A4 – Policy Gradients

Q1 – Cartpole V1

1. VPG

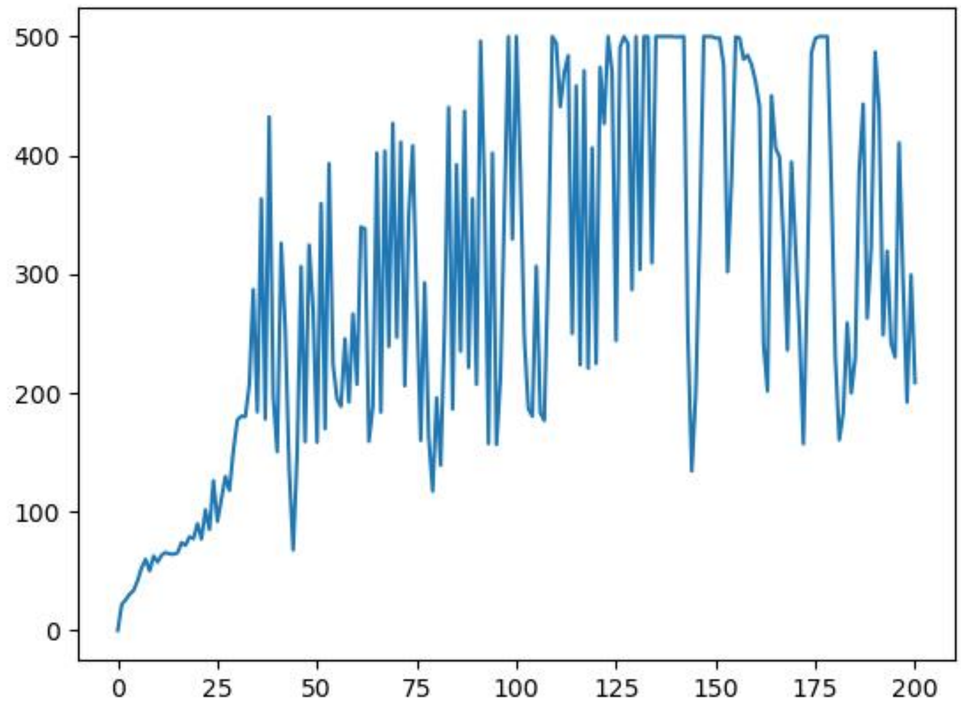

Figure 1: Average total reward per iteration for q1.1

2. VPG using a gradient only dependent on future rewards

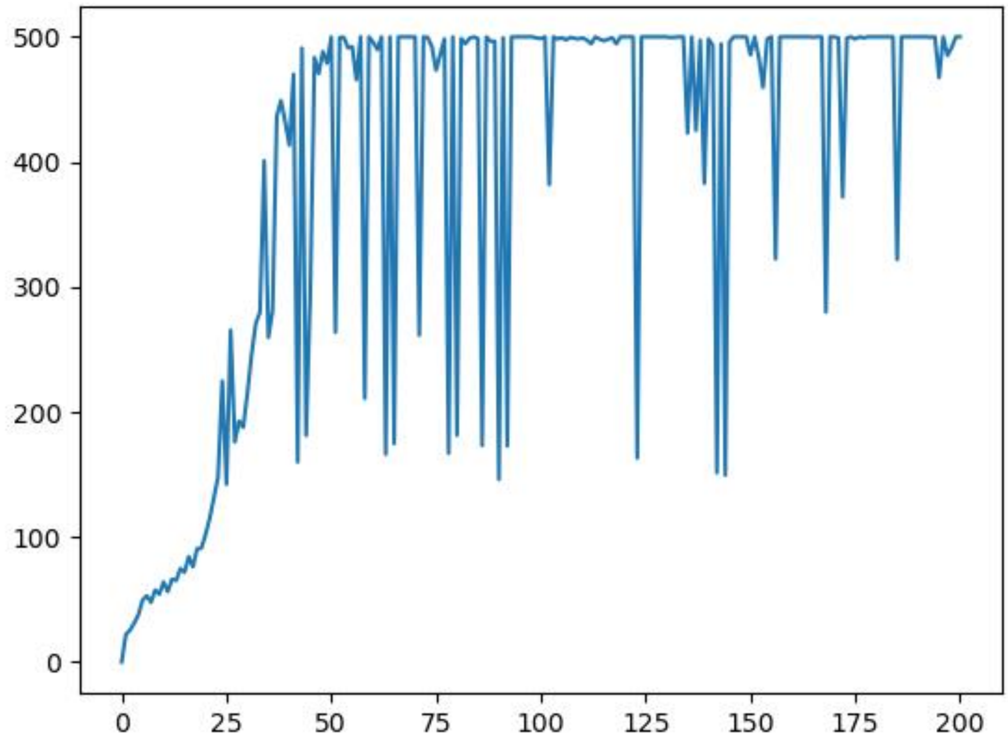

Figure 2: Average total reward per iteration for q1.2

Comparison with Figure 1: Figure 2 shows better results in comparison to figure 1 as it is more stably and consistently hitting the maximum possible rewards each iteration. But there is still lots of variance in both figures.

3. VPG with modified gradient and add baseline

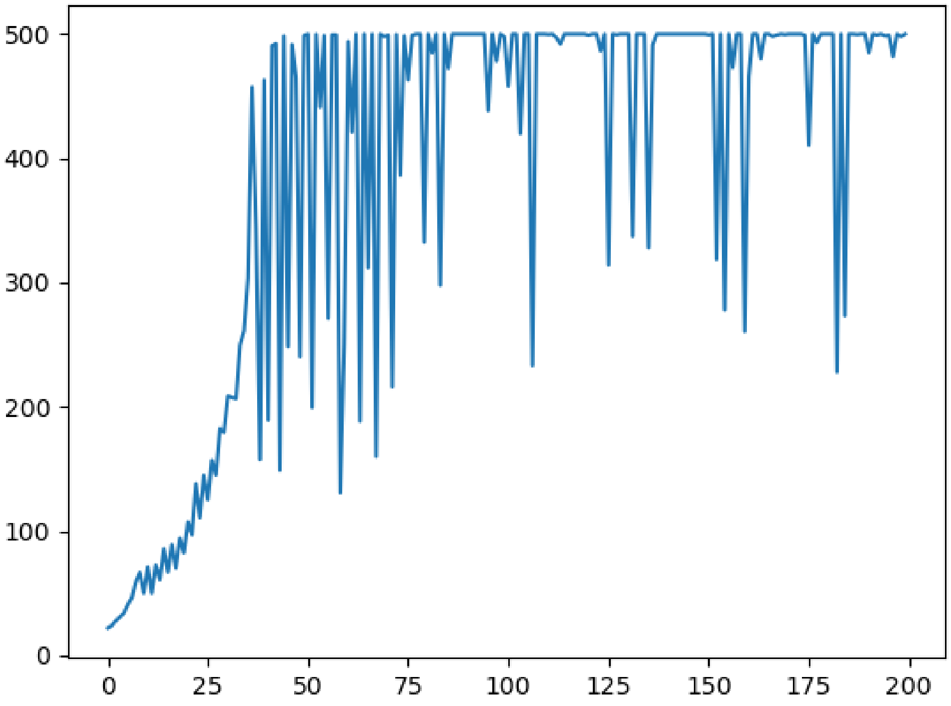

Figure 3: Average total reward per iteration for q1.3

Comparison with Figure 1, 2 and 3: These results are slightly better than the previous results from q1.1 and q1.2. This is because q1.3 reaches the first 500 slightly faster and has less variance in general. There might have been some bad runs, because technically, there should be even less variance (see figure 5 and 6 below).

4. VPG changing number of episodes per iteration

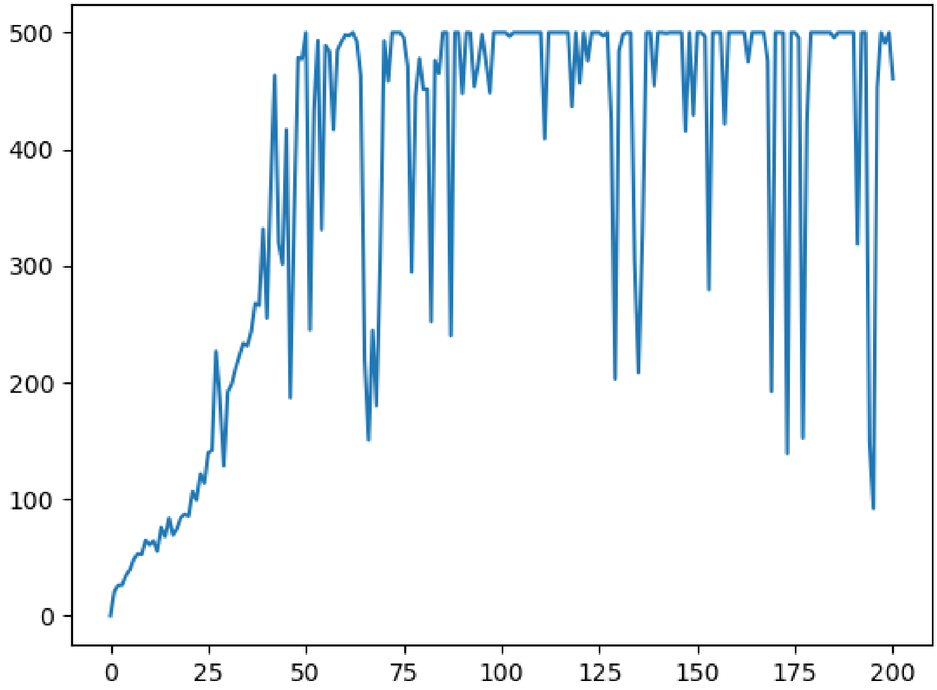

Figure 4: Average total reward per iteration for q1.3 using 100 episodes

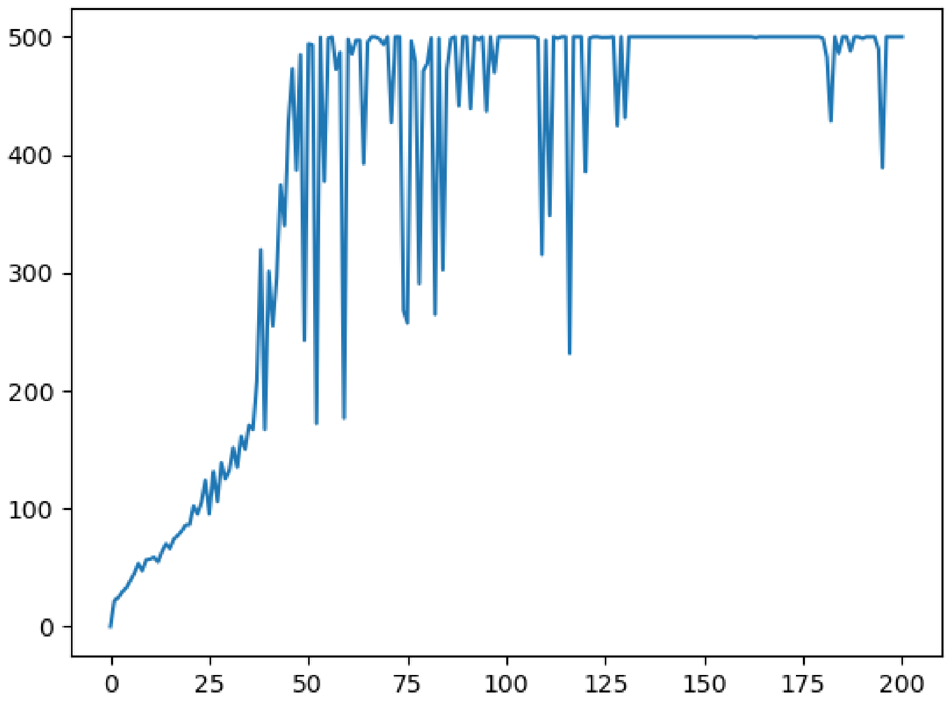

Figure 5: Average total reward per iteration for q1.3 using 300 episodes

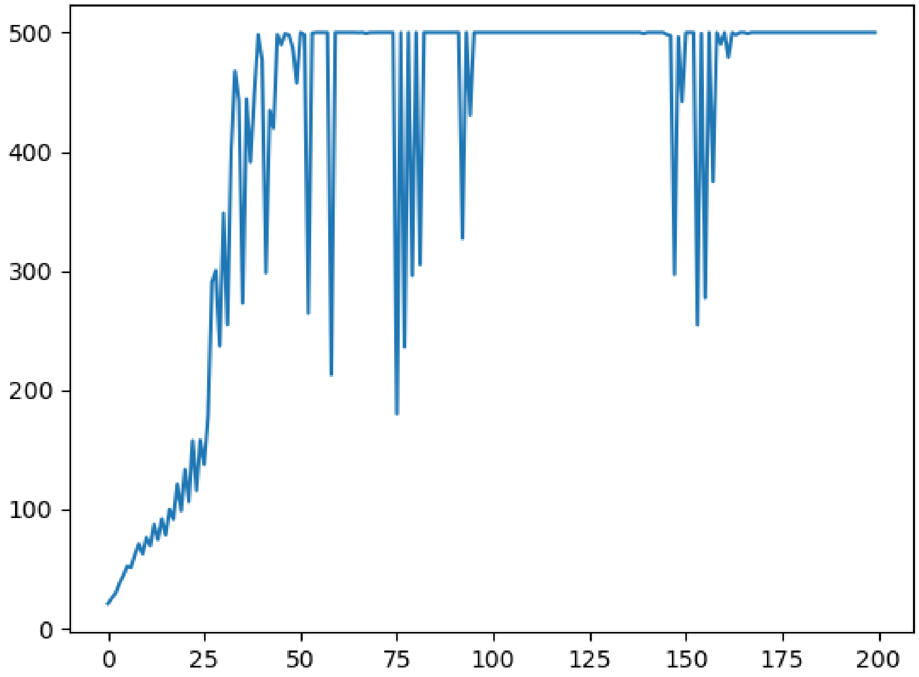

Figure 6: Average total reward per iteration for q1.3 using 1000 episodes

Comparison between using different number of episodes per iteration:

As seen in figure 4, 5 and 6: the larger number of episodes means a lower variance and more consistently hitting the maximum (500) rewards. Therefore, increasing episode count also improves training.

Q2 – 2 Link Arm

The q1.3 algorithm was modified and used to train the arm. Different number of episodes were tested until failure (i.e.: average reward per iteration does not stabilize around –3.5), as shown below:

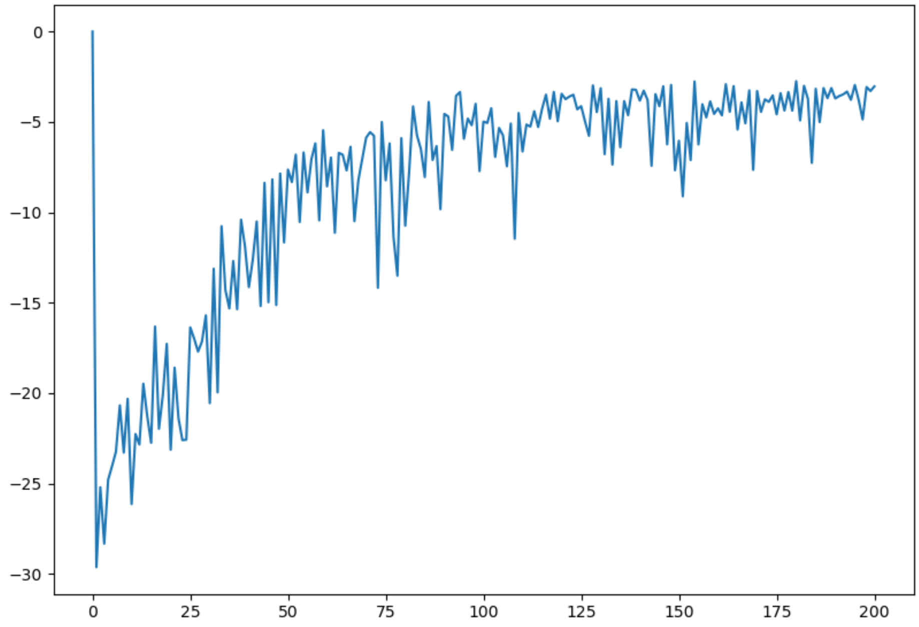

Figure 7: 100 episodes

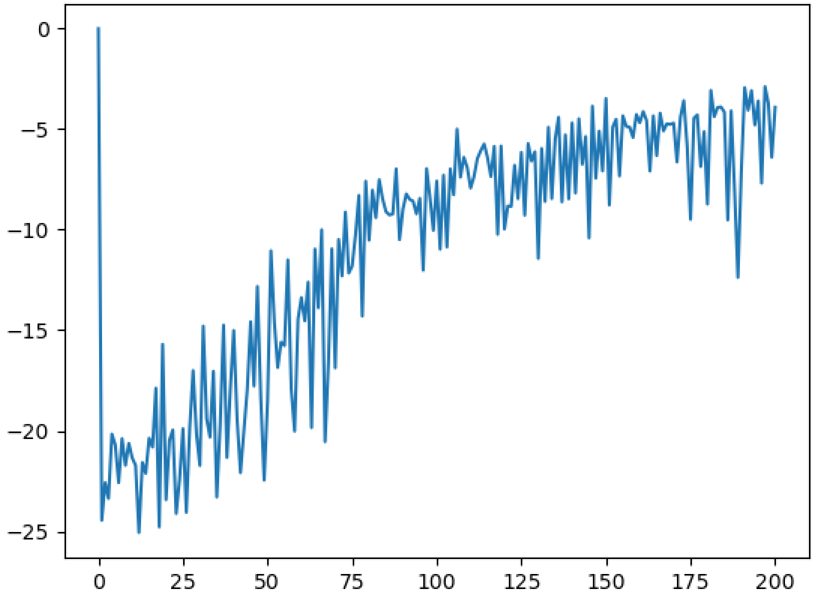

Figure 8: 80 episodes

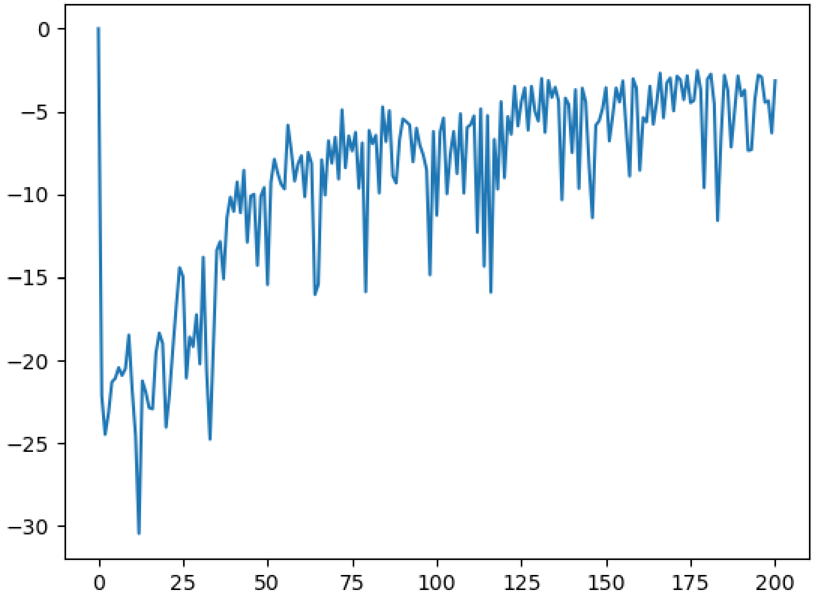

Figure 9: 50episodes

The model was tested with 30 episodes and failed. Thus, the model using 50 episodes was used to create the GIF (see video in uploaded zipfile). Time.sleep(1) was used to see well each of the actions performed by the arm for the episode tested.