_r),
        1:np.array(bg.agents[1].cum_r),
        2:np.array(bg.agents[2].cum_r),
        3:np.array(bg.agents[3].cum_r)}
import pickle
with open("./data/4simple_policy_1000.pickle", "wb+") as f:
    pickle.dump(data, f)
```

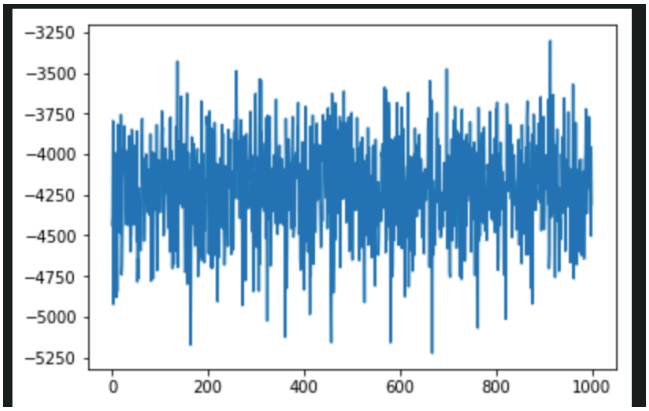
总体cost



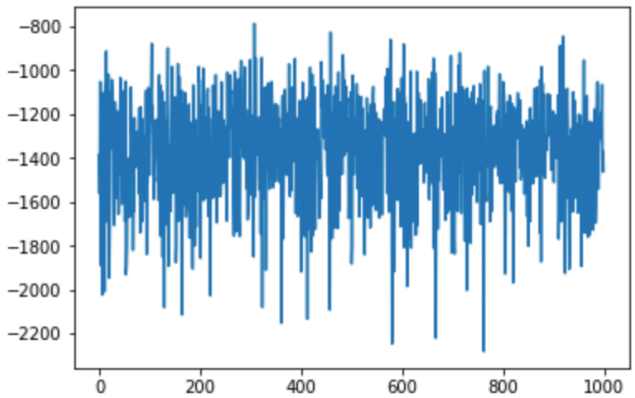
第二个实验
零售agent ar1
数据：3simple_policy_1ar_1000.pickle

```
data = {0:np.array(bg.agents[0].cum_r),
        1:np.array(bg.agents[1].cum_r),
        2:np.array(bg.agents[2].cum_r),
        3:np.array(bg.agents[3].cum_r)}
import pickle
with open("./data/3simple_policy_1ar_1000.pickle", "wb+") as f:
    pickle.dump(data, f)
```

总体cost

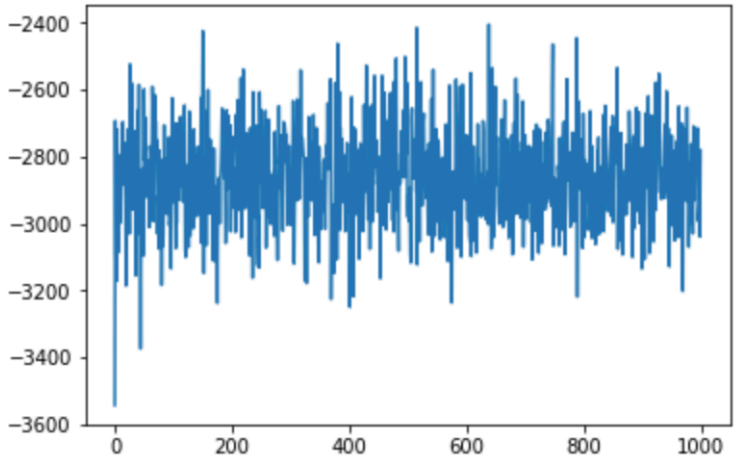


厂商的cost，比较大

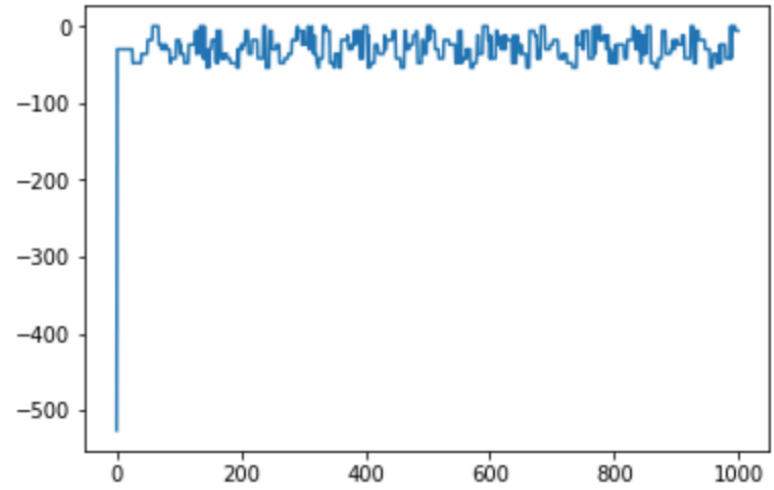


第三个实验
零售agent ar1，政委dqn

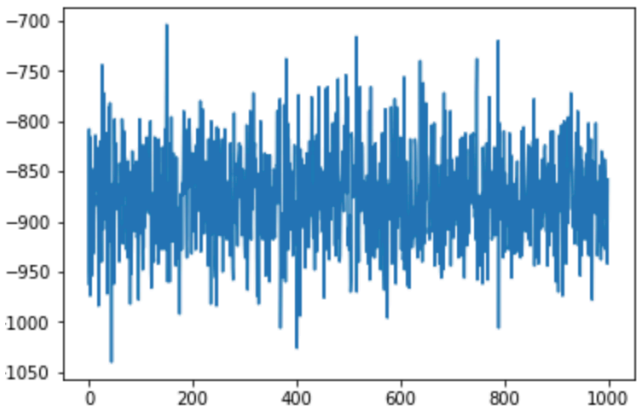
总cost显著地比前两个好了。



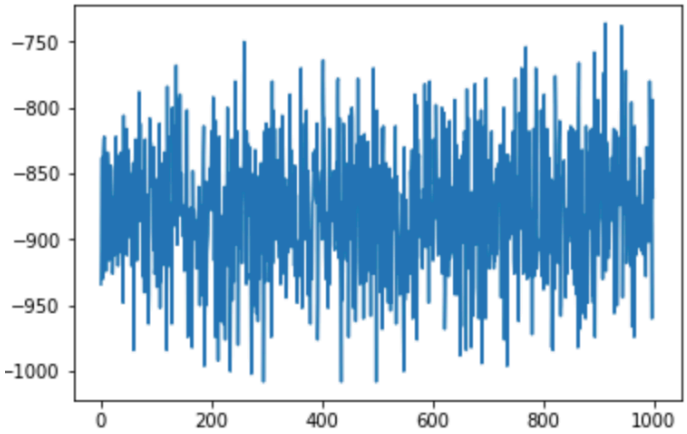
这里的厂商表现很好



政委dqn的表现, -879,和第二个实验原来差不多（图在下面），看来主要是厂家的原因，使得整体变好了。



政委在第二个实验中的表现， 均值 -877

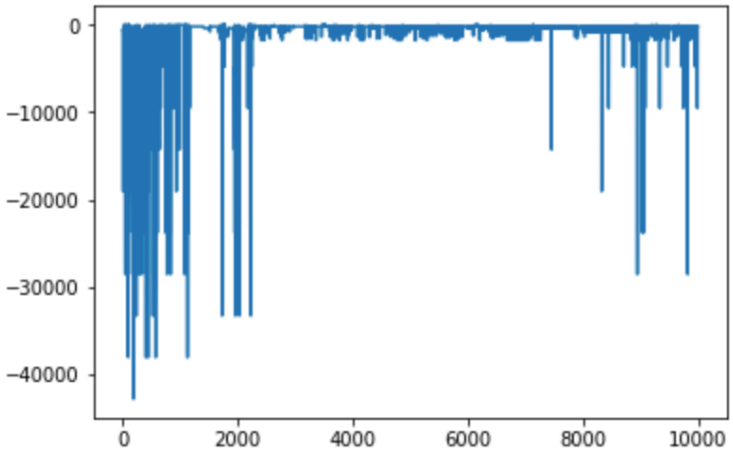


第四个实验

4dqn_10000 是4个agent使用dqn的数据，10000个训练次数的数据的pickle。其中数据是这样保存的

```
data = {0:np.array(bg.agents[0].cum_r),
        1:np.array(bg.agents[1].cum_r),
        2:np.array(bg.agents[2].cum_r),
        3:np.array(bg.agents[3].cum_r)}
import pickle
with open("./data/savedata_4dqn_10000.pickle", "wb+") as f:
    pickle.dump(data, f)
```

厂商容易血崩

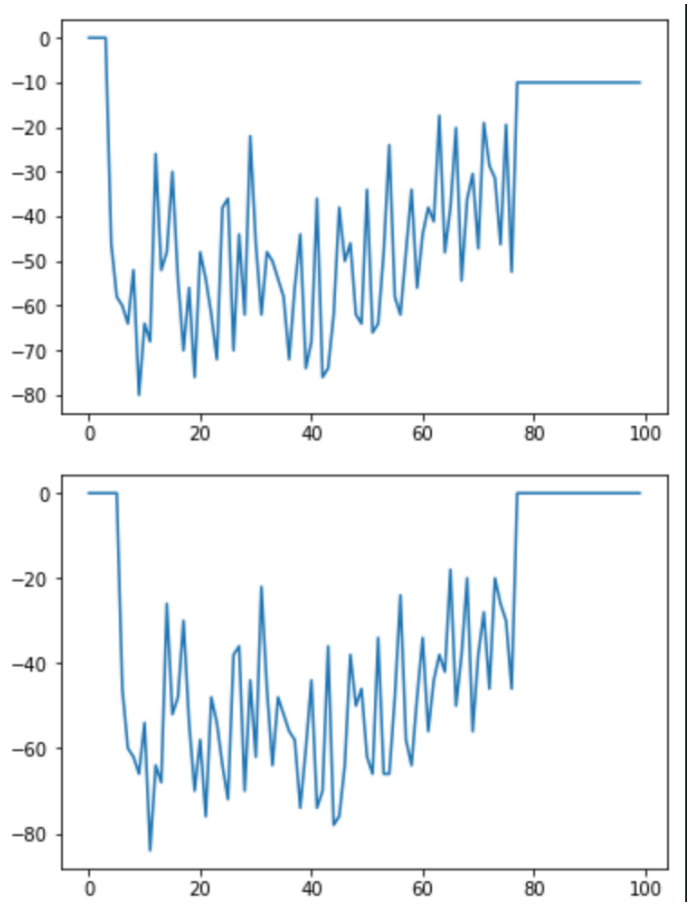


第五个实验 分 4 个 step 的
数据储存：

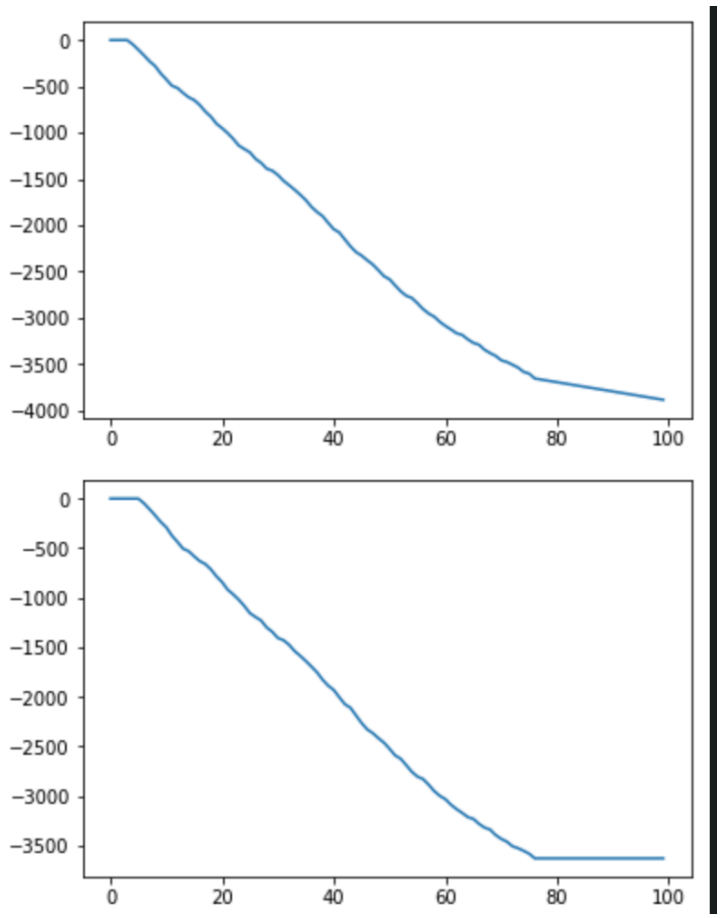
```
data = {0:np.array(bg.cum_r_0),
        1:np.array(bg.cum_r_1),
        2:np.array(bg.cum_r_2),
        3:np.array(bg.cum_r_3),
        "0":np.array(bg.r_0),
        "1":np.array(bg.r_1),
        "2":np.array(bg.r_2),
        "3":np.array(bg.r_3),}


import pickle
with open("./data/4step.pickle", "wb+") as f:
    pickle.dump(data, f)
```

cum_r_0 是总的成本
r_0 是每次的成本记录

零售商(agent0)和批发商(agent1)用了 dqn 每次的 reward 在变好



选了最后 5 次实验的 cost 平均发现，斜率是降低了,说明是好一点了。


0
−500
−1000
−1500
−2000
−2500
−3000
−3500
−4000
0
20
40
60
80
100
0
−500
−1000
−1500
−2000
−2500
−3000
−3500
0
20
40
60
80
100