

实验3. 强化学习实践

MG1733099, 周天烁, tianshuo.zhou@smail.nju.edu.cn

2018 年 1 月 2 日

1 综述

通过解决Gym为用户提供了一些基础强化学习环境，如CartPole-v0、MountainCar-v0和Acrobot-v1等，完成强化学习的任务，加深对强化学习的理解。

2 实验二.

通过连续空间离散化的方式，实现Q-learning算法，并用于求解CartPole-v0、MountainCar-v0和Acrobot-v1三个学习任务。由于三个任务的状态空间都是连续的，所以最直接的方式就是将状态空间等分，把每个状态映射到等分的一个个区间里。具体等分方式视任务而定，即调参过程。

为了调用统一的函数，对于3个任务，每个任务都是训练2000轮（episode），测试100轮（episode），打印结果为每个每轮测试结果及其均值和标准差。

2.1 CartPole-v0

状态空间有4个维度，等分方式为（1,1,6,5）。实验运行结果如图1所示。

```
Episode 76 finished after 2998.000000 time steps
Episode 77 finished after 3749.000000 time steps
Episode 78 finished after 3971.000000 time steps
Episode 79 finished after 6087.000000 time steps
Episode 80 finished after 2838.000000 time steps
Episode 81 finished after 5205.000000 time steps
Episode 82 finished after 3890.000000 time steps
Episode 83 finished after 4144.000000 time steps
Episode 84 finished after 2981.000000 time steps
Episode 85 finished after 3013.000000 time steps
Episode 86 finished after 3355.000000 time steps
Episode 87 finished after 7060.000000 time steps
Episode 88 finished after 4020.000000 time steps
Episode 89 finished after 7974.000000 time steps
Episode 90 finished after 5512.000000 time steps
Episode 91 finished after 10427.000000 time steps
Episode 92 finished after 5063.000000 time steps
Episode 95 finished after 7547.000000 time steps
Episode 96 finished after 6596.000000 time steps
Episode 97 finished after 7835.000000 time steps
Episode 98 finished after 14741.000000 time steps
Episode 99 finished after 5168.000000 time steps
mean: 5739.460674 ; var: 3458.303172
```

图 1: Q-learning: CartPole-v0

2.2 MountainCar-v0

状态空间有2个维度，等分方式为 (20,20)。实验运行结果如图2所示。

```
Episode 78 finished after 113.000000 time steps
Episode 79 finished after 168.000000 time steps
Episode 80 finished after 115.000000 time steps
Episode 81 finished after 111.000000 time steps
Episode 82 finished after 164.000000 time steps
Episode 83 finished after 115.000000 time steps
Episode 84 finished after 178.000000 time steps
Episode 85 finished after 156.000000 time steps
Episode 86 finished after 158.000000 time steps
Episode 87 finished after 165.000000 time steps
Episode 88 finished after 165.000000 time steps
Episode 89 finished after 154.000000 time steps
Episode 90 finished after 154.000000 time steps
Episode 91 finished after 113.000000 time steps
Episode 92 finished after 155.000000 time steps
Episode 93 finished after 156.000000 time steps
Episode 94 finished after 156.000000 time steps
Episode 95 finished after 157.000000 time steps
Episode 96 finished after 155.000000 time steps
Episode 97 finished after 152.000000 time steps
Episode 98 finished after 156.000000 time steps
Episode 99 finished after 167.000000 time steps
mean: 143.170000 ; var: 21.929913
```

图 2: Q-learning: MountainCar-v0

2.3 Acrobot-v1

状态空间有6个维度，等分方式为 (6,6,6,6,6,6)。实验运行结果如图3所示。

```
Episode 78 finished after 231.000000 time steps
Episode 79 finished after 186.000000 time steps
Episode 80 finished after 156.000000 time steps
Episode 81 finished after 303.000000 time steps
Episode 82 finished after 196.000000 time steps
Episode 83 finished after 172.000000 time steps
Episode 84 finished after 166.000000 time steps
Episode 85 finished after 286.000000 time steps
Episode 86 finished after 230.000000 time steps
Episode 87 finished after 197.000000 time steps
Episode 88 finished after 268.000000 time steps
Episode 89 finished after 187.000000 time steps
Episode 90 finished after 228.000000 time steps
Episode 91 finished after 224.000000 time steps
Episode 92 finished after 263.000000 time steps
Episode 93 finished after 172.000000 time steps
Episode 94 finished after 295.000000 time steps
Episode 95 finished after 162.000000 time steps
Episode 96 finished after 164.000000 time steps
Episode 97 finished after 207.000000 time steps
Episode 98 finished after 206.000000 time steps
Episode 99 finished after 257.000000 time steps
mean: 220.030000 ; var: 54.038219
```

图 3: Q-learning: Acrobot-v1

3 实验三.

利用Deep Q-learning(DQN)算法直接在连续的空间中进行学习，完成CartPole-v0、MountainCar-v0和Acrobot-v1这三个任务，在实验三我们直接在连续空间中学习。DQN 的算法主要是实现一个Q 函数网络。本实验采用的Q值网络为Multi-layer Perceptron (MLP)，通过学习一组深度网络参数使Q函数网络的输出近似于真实的Q值。

3.1 CartPole-v0

基本参数和实验二一致，MLP含有一个输入层，一个输出层和两个隐层。训练过程中损失(loss)变化如图4所示,累计奖赏(sum of reward)变化如图5所示。实验运行结果如图6所示。

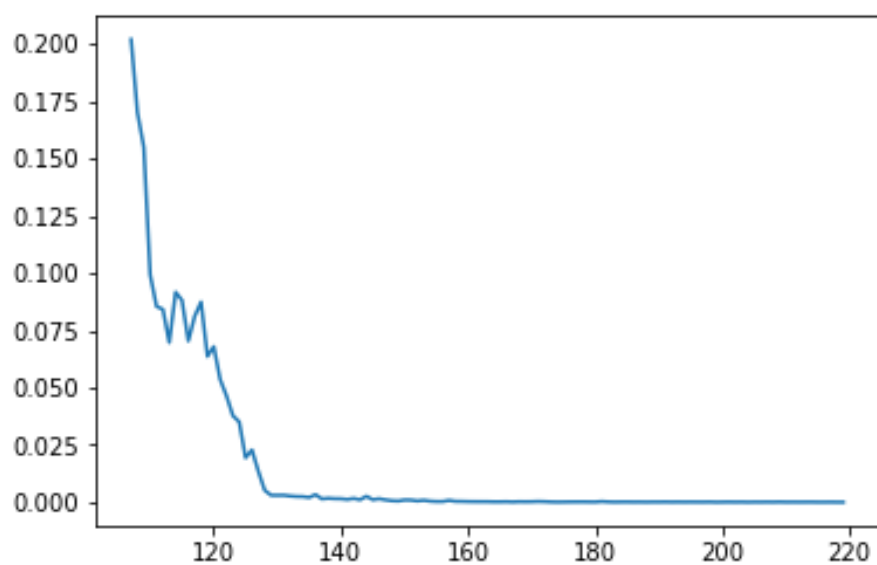


图 4: DQN: CartPole-v0 loss曲线

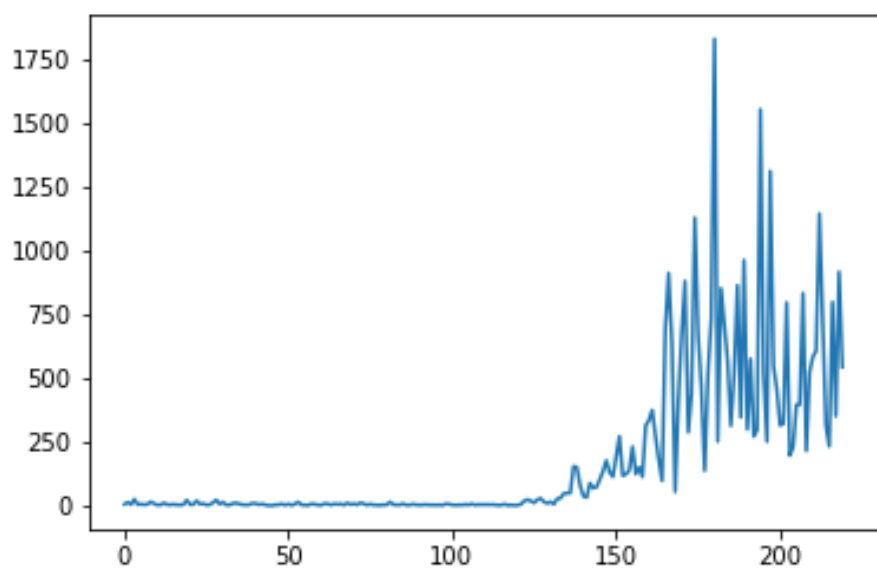


图 5: DQN: CartPole-v0 sum of reward曲线

```
Episode 76 finished after 2094.000000 time steps
Episode 77 terminated after 19999.000000 MAX steps
Episode 78 terminated after 19999.000000 MAX steps
Episode 79 terminated after 19999.000000 MAX steps
Episode 80 terminated after 19999.000000 MAX steps
Episode 81 finished after 1838.000000 time steps
Episode 82 terminated after 19999.000000 MAX steps
Episode 83 terminated after 19999.000000 MAX steps
Episode 84 terminated after 19999.000000 MAX steps
Episode 85 terminated after 19999.000000 MAX steps
Episode 86 terminated after 19999.000000 MAX steps
Episode 87 terminated after 19999.000000 MAX steps
Episode 88 terminated after 19999.000000 MAX steps
Episode 89 terminated after 19999.000000 MAX steps
Episode 90 terminated after 19999.000000 MAX steps
Episode 91 terminated after 19999.000000 MAX steps
Episode 92 terminated after 19999.000000 MAX steps
Episode 93 terminated after 19999.000000 MAX steps
Episode 94 terminated after 19999.000000 MAX steps
Episode 95 terminated after 19999.000000 MAX steps
Episode 96 finished after 4630.000000 time steps
Episode 97 terminated after 19999.000000 MAX steps
Episode 98 terminated after 19999.000000 MAX steps
Episode 99 terminated after 19999.000000 MAX steps
mean: 18765.290000 ; std: 4505.866073
```

图 6: DQN: CartPole-v0

3.2 MountainCar-v0

基本参数和实验二一致，MLP含有一个输入层，一个输出层和两个隐层。训练过程中损失(loss)变化如图7所示,累计奖赏(sum of reward)变化如图8所示。实验运行结果如图9所示。

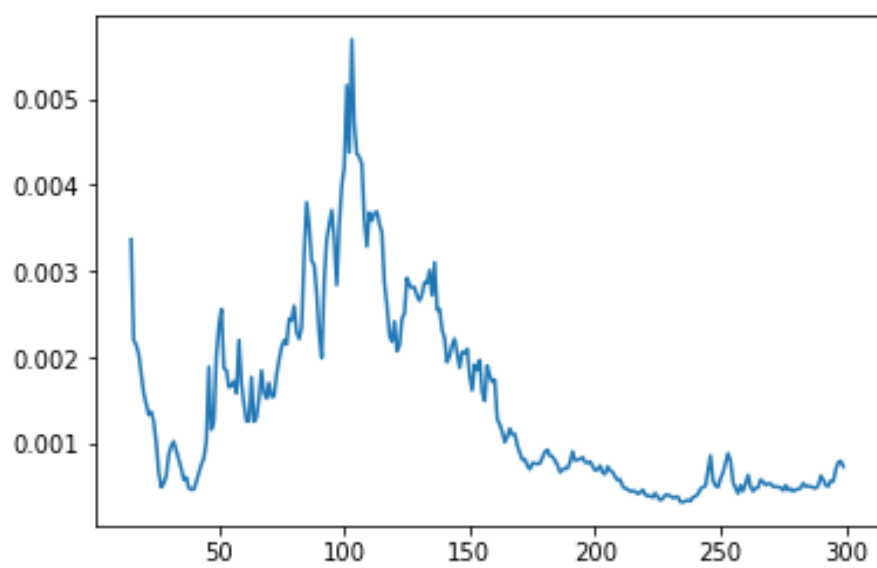


图 7: DQN: MountainCar-v0 loss曲线

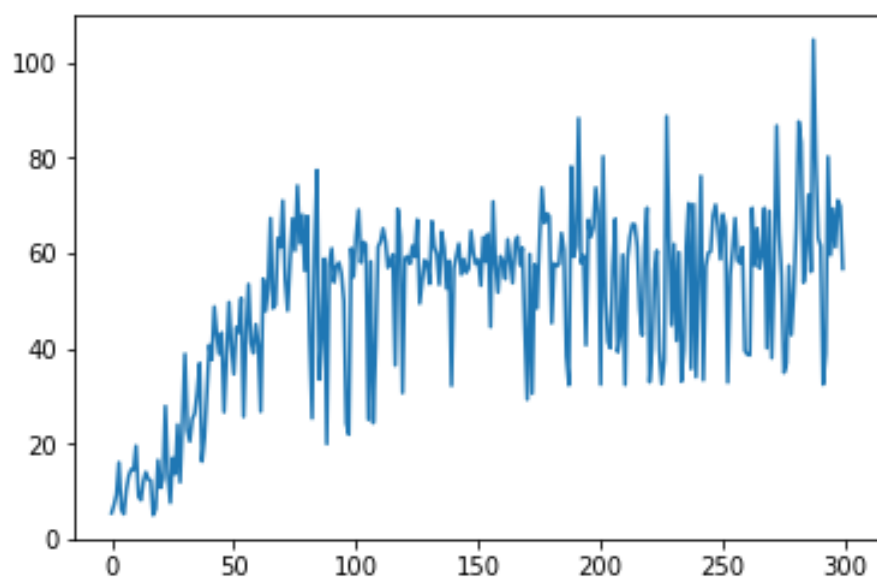


图 8: DQN: MountainCar-v0 sum of reward曲线

```
Episode 79 finished after 176.000000 time steps
Episode 80 finished after 189.000000 time steps
Episode 81 finished after 175.000000 time steps
Episode 82 finished after 102.000000 time steps
Episode 83 finished after 175.000000 time steps
Episode 84 finished after 176.000000 time steps
Episode 85 finished after 177.000000 time steps
Episode 86 finished after 176.000000 time steps
Episode 87 finished after 145.000000 time steps
Episode 88 finished after 186.000000 time steps
Episode 89 finished after 198.000000 time steps
Episode 90 finished after 188.000000 time steps
Episode 91 finished after 178.000000 time steps
Episode 92 finished after 176.000000 time steps
Episode 93 finished after 175.000000 time steps
Episode 94 finished after 177.000000 time steps
Episode 95 finished after 175.000000 time steps
Episode 96 finished after 175.000000 time steps
Episode 97 finished after 189.000000 time steps
Episode 98 finished after 175.000000 time steps
Episode 99 finished after 177.000000 time steps
mean: 163.040000 ; std: 33.387698
```

图 9: DQN: MountainCar-v0

3.3 Acrobot-v1

基本参数和实验二一致，MLP含有一个输入层，一个输出层和两个隐层。训练过程中损失(loss)变化如图10所示,累计奖赏(sum of reward)变化如图11所示。实验运行结果如图12所示。

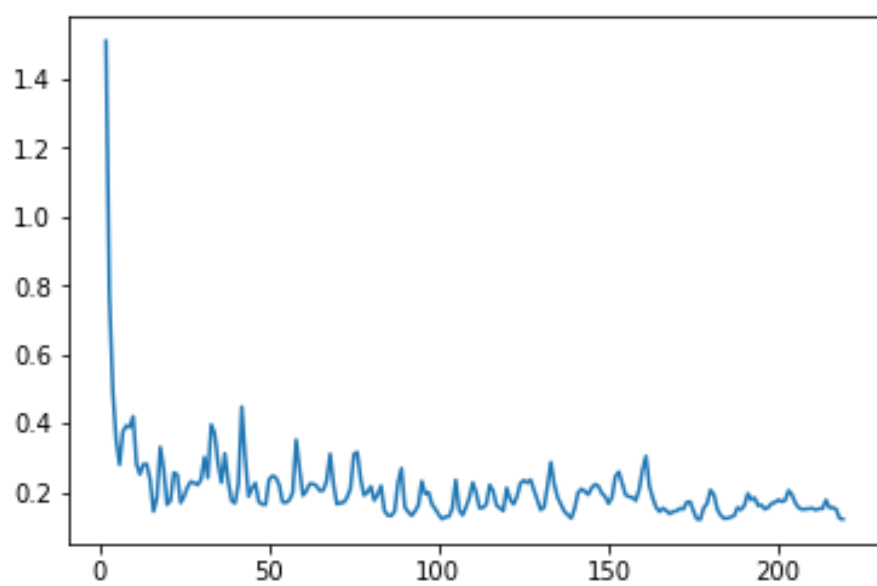


图 10: DQN: Acrobot-v1 loss曲线

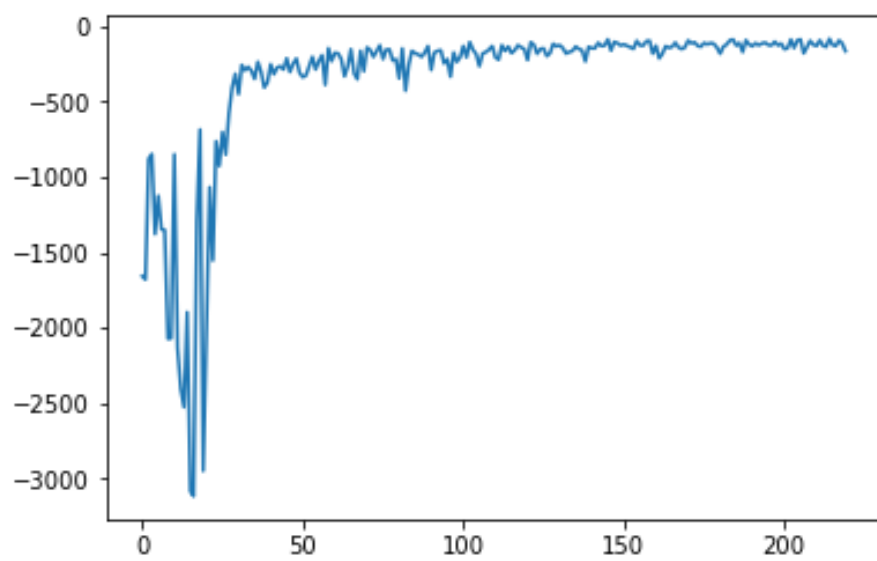


图 11: DQN: Acrobot-v1 sum of reward曲线

```
Episode 77 finished after 106.000000 time steps
Episode 78 finished after 93.000000 time steps
Episode 79 finished after 113.000000 time steps
Episode 80 finished after 108.000000 time steps
Episode 81 finished after 113.000000 time steps
Episode 82 finished after 93.000000 time steps
Episode 83 finished after 107.000000 time steps
Episode 84 finished after 98.000000 time steps
Episode 85 finished after 108.000000 time steps
Episode 86 finished after 106.000000 time steps
Episode 87 finished after 106.000000 time steps
Episode 88 finished after 97.000000 time steps
Episode 89 finished after 118.000000 time steps
Episode 90 finished after 91.000000 time steps
Episode 91 finished after 108.000000 time steps
Episode 92 finished after 99.000000 time steps
Episode 93 finished after 95.000000 time steps
Episode 94 finished after 102.000000 time steps
Episode 95 finished after 99.000000 time steps
Episode 96 finished after 98.000000 time steps
Episode 97 finished after 106.000000 time steps
Episode 98 finished after 78.000000 time steps
Episode 99 finished after 103.000000 time steps
mean: 104.000000 ; std: 11.581882
```

图 12: DQN: Acrobot-v1

4 实验四.

对实验三的Deep Q-learning(DQN)算法进行改进。唯一的区别是Q值网络函数的更新方式不同，其它参数基本和实验三相同。

4.1 CartPole-v0

训练过程中损失(loss)变化如图13所示,累计奖赏(sum of reward)变化如图14所示。实验运行结果如图15所示。

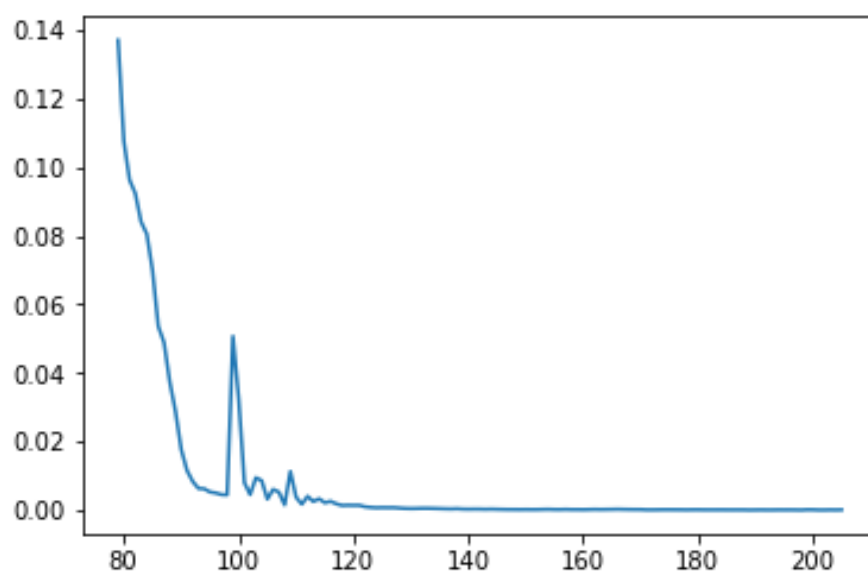


图 13: MyImprovedDQN: CartPole-v0 loss曲线

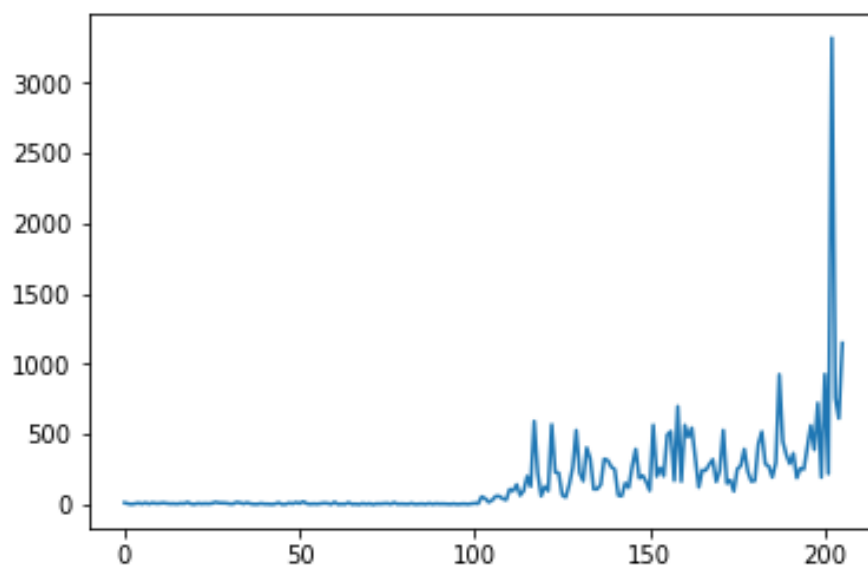


图 14: MyImprovedDQN: CartPole-v0 sum of reward曲线

```
Episode 80 terminated after 1999.000000 MAX steps
Episode 81 terminated after 1999.000000 MAX steps
Episode 82 terminated after 1999.000000 MAX steps
Episode 83 terminated after 1999.000000 MAX steps
Episode 84 terminated after 1999.000000 MAX steps
Episode 85 terminated after 1999.000000 MAX steps
Episode 86 terminated after 1999.000000 MAX steps
Episode 87 terminated after 1999.000000 MAX steps
Episode 88 terminated after 1999.000000 MAX steps
Episode 89 terminated after 1999.000000 MAX steps
Episode 90 terminated after 1999.000000 MAX steps
Episode 91 terminated after 1999.000000 MAX steps
Episode 92 terminated after 1999.000000 MAX steps
Episode 93 terminated after 1999.000000 MAX steps
Episode 94 terminated after 1999.000000 MAX steps
Episode 95 terminated after 1999.000000 MAX steps
Episode 96 terminated after 1999.000000 MAX steps
Episode 97 terminated after 1999.000000 MAX steps
Episode 98 terminated after 1999.000000 MAX steps
Episode 99 terminated after 1999.000000 MAX steps
mean: 1999.000000 ; std: 0.000000
```

图 15: MyImprovedDQN: CartPole-v0

4.2 MountainCar-v0

训练过程中损失(loss)变化如图16所示,累计奖赏(sum of reward)变化如图17所示。实验运行结果如图18所示。

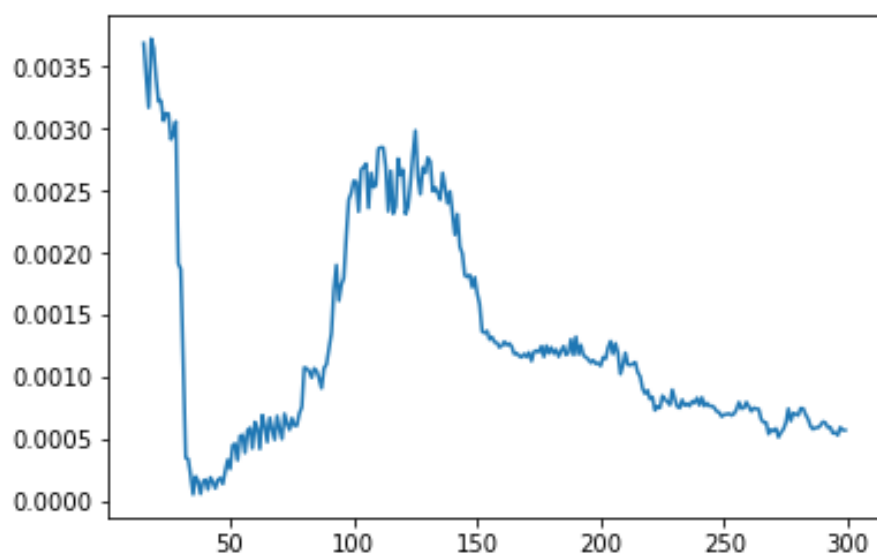


图 16: MyImprovedDQN: MountainCar-v0 loss曲线

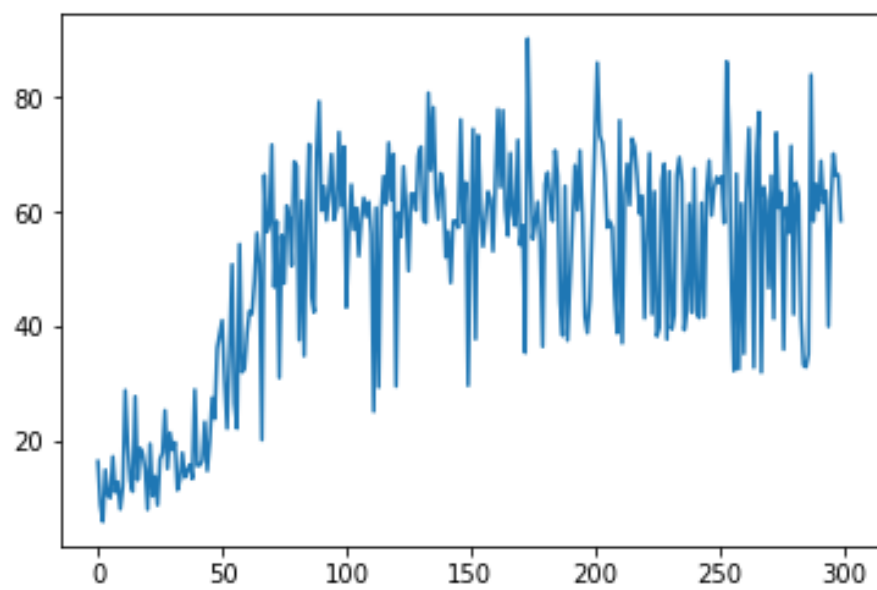


图 17: MyImprovedDQN: MountainCar-v0 sum of reward曲线

```
Episode 80 finished after 212.000000 time steps
Episode 81 finished after 174.000000 time steps
Episode 82 finished after 92.000000 time steps
Episode 83 finished after 172.000000 time steps
Episode 84 finished after 171.000000 time steps
Episode 85 finished after 94.000000 time steps
Episode 86 finished after 98.000000 time steps
Episode 87 finished after 90.000000 time steps
Episode 88 finished after 100.000000 time steps
Episode 89 finished after 171.000000 time steps
Episode 90 finished after 93.000000 time steps
Episode 91 finished after 89.000000 time steps
Episode 92 finished after 94.000000 time steps
Episode 93 finished after 173.000000 time steps
Episode 94 finished after 93.000000 time steps
Episode 95 finished after 173.000000 time steps
Episode 96 finished after 117.000000 time steps
Episode 97 finished after 89.000000 time steps
Episode 98 finished after 171.000000 time steps
Episode 99 finished after 194.000000 time steps
mean: 137.820000 ; std: 43.567965
```

图 18: MyImprovedDQN: MountainCar-v0

4.3 Acrobot-v1

训练过程中损失(loss)变化如图19所示,累计奖赏(sum of reward)变化如图20所示。实验运行结果如图21所示。

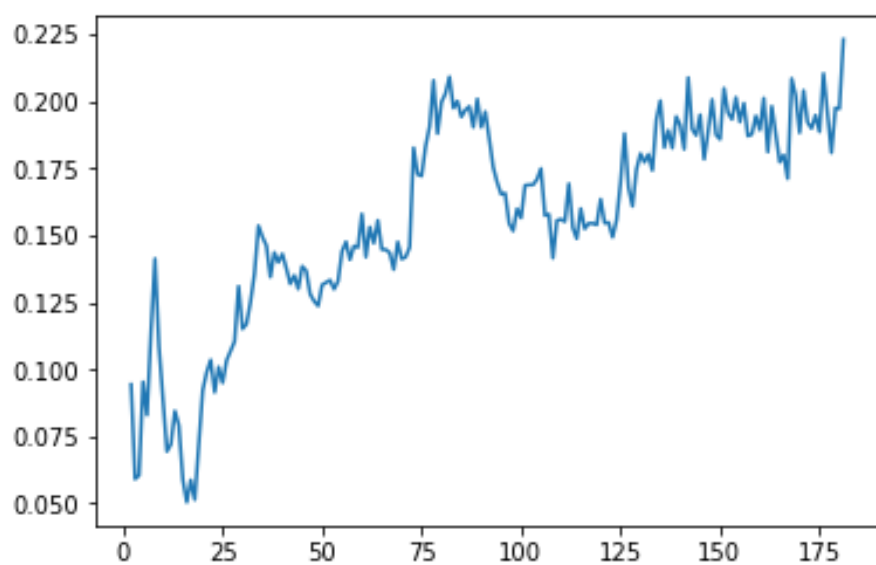


图 19: MyImprovedDQN: Acrobot-v1 loss曲线

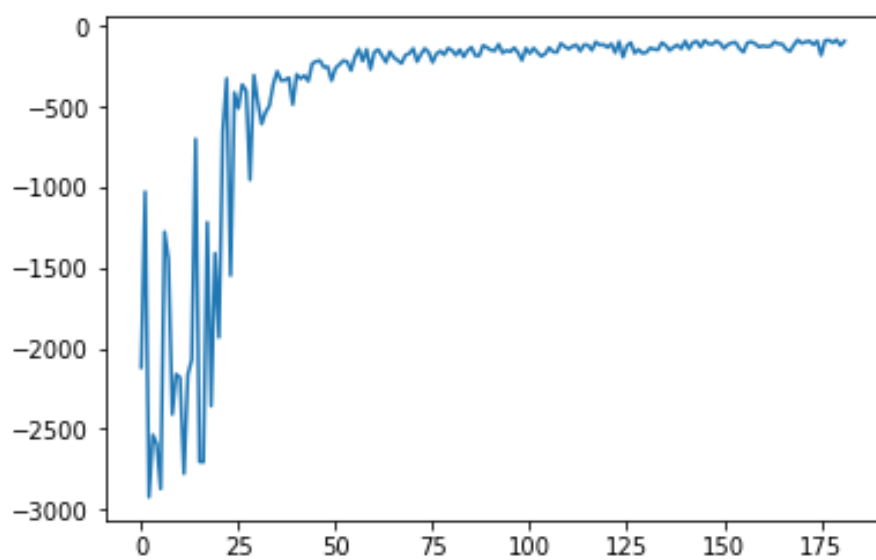


图 20: MyImprovedDQN: Acrobot-v1 sum of reward曲线

```
Episode 80 finished after 112.000000 time steps
Episode 81 finished after 131.000000 time steps
Episode 82 finished after 85.000000 time steps
Episode 83 finished after 130.000000 time steps
Episode 84 finished after 91.000000 time steps
Episode 85 finished after 82.000000 time steps
Episode 86 finished after 79.000000 time steps
Episode 87 finished after 93.000000 time steps
Episode 88 finished after 110.000000 time steps
Episode 89 finished after 225.000000 time steps
Episode 90 finished after 126.000000 time steps
Episode 91 finished after 144.000000 time steps
Episode 92 finished after 150.000000 time steps
Episode 93 finished after 136.000000 time steps
Episode 94 finished after 140.000000 time steps
Episode 95 finished after 122.000000 time steps
Episode 96 finished after 97.000000 time steps
Episode 97 finished after 117.000000 time steps
Episode 98 finished after 65.000000 time steps
Episode 99 finished after 119.000000 time steps
mean: 105.940000 ; std: 28.533076
```

图 21: MyImprovedDQN: Acrobot-v1

5 调参过程.

本部分按照介绍各个实验具体的调参过程中的问题和解决方式。

5.1 实验二.MyQlearning

主要问题就是把各个连续状态空间划分为合适的区间。本实验采用的是等分的方式，未修改奖赏（reward）。因此主要方法就是比较笨的尝试几组参数，然后选择一个较好的，因此两个任务分别是（20,20）和（6,6,6,6,6,6）。比较奇怪的四第一个任务划分（1,1,6,5），因为发现前两个维度的划分会使实验效果下降，而最后一个维度取5和6差别很大，可能是因为区间划分的奇偶性以及状态的数量对该实验的收敛性有所影响。

5.2 实验三.MyDQN

主要问题是发现如果不修改奖赏算法很难收敛，或者训练过程中表现收敛但是测试时发现和没有训练没什么区别（比如任务三）。考虑到网络的结构为全连接层构成的MLP，各层均为线性，因此感觉上输入的状态以及输出的奖励都应该是连续的能反应实际情景的变

量，因此对于三个任务，都修改了每一步的奖赏（reward），对于任务三，还重新定义了状态空间。具体如下：

- 任务一的奖赏设为cartpole的角度和位置偏离中心的程度。
- mountaincar设置为其小车相对底部的高度，高度越大，奖赏越大。
- 通过阅读任务三的源代码，发现其状态空间各的分别是杆件一的角度、杆件二相对于杆件一的角度、两者的角速度。其终止条件是两个杆件的高度之和大于1，因此修改奖励为两个杆件的高度之和，并输入状态为两个杆件的绝对角度的正弦和余弦值。

5.3 实验四.MyImprovedDQN

和实验三参数基本一致。不同的是多了一个网络更新参数，在此统一设置为300。

6 实验对比.

- 从训练程度上，实验二的训练轮数基本在两千步到三千步，而实验三和四的训练轮数基本在两百步到三百步。
- 从训练结果上，有实验运行结果截图可以看出实验三四的效果明显好于实验二。实验三和四的差别不大，实验四略微好点。
- 从稳定性上，实验二的运行非常稳定，结果基本可以复现；实验三和四的网络训练存在很多不稳定性训练，可能需要运行几次才能复现报告的结果。