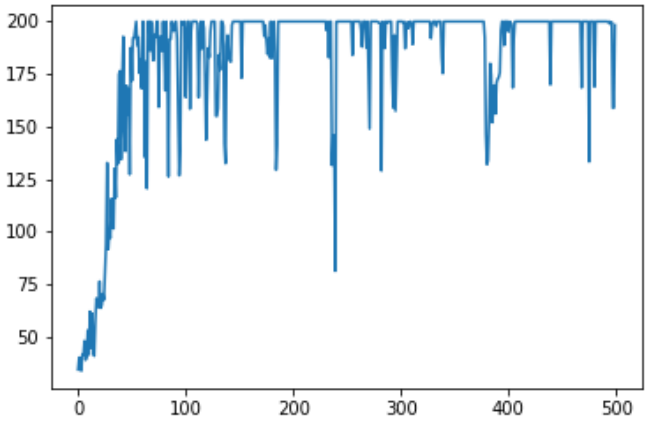
Reinforcement Learning HW-2

Name: Li, Jieda
NetID: jlg7773

Problem 1: Cart-Pole Game

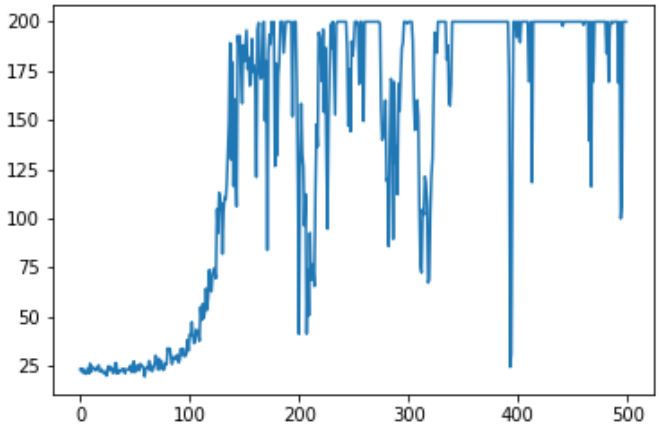
https://github.com/jiedali/reinforcement_learning_jieda_li/blob/master/hw2/cart_pole/cart_pole_baseline.ipynb

I implemented the Policy Gradient algorithm with a Value Function (approximated by a Neuron Network) as baseline. The average episode reward vs training epoch is shown below. (Training epoch: 500)



With Baseline

As a comparison, I did the policy gradient without baseline and clearly the convergence is much slower and the variance is much bigger.



No Baseline

Problem 2: Pong Game

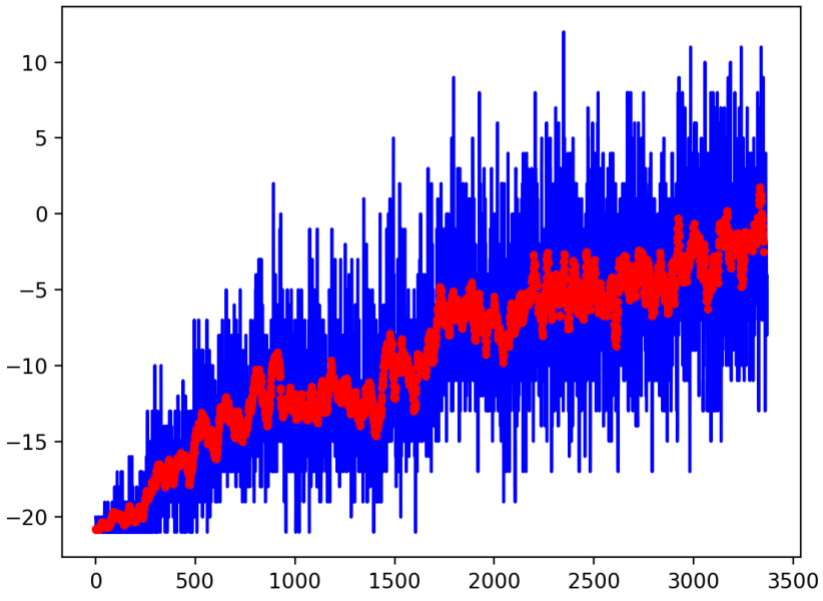
Code:

https://github.com/jiedali/reinforcement_learning_jieda_li/blob/master/hw2/pong/pong_FINAL_baseline_best_params.ipynb

For this game, I implemented policy gradient with a Value function baseline approximated by a neuron network. Both policy network and value network are using a CNN structure, for policy network, the output is 2 logits for the 2 possible actions and for value network, the output is a single value.

I experimented with a few different learning rate and random seed, the final selected learning rate is: 0.0001, tf seed:1, gym seed=666. (detailed experiment record in the appendix)

The average episode reward vs training epoch over 48 hrs training wall time is shown below. (One minibatch in my implementation is 1 episode, the red curve is a moving average of 20 episode). The best average episode reward is around **0**. The best single episode reward is ***13***.



Appendix: Pong Game parameters tuning

Here is a list of parameters I tried
Run1: lr=0.002, seed=325, can't learn (appendix figure 1)
Run2: lr=0.0005, seed = 666, can't learn (appendix figure 2)
Run3: lr=0.0001, seed=666, yielded learning (CHOSEN parameters)
Run4: lr=0.0005, seed=123, can't learn (at end of 1162 epoch, average episode reward is still -21)

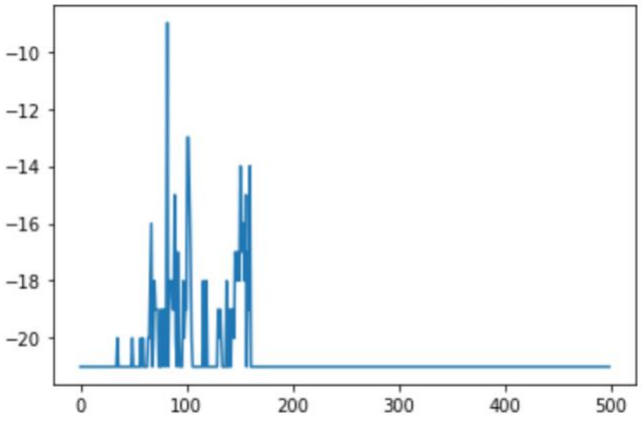


Figure 1: Run 1 results (500 epoch)

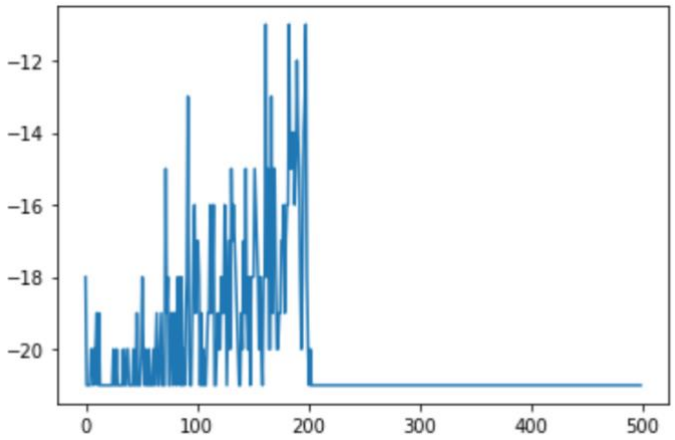


Figure 2: Run 2 results (500 epoch)