在 gym 中可以获得讯息

- 一辆车在一个二维的轨道上，位于两座山之间。
- 目标是冲上右边的山头。
- 动作集：一维的离散空间：$motor=(left,neutral,right)$
- 状态变量：二维的连续状态空间：
- $Velocity=(-0.07,0.07)$ ，$Position=(-1.2,0.6)$
- 奖励：每过去一个时间，$reward=-1$
- 状态更新：

  $Velocity=Velocity+(Action)*0.001+\cos(3*Position)*(-0.0025)$Velocity=Velocity+(Action)*0.001+cos(3*Position)*(–0.0025)，

  $Position=Position+Velocity$Position=Position+Velocity
- 初始状态：$Position=-0.5$ Position=–0.5， $Velocity=0.0$ Velocity=0.0
- 终止状态：$Position\geq 0.6$

1. 为什么选择 DQN,问题本身是 MDP 的知道状态和连续动作，可以用物理学推导出下一个状态和连续动作。但是问题是，状态是连续的不能用单一的一个大 Q 表表示。所以选择使用 DQN

   Policy gradient:

   选择的理由：更好的收敛性质

   能有效处理连续空间，高维度资讯

   能学习复杂的策略

2. Q-learning is an off-policy algorithm, since it updates the Q values without

making any assumptions about the actual policy being followed.

Policy gradient is an on-policy algorithm, which will directly learn the policy on each state

3. It uses a replay buffer to store the **experiences** of the agent during training, and then randomly sample experiences to use for learning in order to break up the temporal correlations within different training episodes. This technique is known as **experience replay**.

实验结果分析：

不同 batch_size 对 train 影响：

Batch_size 32:

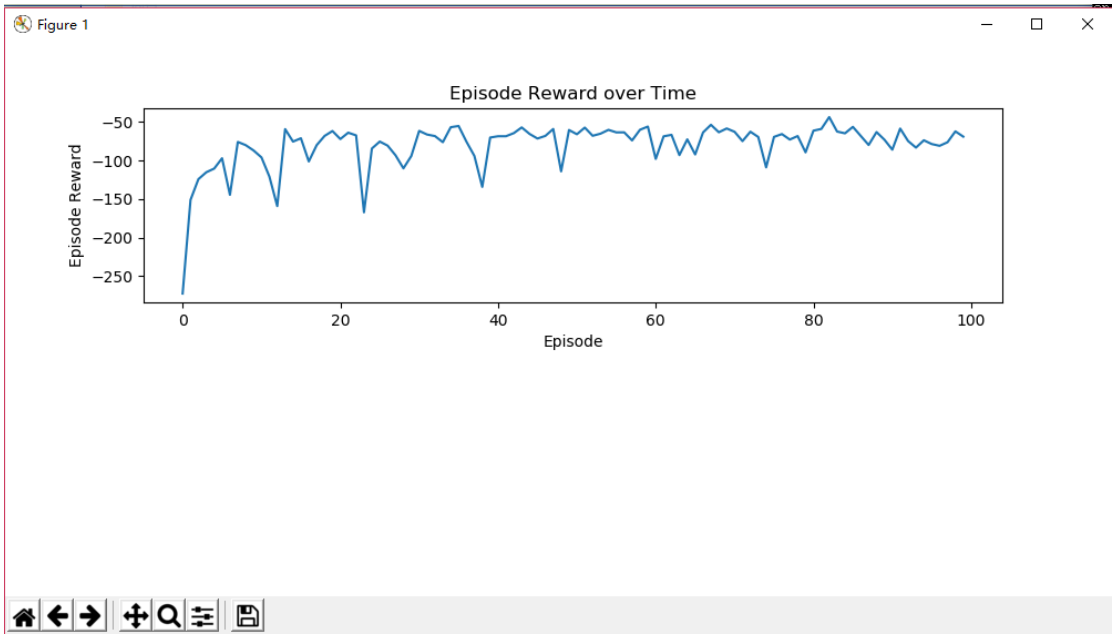


Figure 1
Episode Reward over Time
Episode Reward
−50
−100
−150
−200
−250
0
20
40
60
80
100
Episode

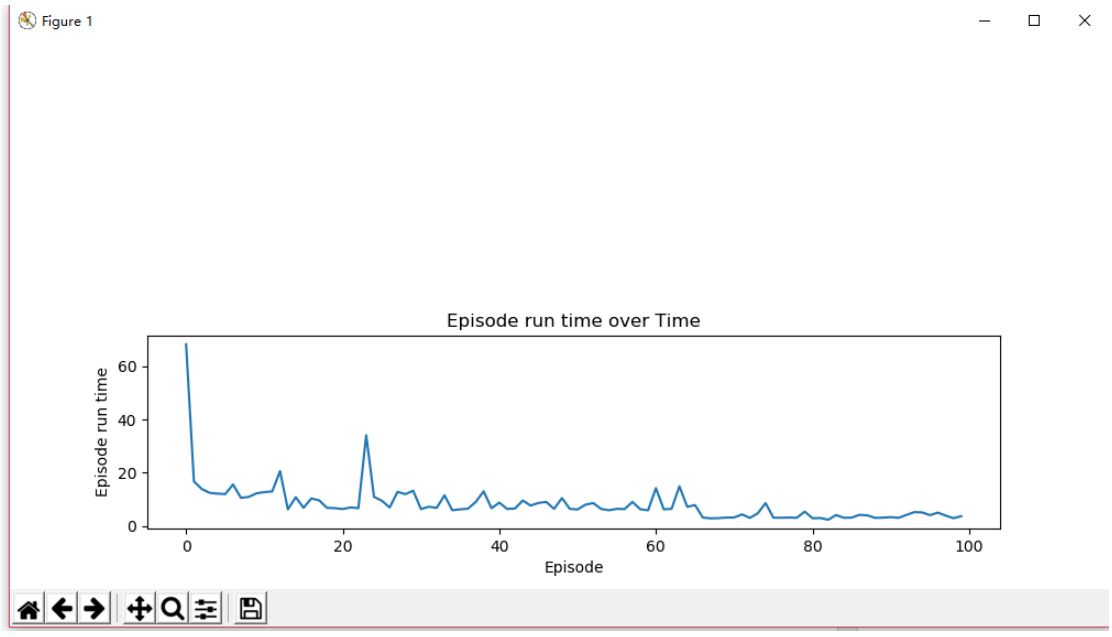
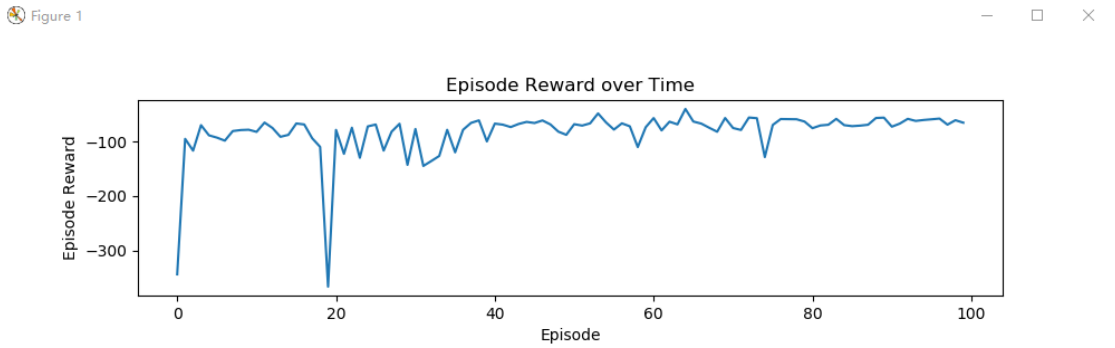
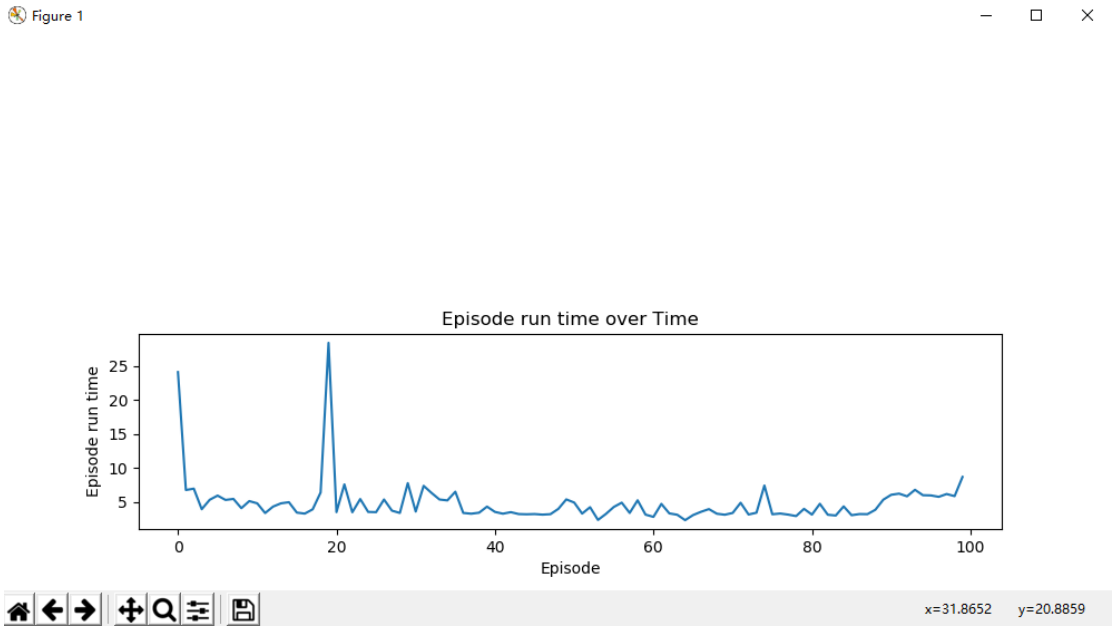


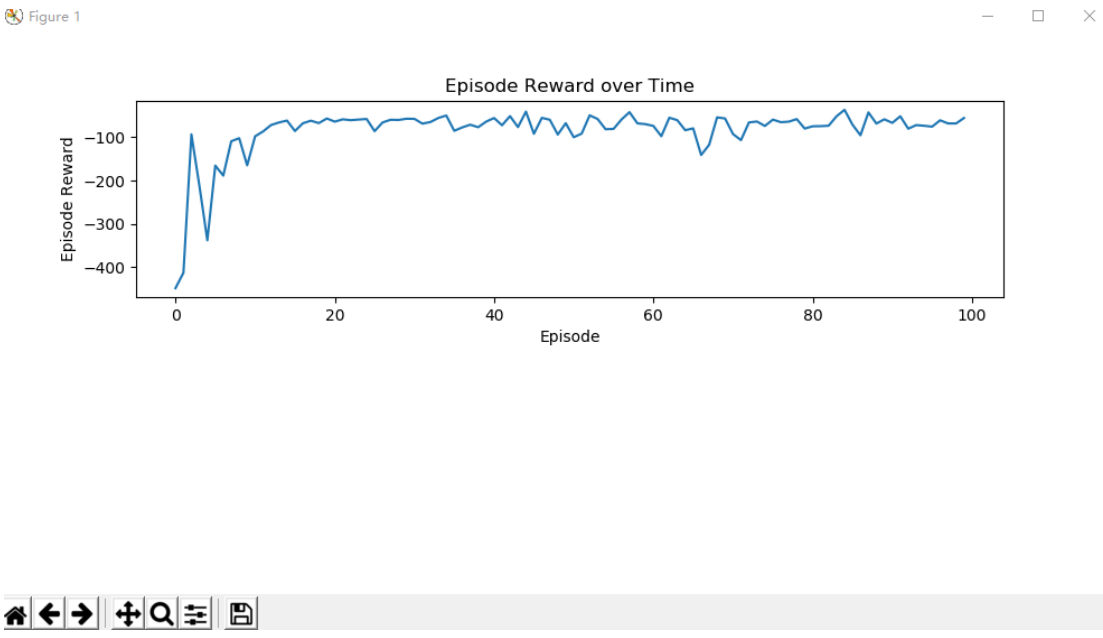
Figure 1
Episode run time over Time
Episode run time
60
40
20
0
0
20
40
60
80
100
Episode

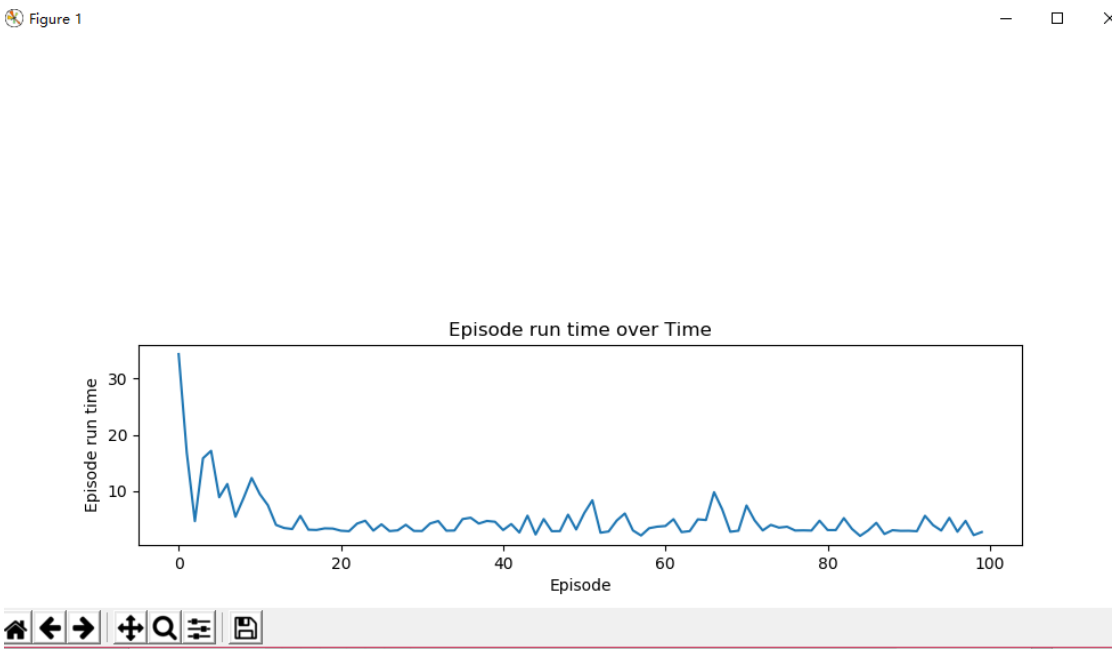
Batch_size:128





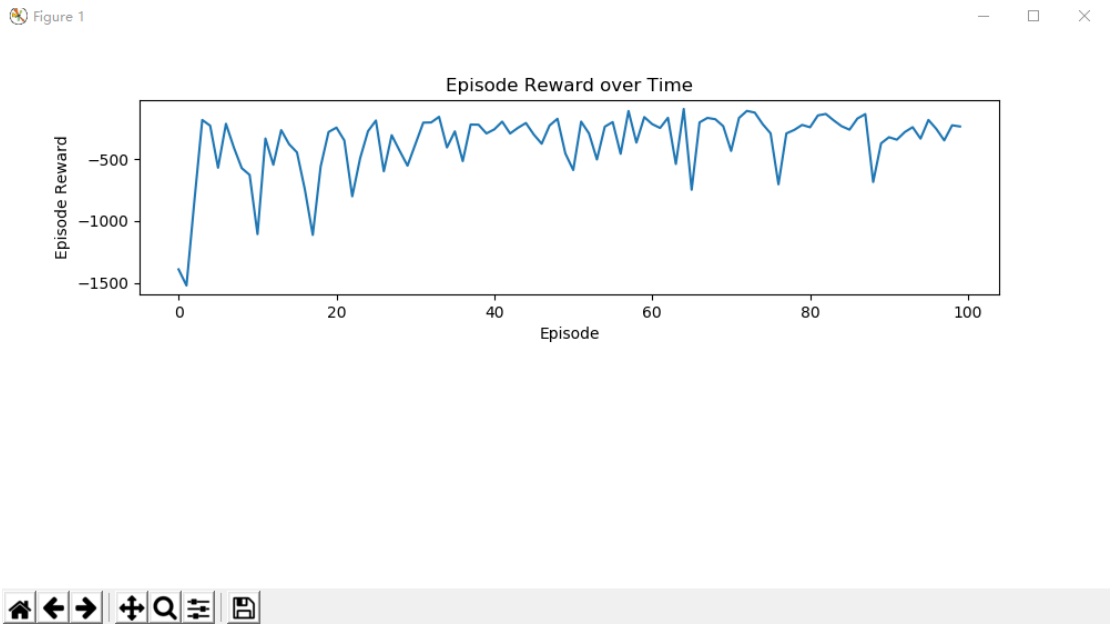
Batch_size 16:

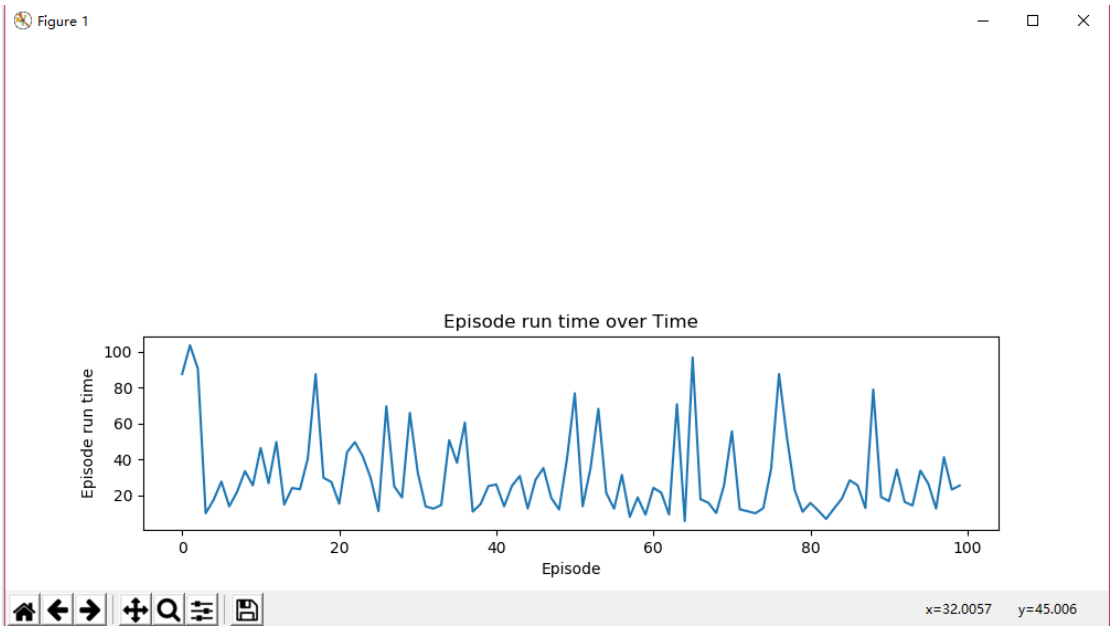




实验结果显示：batch_size 越大收敛的越快，但在训练过程中会出现比较大的震荡，小的 batch_size 会比较平滑的收敛。

Learning_rate:0.1





Learning rate:0.001

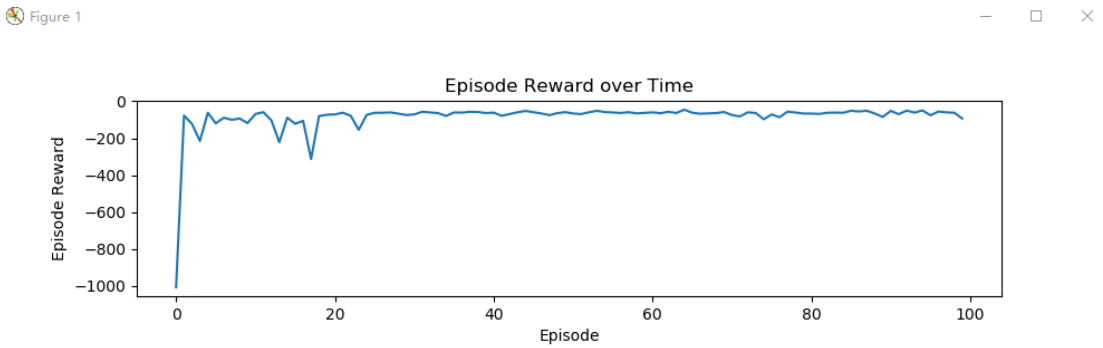


Figure 1
Episode Reward over Time
Episode Reward
0
−200
−400
−600
−800
−1000
0
20
40
60
80
100
Episode

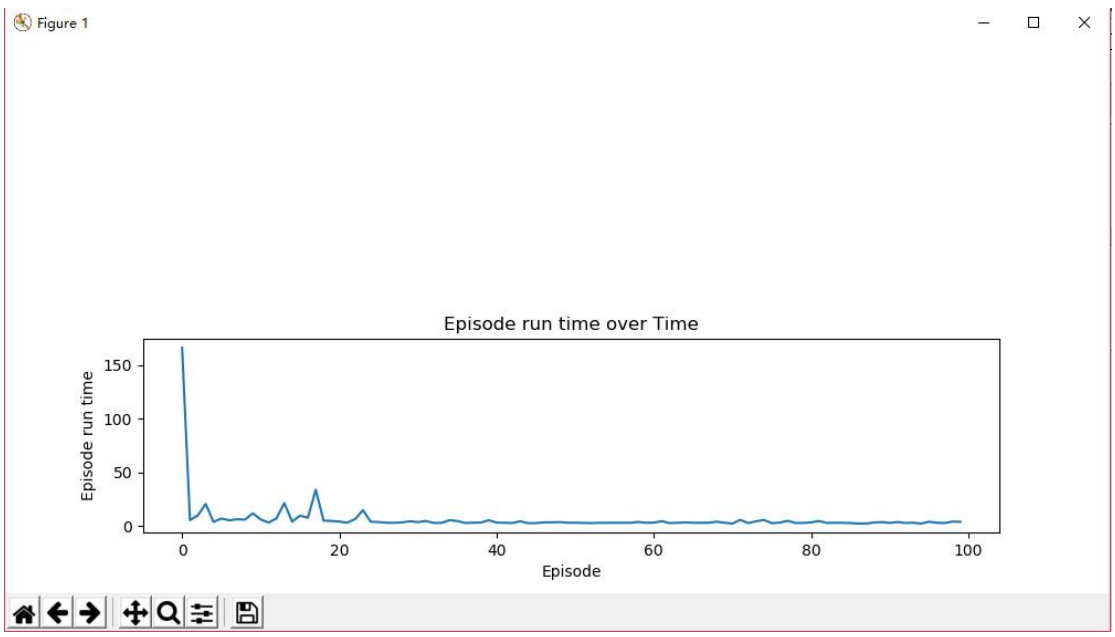
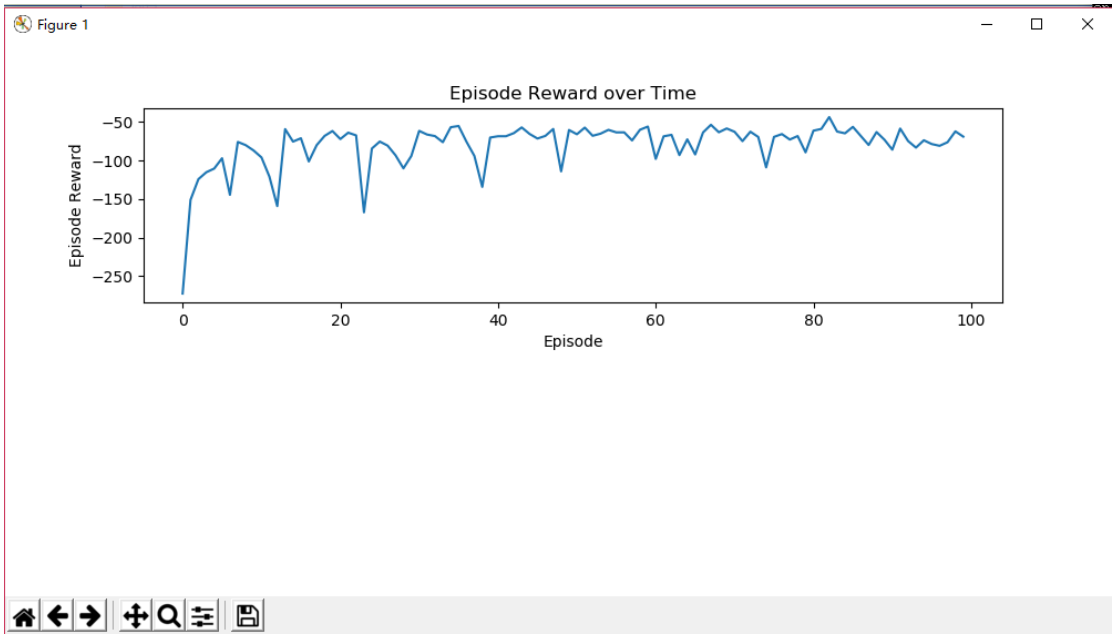
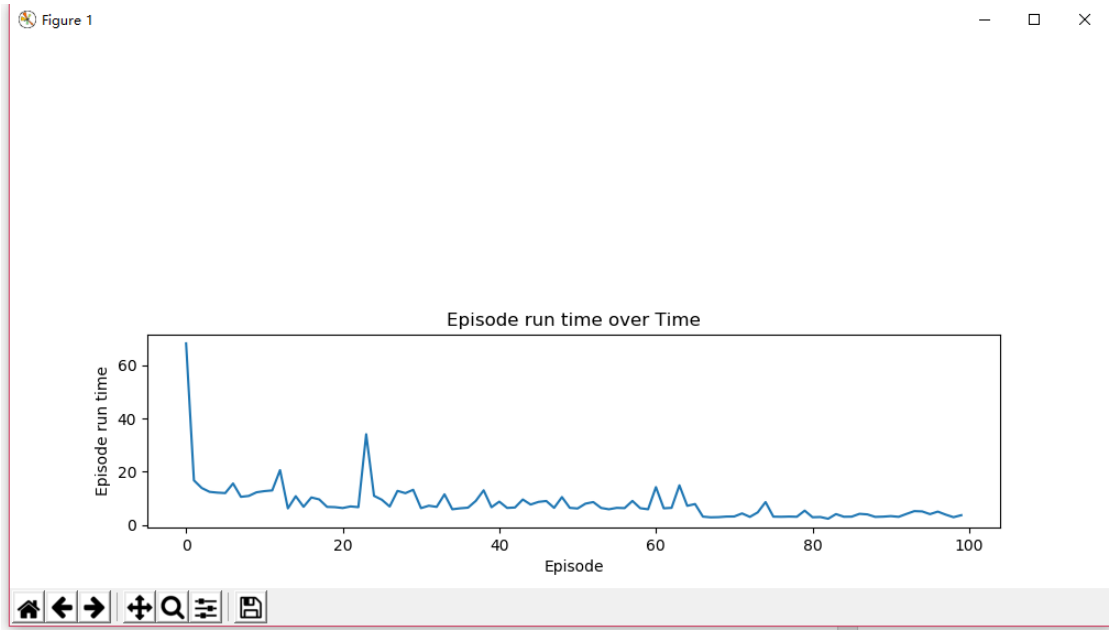


Figure 1
Episode run time over Time
Episode run time
150
100
50
0
0
20
40
60
80
100
Episode

Learning rate0.01：

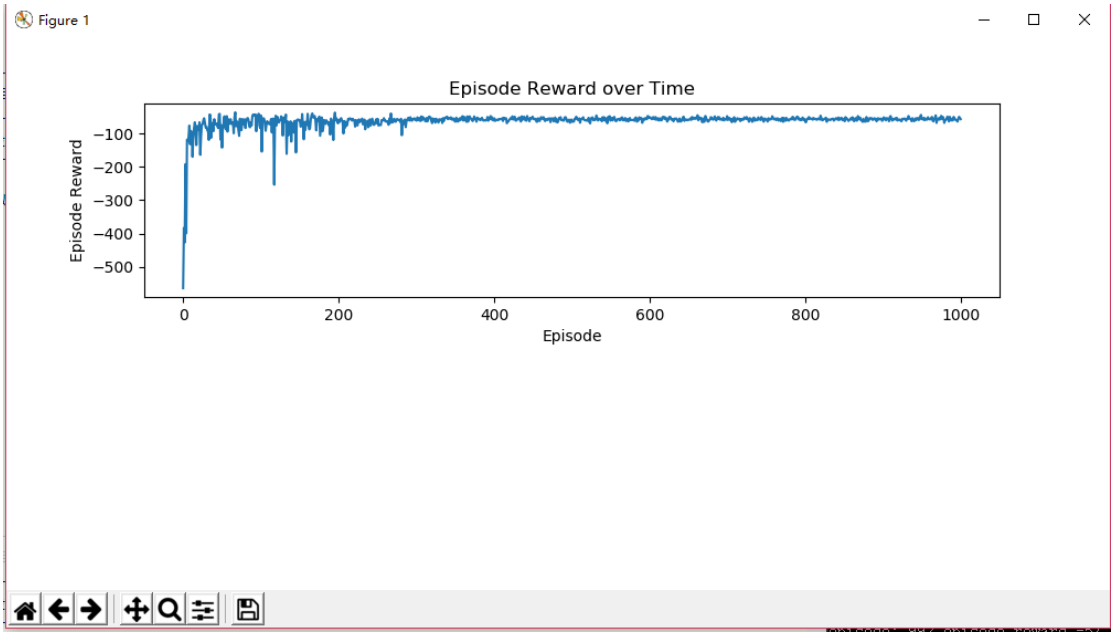


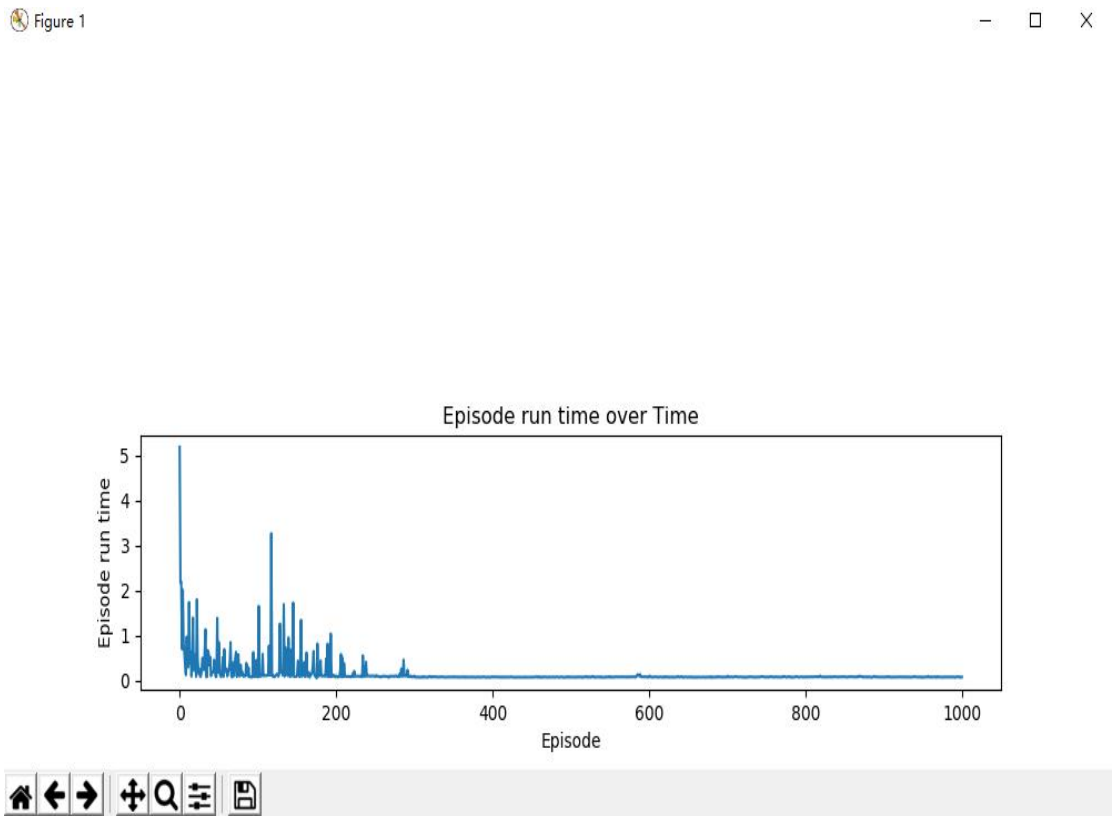


Learning rate0.1 可以比较快速的收敛，但是很难学习到 global minimal，

Learning rate0.01，和 0.001 都可以收敛到比较好的结果，并且 learning rate 0.01 有更快的收敛速度，但 learning rate 0.001 有更少的震荡 故选择 learning rate 0.001

Q-learning 1000 episodes over time:





Q_LEARNING TEST EVERY 100 EPISODES:

100:

```
episode: 0  episode_reward -281.00 epsilon 0.90
episode: 1  episode_reward -276.00 epsilon 0.90
episode: 2  episode_reward -290.00 epsilon 0.90
episode: 3  episode_reward -217.00 epsilon 0.90
episode: 4  episode_reward -1403.00 epsilon 0.90
episode: 5  episode_reward -234.00 epsilon 0.90
episode: 6  episode_reward -1203.00 epsilon 0.90
episode: 7  episode_reward -215.00 epsilon 0.90
episode: 8  episode_reward -3315.00 epsilon 0.90
episode: 9  episode_reward -170.00 epsilon 0.90
episode: 101 episode_reward -153.65 epsilon 0.90
```

200:

```
episode: 0  episode_reward -172.00 epsilon 0.90
episode: 1  episode_reward -176.00 epsilon 0.90
episode: 2  episode_reward -171.00 epsilon 0.90
episode: 3  episode_reward -170.00 epsilon 0.90
episode: 4  episode_reward -173.00 epsilon 0.90
episode: 5  episode_reward -170.00 epsilon 0.90
episode: 6  episode_reward -172.00 epsilon 0.90
episode: 7  episode_reward -180.00 epsilon 0.90
episode: 8  episode_reward -191.00 epsilon 0.90
episode: 9  episode_reward -166.00 epsilon 0.90
```

300:

```
episode: 0  episode_reward -167.00 epsilon 0.90
episode: 1  episode_reward -161.00 epsilon 0.90
episode: 2  episode_reward -165.00 epsilon 0.90
episode: 3  episode_reward -160.00 epsilon 0.90
episode: 4  episode_reward -168.00 epsilon 0.90
episode: 5  episode_reward -168.00 epsilon 0.90
episode: 6  episode_reward -157.00 epsilon 0.90
episode: 7  episode_reward -167.00 epsilon 0.90
episode: 8  episode_reward -167.00 epsilon 0.90
episode: 9  episode_reward -164.00 epsilon 0.90
```

400:

```
episode: 0  episode_reward -155.00 epsilon 0.90
episode: 1  episode_reward -162.00 epsilon 0.90
episode: 2  episode_reward -161.00 epsilon 0.90
episode: 3  episode_reward -162.00 epsilon 0.90
episode: 4  episode_reward -158.00 epsilon 0.90
episode: 5  episode_reward -158.00 epsilon 0.90
episode: 6  episode_reward -152.00 epsilon 0.90
episode: 7  episode_reward -163.00 epsilon 0.90
episode: 8  episode_reward -157.00 epsilon 0.90
episode: 9  episode_reward -157.00 epsilon 0.90
```

500:

```
episode: 0  episode_reward -162.00 epsilon 0.90
episode: 1  episode_reward -156.00 epsilon 0.90
episode: 2  episode_reward -161.00 epsilon 0.90
episode: 3  episode_reward -156.00 epsilon 0.90
episode: 4  episode_reward -163.00 epsilon 0.90
episode: 5  episode_reward -163.00 epsilon 0.90
episode: 6  episode_reward -151.00 epsilon 0.90
episode: 7  episode_reward -157.00 epsilon 0.90
episode: 8  episode_reward -166.00 epsilon 0.90
episode: 9  episode_reward -162.00 epsilon 0.90
```

600:

```
episode: 0  episode_reward -154.00 epsilon 0.90
episode: 1  episode_reward -160.00 epsilon 0.90
episode: 2  episode_reward -161.00 epsilon 0.90
episode: 3  episode_reward -160.00 epsilon 0.90
episode: 4  episode_reward -162.00 epsilon 0.90
episode: 5  episode_reward -167.00 epsilon 0.90
episode: 6  episode_reward -168.00 epsilon 0.90
episode: 7  episode_reward -157.00 epsilon 0.90
episode: 8  episode_reward -158.00 epsilon 0.90
episode: 9  episode_reward -164.00 epsilon 0.90
episode: 601 episode_reward -52.87 epsilon 0.90
```

700:

```
episode: 0  episode_reward -163.00 epsilon 0.90
episode: 1  episode_reward -158.00 epsilon 0.90
episode: 2  episode_reward -157.00 epsilon 0.90
episode: 3  episode_reward -163.00 epsilon 0.90
episode: 4  episode_reward -160.00 epsilon 0.90
episode: 5  episode_reward -148.00 epsilon 0.90
episode: 6  episode_reward -157.00 epsilon 0.90
episode: 7  episode_reward -161.00 epsilon 0.90
episode: 8  episode_reward -158.00 epsilon 0.90
episode: 9  episode_reward -163.00 epsilon 0.90
```

800:

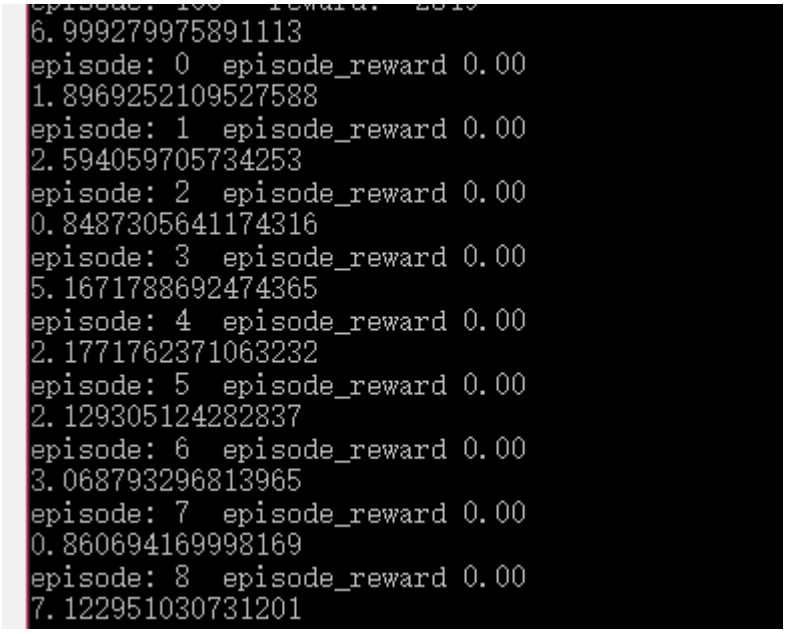
```
episode: 0  episode_reward -160.00 epsilon 0.90
episode: 1  episode_reward -164.00 epsilon 0.90
episode: 2  episode_reward -151.00 epsilon 0.90
episode: 3  episode_reward -165.00 epsilon 0.90
episode: 4  episode_reward -163.00 epsilon 0.90
episode: 5  episode_reward -160.00 epsilon 0.90
episode: 6  episode_reward -162.00 epsilon 0.90
episode: 7  episode_reward -156.00 epsilon 0.90
episode: 8  episode_reward -161.00 epsilon 0.90
episode: 9  episode_reward -162.00 epsilon 0.90
episode: 801 episode_reward -58.61 epsilon 0.90
```

900:

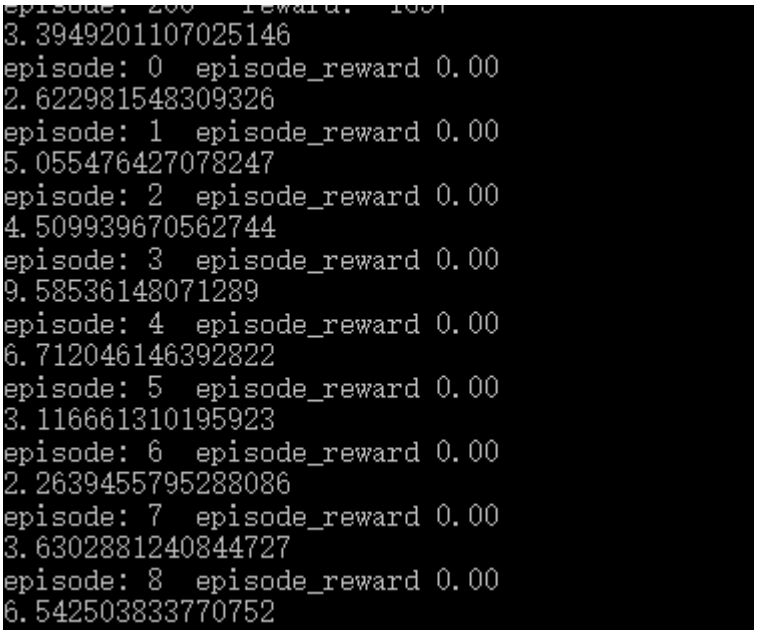
```
episode: 0  episode_reward -160.00 epsilon 0.90
episode: 1  episode_reward -149.00 epsilon 0.90
episode: 2  episode_reward -158.00 epsilon 0.90
episode: 3  episode_reward -162.00 epsilon 0.90
episode: 4  episode_reward -157.00 epsilon 0.90
episode: 5  episode_reward -162.00 epsilon 0.90
episode: 6  episode_reward -158.00 epsilon 0.90
episode: 7  episode_reward -152.00 epsilon 0.90
episode: 8  episode_reward -156.00 epsilon 0.90
episode: 9  episode_reward -158.00 epsilon 0.90
episode: 901 episode_reward -51.69 epsilon 0.90
```

Test every 100 epoch in policy gradient：

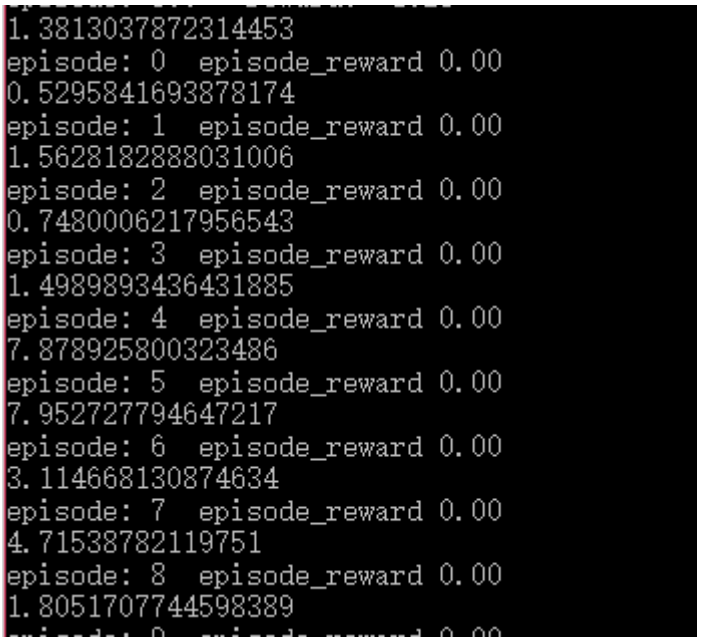
100：



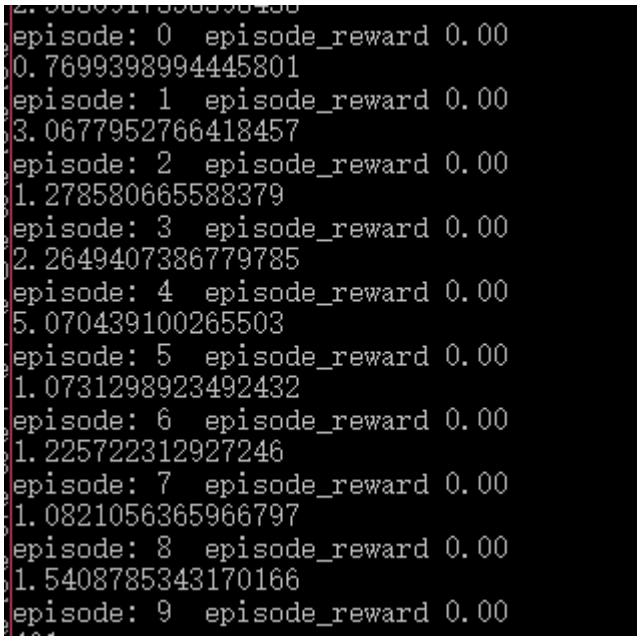
200：



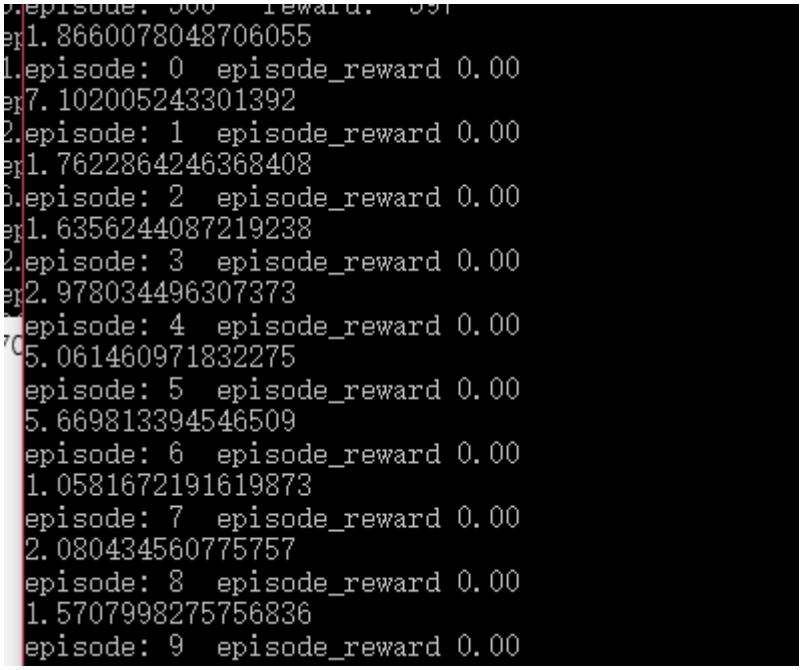
300:


1.3813037872314453
episode: 0 episode_reward 0.00
0.5295841693878174
episode: 1 episode_reward 0.00
1.5628182888031006
episode: 2 episode_reward 0.00
0.7480006217956543
episode: 3 episode_reward 0.00
1.4989893436431885
episode: 4 episode_reward 0.00
7.878925800323486
episode: 5 episode_reward 0.00
7.952727794647217
episode: 6 episode_reward 0.00
3.1146681308746334
episode: 7 episode_reward 0.00
4.71538782119751
episode: 8 episode_reward 0.00
1.8051707744598389

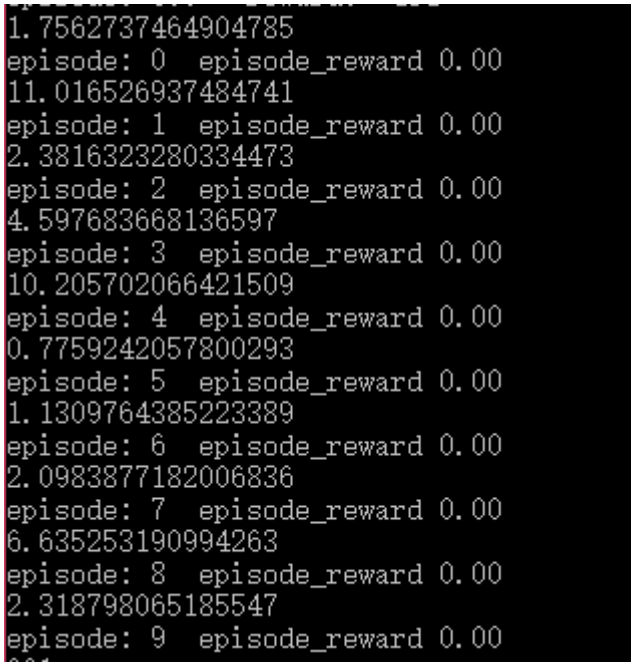
400:


episode: 0 episode_reward 0.00
0.7699398994445801
episode: 1 episode_reward 0.00
3.0677952766418457
episode: 2 episode_reward 0.00
1.278580665588379
episode: 3 episode_reward 0.00
2.2649407386779785
episode: 4 episode_reward 0.00
5.070439100265503
episode: 5 episode_reward 0.00
1.0731298923492432
episode: 6 episode_reward 0.00
1.225722312927246
episode: 7 episode_reward 0.00
1.0821056365966797
episode: 8 episode_reward 0.00
1.5408785343170166
episode: 9 episode_reward 0.00

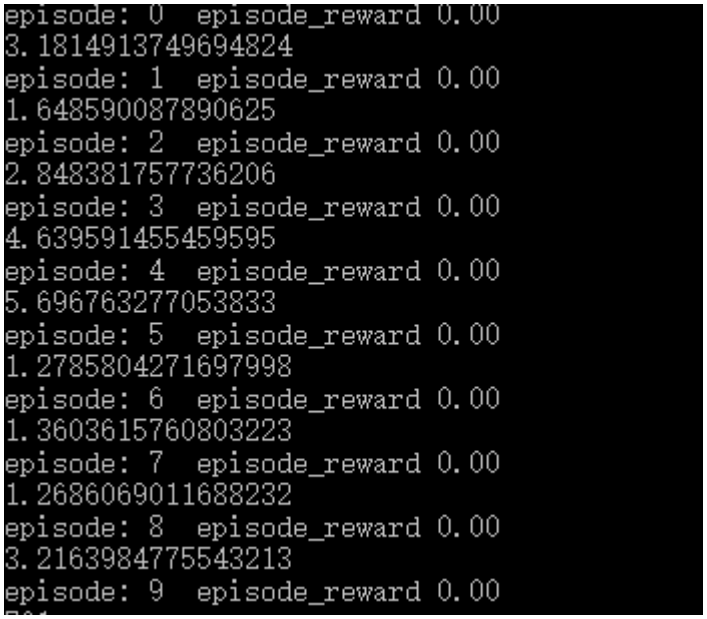
500:


1.8660078048706055
episode: 0 episode_reward 0.00
7.102005243301392
episode: 1 episode_reward 0.00
1.7622864246368408
episode: 2 episode_reward 0.00
1.6356244087219238
episode: 3 episode_reward 0.00
2.978034496307373
episode: 4 episode_reward 0.00
5.061460971832275
episode: 5 episode_reward 0.00
5.669813394546509
episode: 6 episode_reward 0.00
1.0581672191619873
episode: 7 episode_reward 0.00
2.080434560775757
episode: 8 episode_reward 0.00
1.5707998275756836
episode: 9 episode_reward 0.00

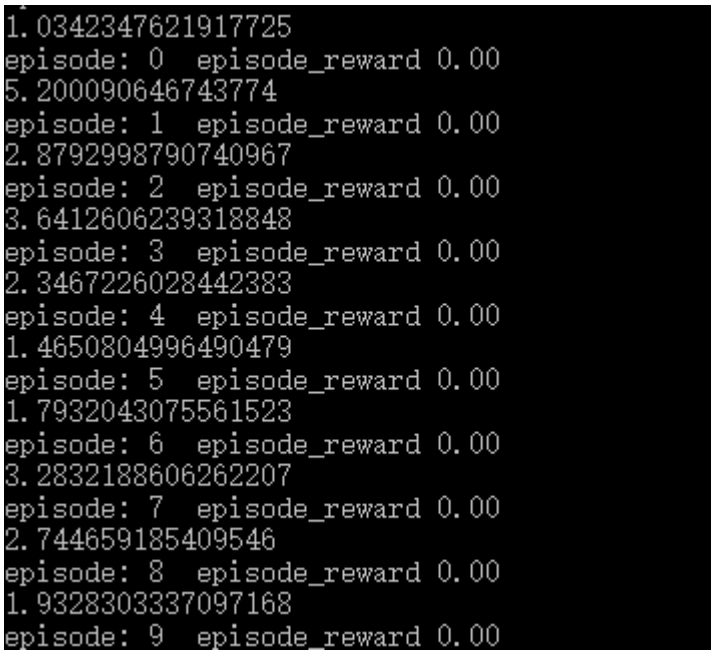
600:


1.7562737464904785
episode: 0 episode_reward 0.00
11.016526937484741
episode: 1 episode_reward 0.00
2.3816323280334473
episode: 2 episode_reward 0.00
4.597683668136597
episode: 3 episode_reward 0.00
10.2057020664215O9
episode: 4 episode_reward 0.00
0.7759242057800293
episode: 5 episode_reward 0.00
1.1309764385223389
episode: 6 episode_reward 0.00
2.0983877182006836
episode: 7 episode_reward 0.00
6.635253190994263
episode: 8 episode_reward 0.00
2.318798065185547
episode: 9 episode_reward 0.00

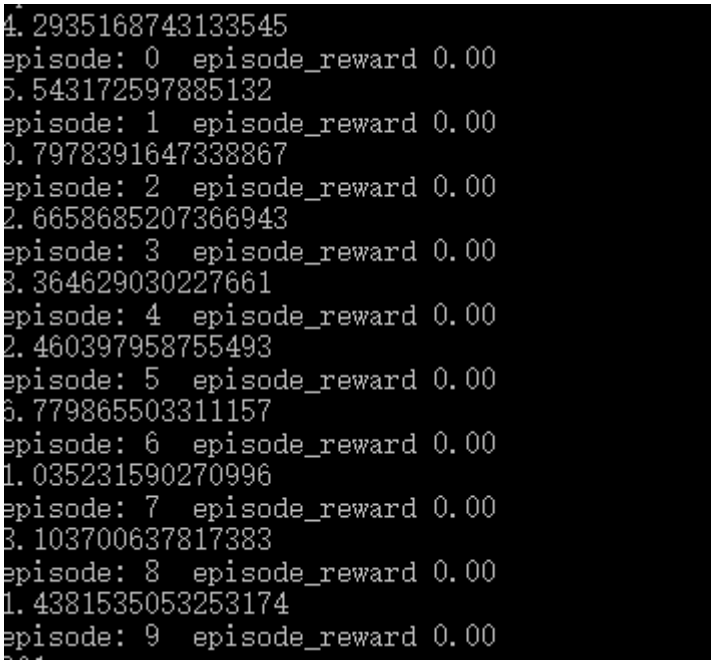
700:


episode: 0 episode_reward 0.00
3.1814913749694824
episode: 1 episode_reward 0.00
1.648590087890625
episode: 2 episode_reward 0.00
2.848381757736206
episode: 3 episode_reward 0.00
4.639591455459595
episode: 4 episode_reward 0.00
5.696763277053833
episode: 5 episode_reward 0.00
1.2785804271697998
episode: 6 episode_reward 0.00
1.3603615760803223
episode: 7 episode_reward 0.00
1.2686069011688232
episode: 8 episode_reward 0.00
3.2163984775543213
episode: 9 episode_reward 0.00

800:


1.0342347621917725
episode: 0 episode_reward 0.00
5.200090646743774
episode: 1 episode_reward 0.00
2.8792998790740967
episode: 2 episode_reward 0.00
3.6412606239318848
episode: 3 episode_reward 0.00
2.3467226028442383
episode: 4 episode_reward 0.00
1.4650804996490479
episode: 5 episode_reward 0.00
1.7932043075561523
episode: 6 episode_reward 0.00
3.2832188606262207
episode: 7 episode_reward 0.00
2.744659185409546
episode: 8 episode_reward 0.00
1.9328303337097168
episode: 9 episode_reward 0.00

900:



Policy gradient: reward time

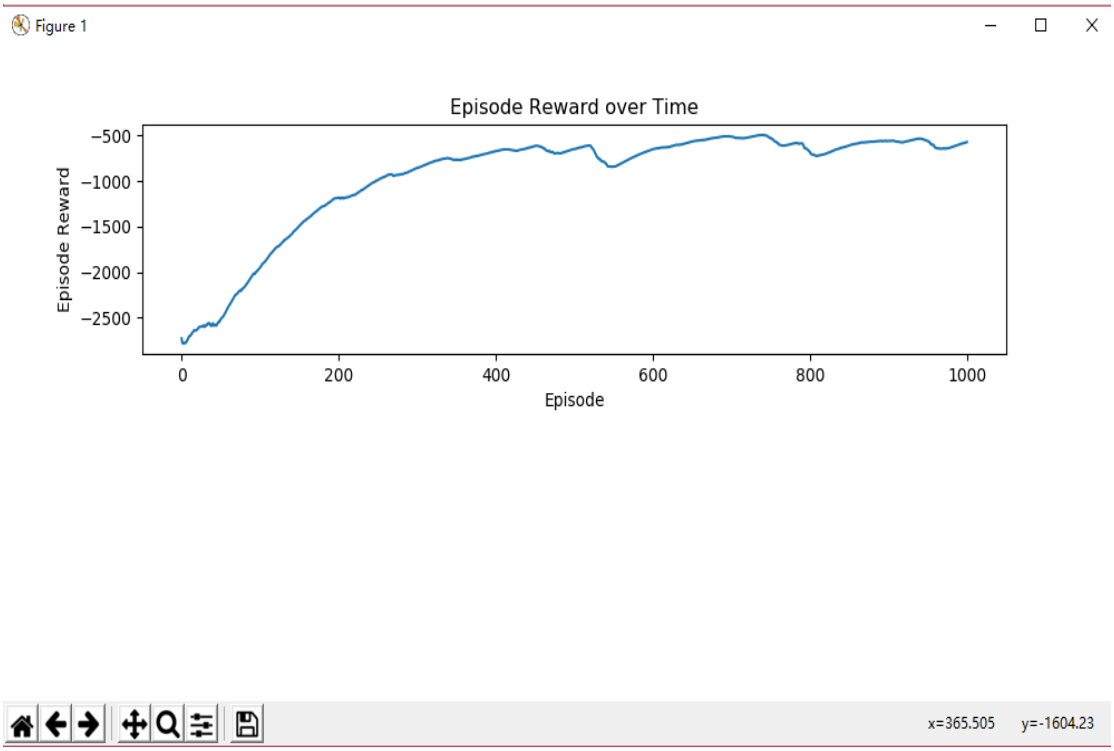


Figure 1
Episode Reward over Time
Episode Reward
−500
−1000
−1500
−2000
−2500
0
200
400
600
800
1000
Episode
x=365.505
y=-1604.23