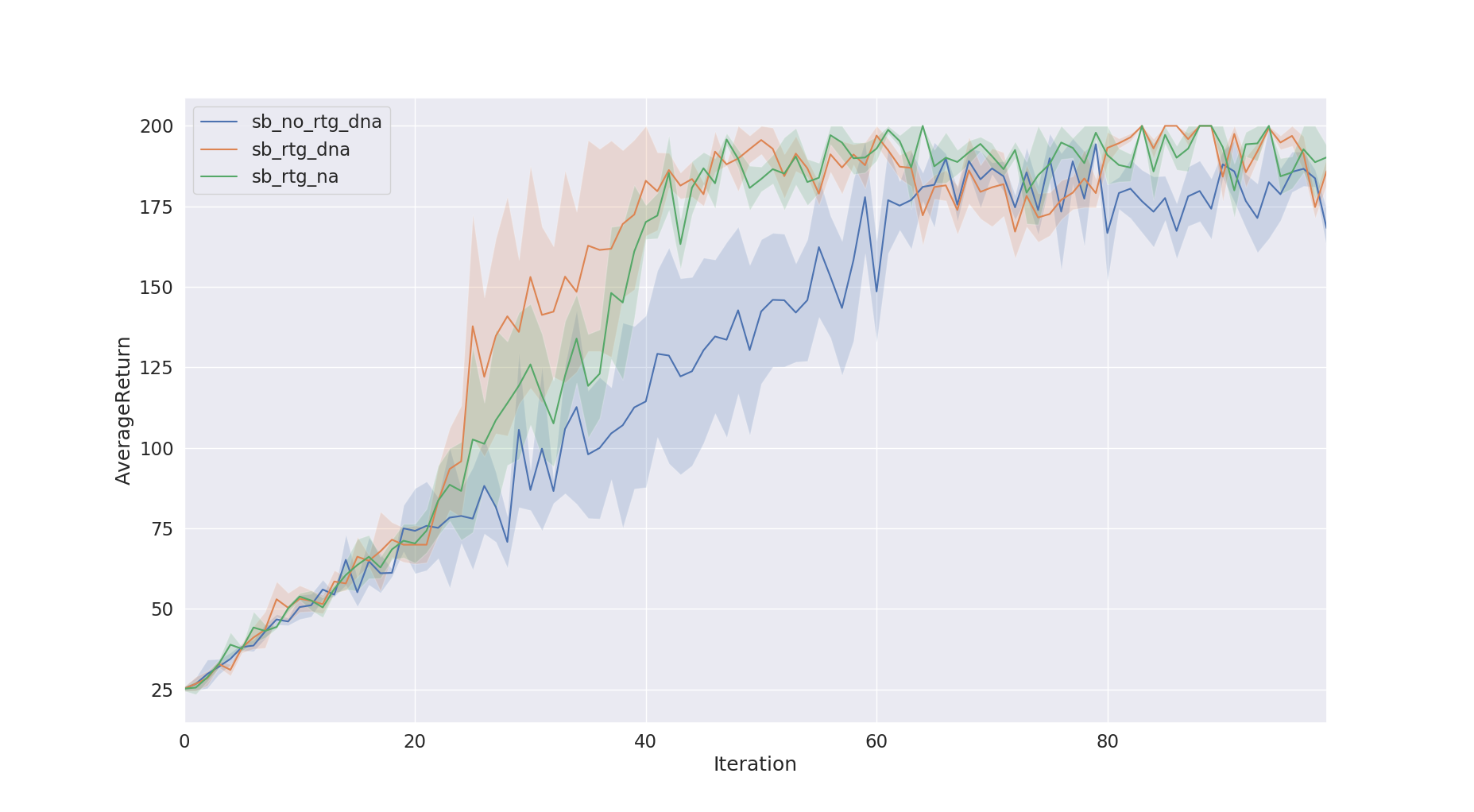
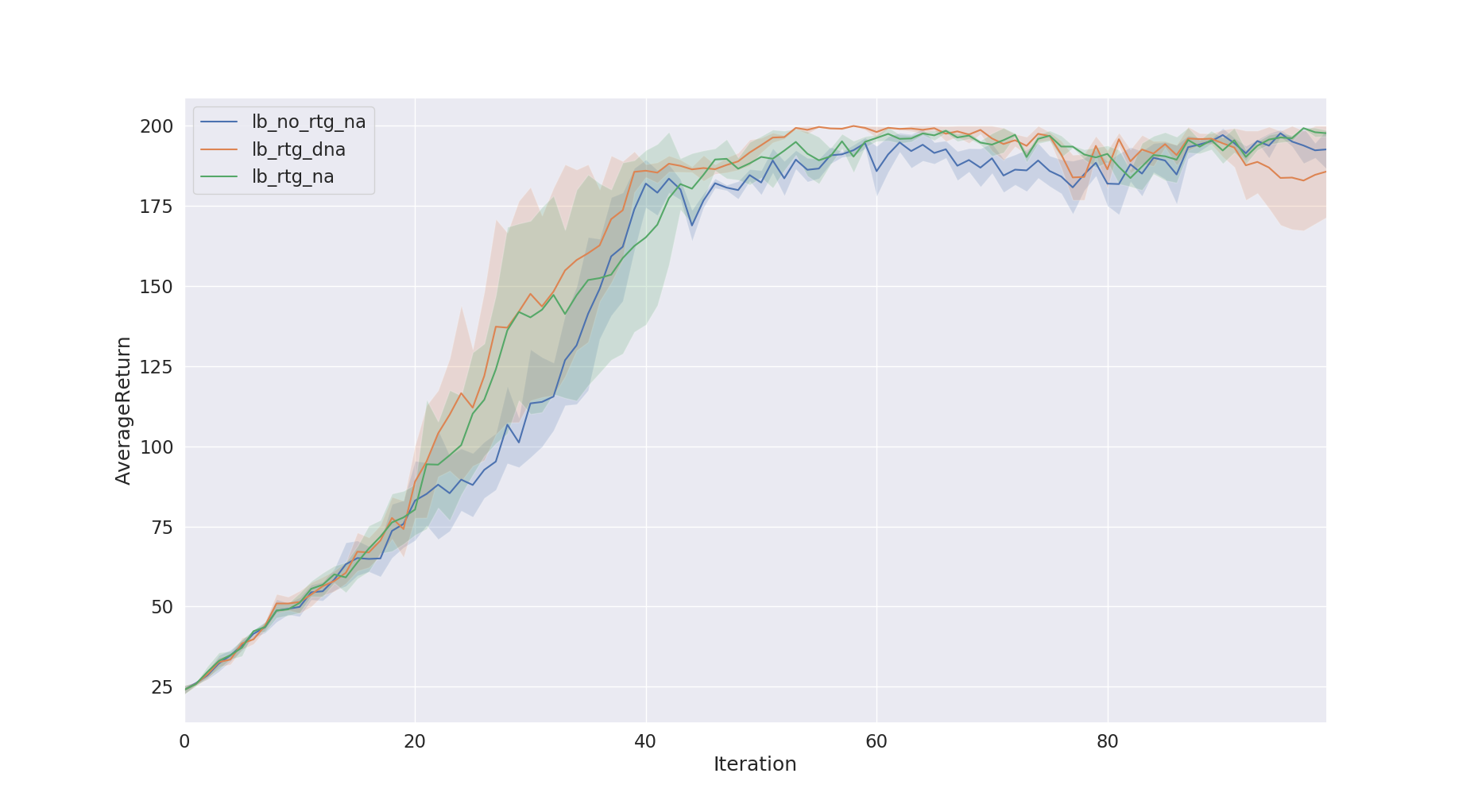
**Problem 4. CartPole**



* The reward-to-go gradient estimator has better performance without advantage-centering.
* The advantage-centering does not help learn but helping reduce the variance.
* The batch size has a clear impact in learning performance, i.e. helping learn faster but not improving performance.

**Command lines:**

python train\_pg\_f18.py CartPole-v0 -n 100 -b 1000 -e 3 -dna --exp\_name sb\_no\_rtg\_dna

python train\_pg\_f18.py CartPole-v0 -n 100 -b 1000 -e 3 -rtg -dna --exp\_name sb\_rtg\_dna

python train\_pg\_f18.py CartPole-v0 -n 100 -b 1000 -e 3 -rtg --exp\_name sb\_rtg\_na

python train\_pg\_f18.py CartPole-v0 -n 100 -b 5000 -e 3 -dna --exp\_name lb\_no\_rtg\_dna

python train\_pg\_f18.py CartPole-v0 -n 100 -b 5000 -e 3 -rtg -dna --exp\_name lb\_rtg\_dna

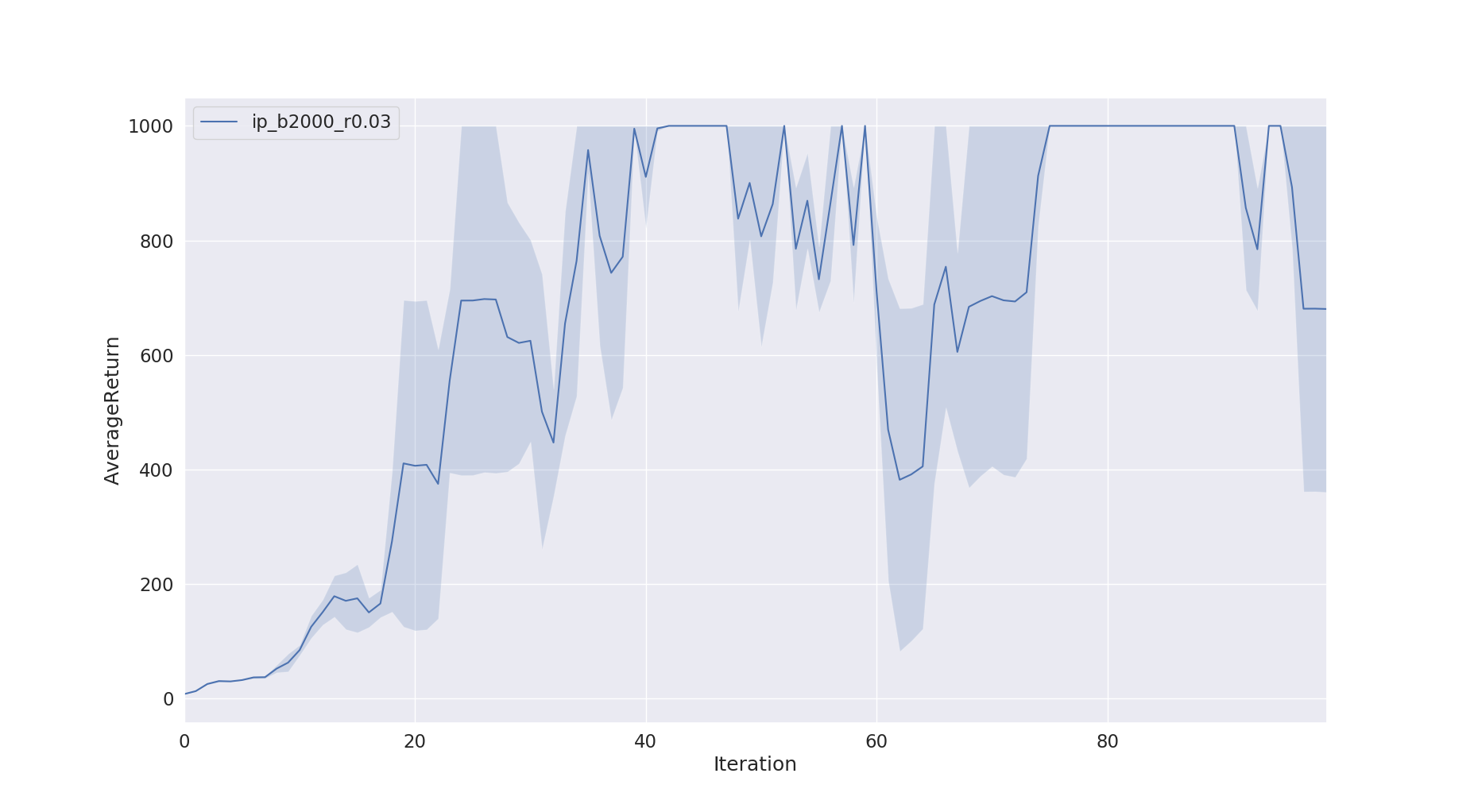
python train\_pg\_f18.py CartPole-v0 -n 100 -b 5000 -e 3 -rtg --exp\_name lb\_rtg\_na

python plot.py data/sb\_no\_rtg\_dna\_CartPole-v0\_19-09-2018\_21-26-20/ data/sb\_rtg\_dna\_CartPole-v0\_19-09-2018\_21-46-46/ data/sb\_rtg\_na\_CartPole-v0\_19-09-2018\_21-49-55/

python plot.py data/lb\_no\_rtg\_dna\_CartPole-v0\_19-09-2018\_22-05-12/ data/lb\_rtg\_dna\_CartPole-v0\_19-09-2018\_22-19-22/ data/lb\_rtg\_na\_CartPole-v0\_19-09-2018\_22-39-08/

**Problem 5. Inverted Pendulum**

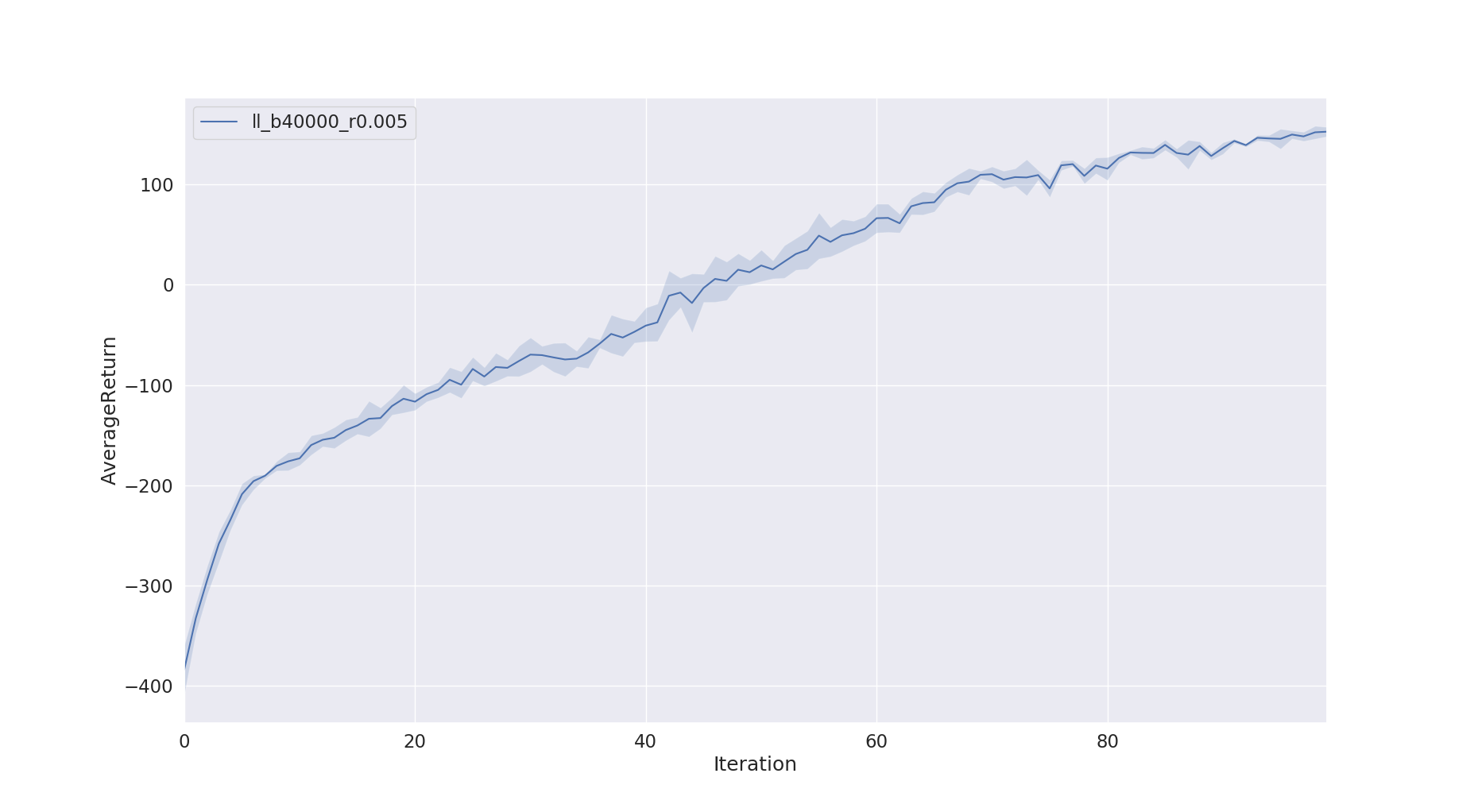
b\* = 2000, lr\* = 0.03

**Command lines:**

python train\_pg\_f18.py InvertedPendulum-v2 -ep 1000 --discount 0.9 -n 100 -e 3 -l 2 -s 64 -b 2000 -lr 0.03 -rtg --exp\_name hc\_b2000\_r0.03

python plot.py data/ip\_b2000\_r0.03\_InvertedPendulum-v2\_20-09-2018\_00-21-21/

**Problem 7. Lunar Lander**



**Command line:**

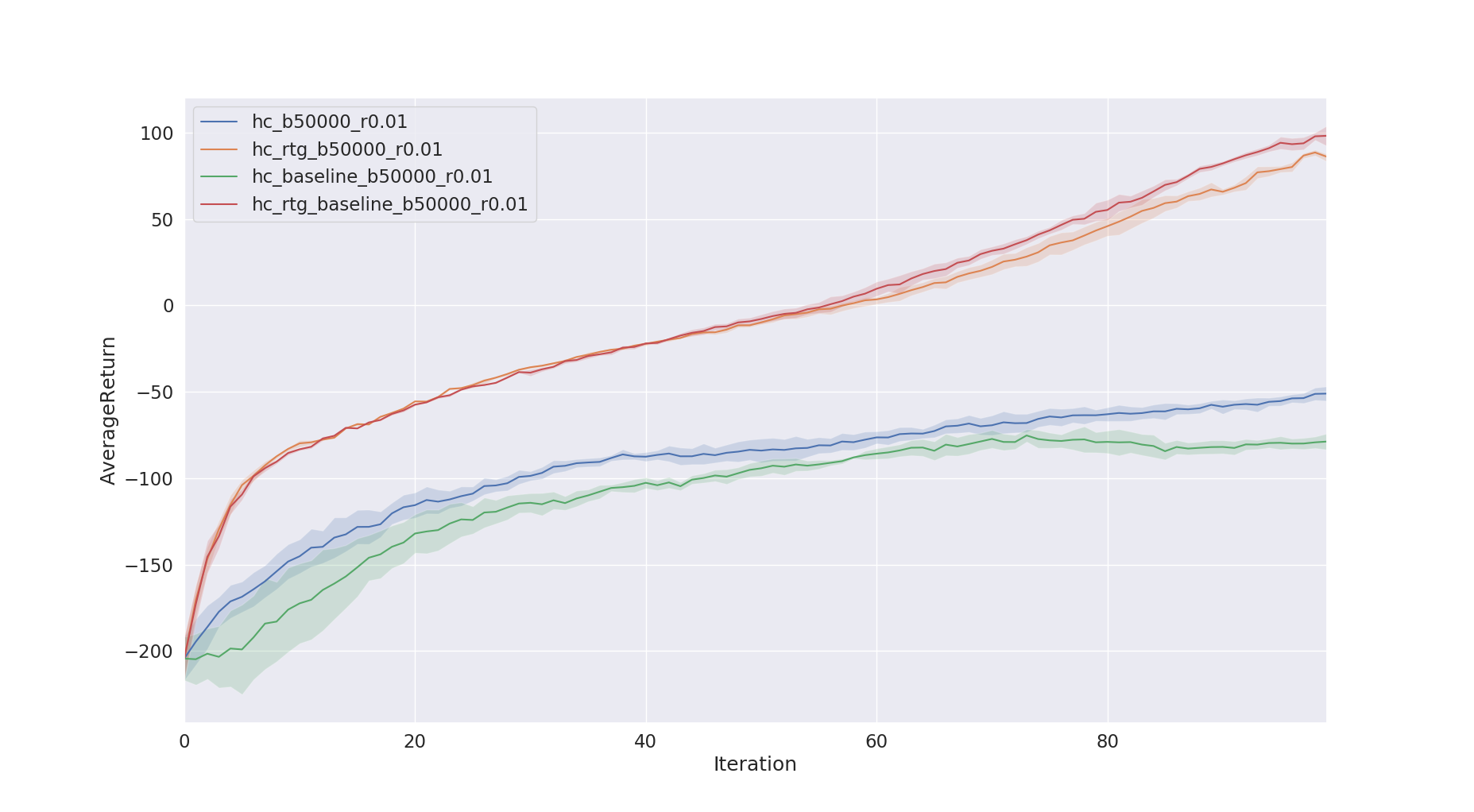
python train\_pg\_f18.py LunarLanderContinuous-v2 -ep 1000 --discount 0.99 -n 100 -e 3 -l 2 -s 64 -b 40000 -lr 0.005 -rtg --nn\_baseline --exp\_name ll\_b40000\_r0.005

python plot.py data/ll\_b40000\_r0.005\_LunarLanderContinuous-v2\_20-09-2018\_01-32-48/

**Problem 8. Half Cheetah**

* The batch size improves the learning performance, i.e. helping learn better, but increasing the learning time.
* The learning rate help learn faster, thus helping learn more given the same time/number of iteration, but if the large learning rate harms the learning performance.

b\* = 50000, lr\* = 0.01. Old-version homework#2 → average reward is around 100.



**Command lines:**

python train\_pg\_f18.py HalfCheetah-v2 -ep 150 --discount 0.9 -n 100 -e 3 -l 2 -s 32 -b 50000 -lr 0.01 --exp\_name hc\_b50000\_r0.01

python train\_pg\_f18.py HalfCheetah-v2 -ep 150 --discount 0.9 -n 100 -e 3 -l 2 -s 32 -b 50000 -lr 0.01 -rtg --exp\_name hc\_rtg\_b50000\_r0.01

python train\_pg\_f18.py HalfCheetah-v2 -ep 150 --discount 0.9 -n 100 -e 3 -l 2 -s 32 -b 50000 -lr 0.01 --nn\_baseline --exp\_name hc\_baseline\_b50000\_r0.01

python train\_pg\_f18.py HalfCheetah-v2 -ep 150 --discount 0.9 -n 100 -e 3 -l 2 -s 32 -b 50000 -lr 0.01 -rtg --nn\_baseline --exp\_name hc\_rtg\_baseline\_b50000\_r0.01

python plot.py data/hc\_b50000\_r0.01\_HalfCheetah-v2\_20-09-2018\_22-24-53/data/hc\_rtg\_b50000\_r0.01\_HalfCheetah-v2\_20-09-2018\_21-12-18/data/hc\_baseline\_b50000\_r0.01\_HalfCheetah-v2\_20-09-2018\_20-01-17/data/hc\_rtg\_baseline\_b50000\_r0.01\_HalfCheetah-v2\_20-09-2018\_18-51-17/