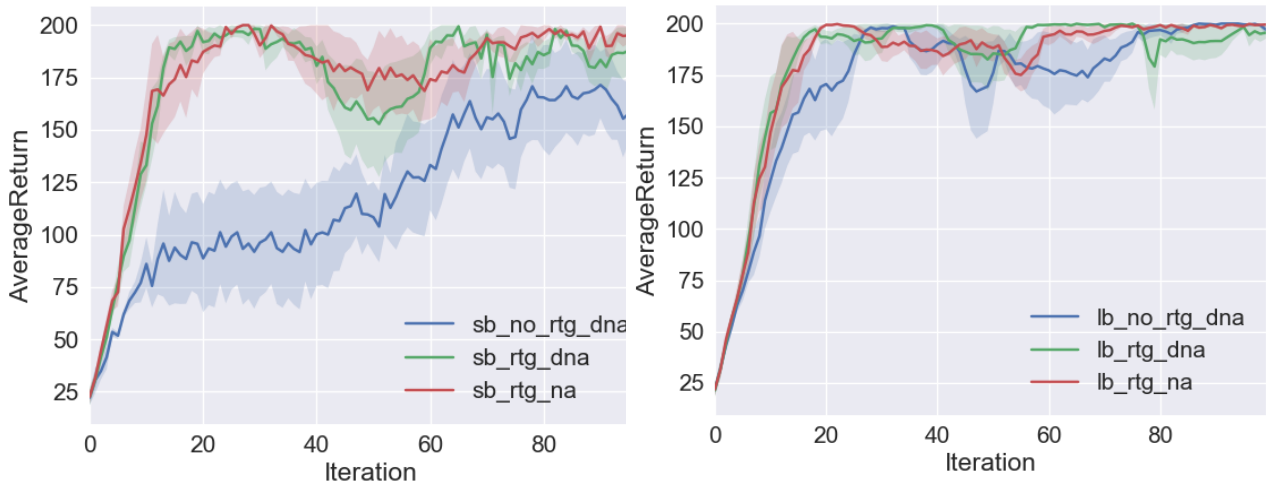
Deep RL Assignment 2 Amin Ghafari Zeydabadi 25911193

Section 4, Implement Policy Gradient



Left: small batch size for Cart, Right: large batch size for Cart

Questions:

- The one with reward to go and advantage centering preform the best.

- The advantage centering helped to decrease the variance of the rewards.

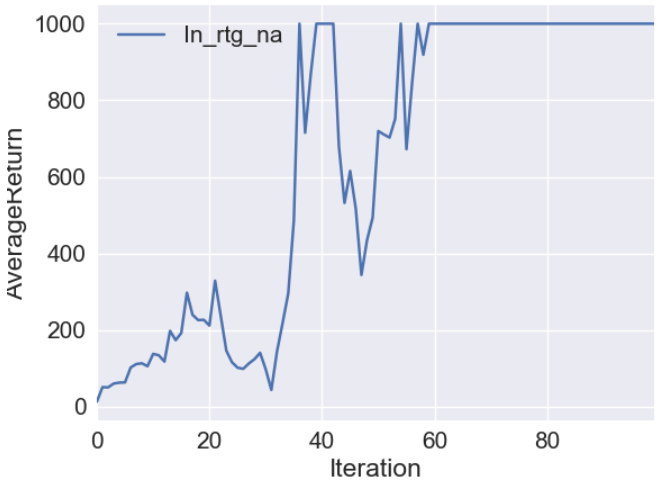
- As we expect by having advantage centering we expect less variance in the results and we actually get it.
And also we know that reward to go ignores the previous rewards that do not effect the future actions. If we consider this we expect to do not see the effect of pass rewards. This means we should see a smoother curve that produces a better results after using this.

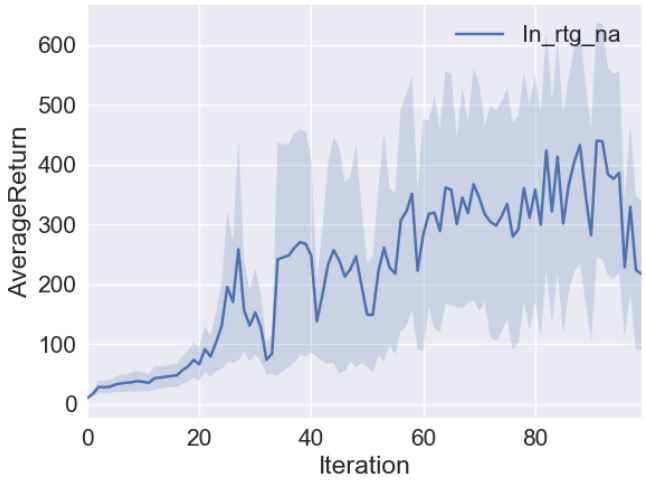
- As the batch size increase the learning curve become smoother and it has less variance.

- Command lines:

```
python train_pg.py CartPole-v0 -n 100 -b 1000 -e 5 -dna –exp_name sb_no_rtg_dna

python train_pg.py CartPole-v0 -n 100 -b 1000 -e 5 -rtg -dna –exp_name sb_rtg_dna

python train_pg.py CartPole-v0 -n 100 -b 1000 -e 5 -rtg –exp_name sb_rtg_na

python train_pg.py CartPole-v0 -n 100 -b 5000 -e 5 -dna –exp_name lb_no_rtg_dna

python train_pg.py CartPole-v0 -n 100 -b 5000 -e 5 -rtg -dna –exp_name lb_rtg_dna

python train_pg.py CartPole-v0 -n 100 -b 5000 -e 5 -rtg --exp_namelb_rtg_na
```

InvertedPendulum-v1





Left: InvertedPendulum, 1 experiment,

Right: Left: InvertedPendulum, 5 experiments

-Command line:

```
python train_pg.py InvertedPendulum-v1  -n 100 -b 2000 -e 5 --learning_rate 0.01
-rtg --exp_name In_rtg_na
```

Section 5, Implement Neural Network Baselines

Left: InvertedPendulum with baseline, Right: InvertedPendulum with W and W/O baseline

- Command line:

```
python train_pg.py InvertedPendulum-v1 -n 100 -b 2000 -e 5 --learning_rate 0.01
-rtg --exp_name In_rtg_na_bl -bl
```

Section 6 HalfCheetah

...