CS285 HW2 Report

Experiment 1

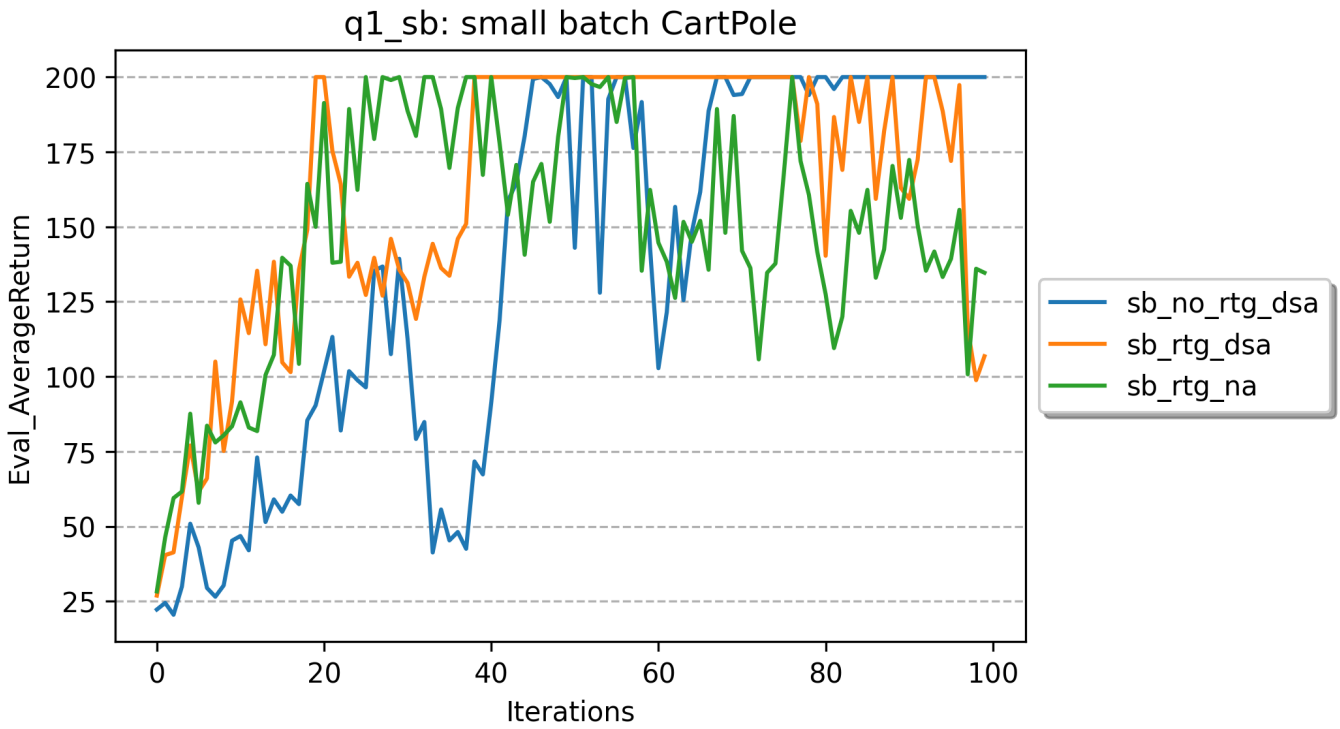


Fig. 1. Learning curves for small batch on CartPole

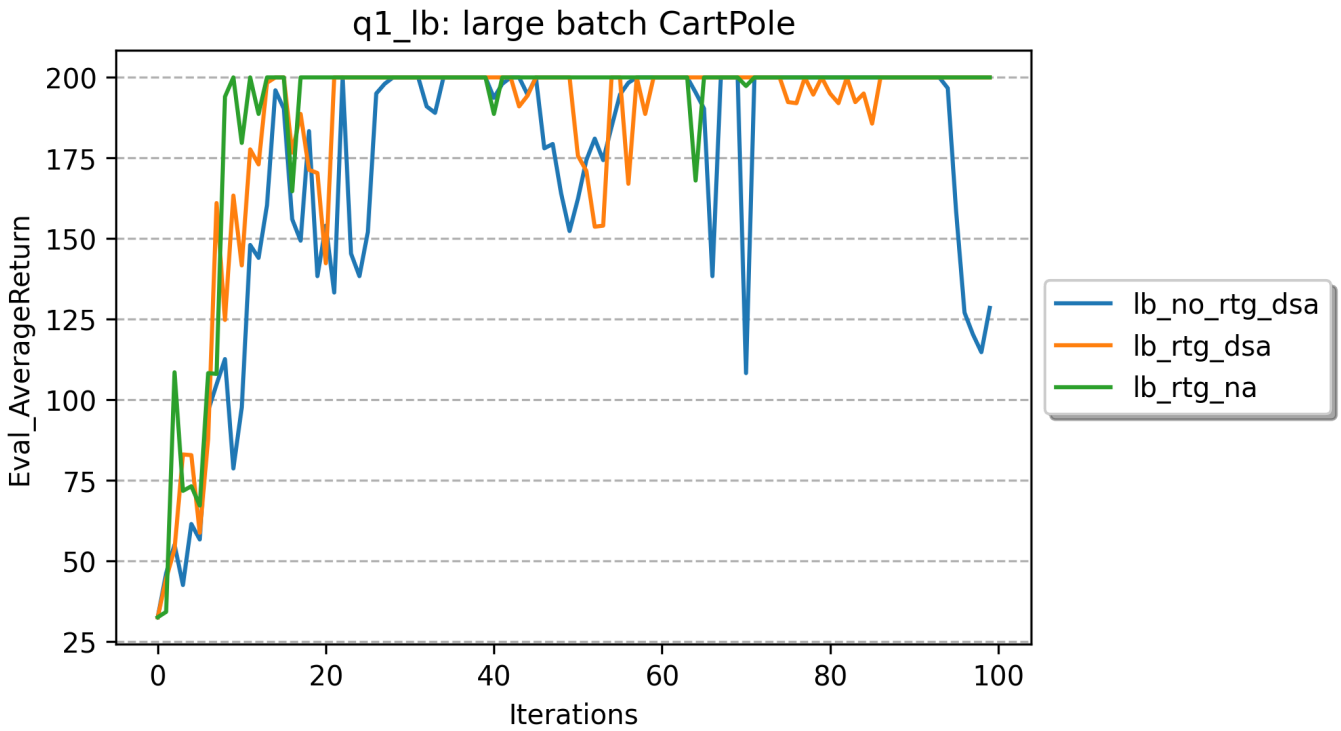


Fig. 2. Learning curves for large batch on CartPole.

Q: Which value estimator has better performance without advantage-standardization: the trajectory-centric one, or the one using reward-to-go?

A: The one using reward-to-go.

Q: Did advantage standardization help?

A: For small batch, it did not help. For large batch, it made the score more stable at a high performance.

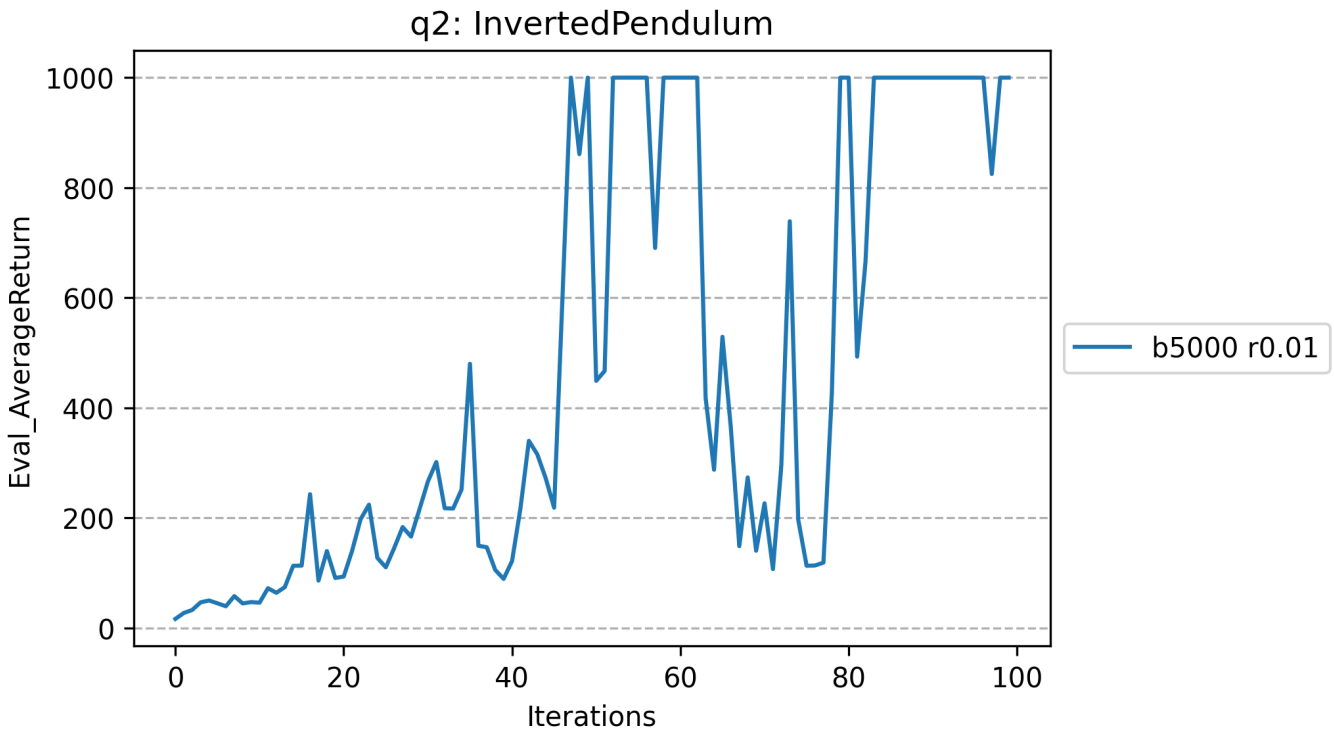
Q: Did the batch size make an impact?

A: Yes. A batch size of 5000 gives a much better performance than that of 1000.

Parameter Configurations

Same as default.

Experiment 2

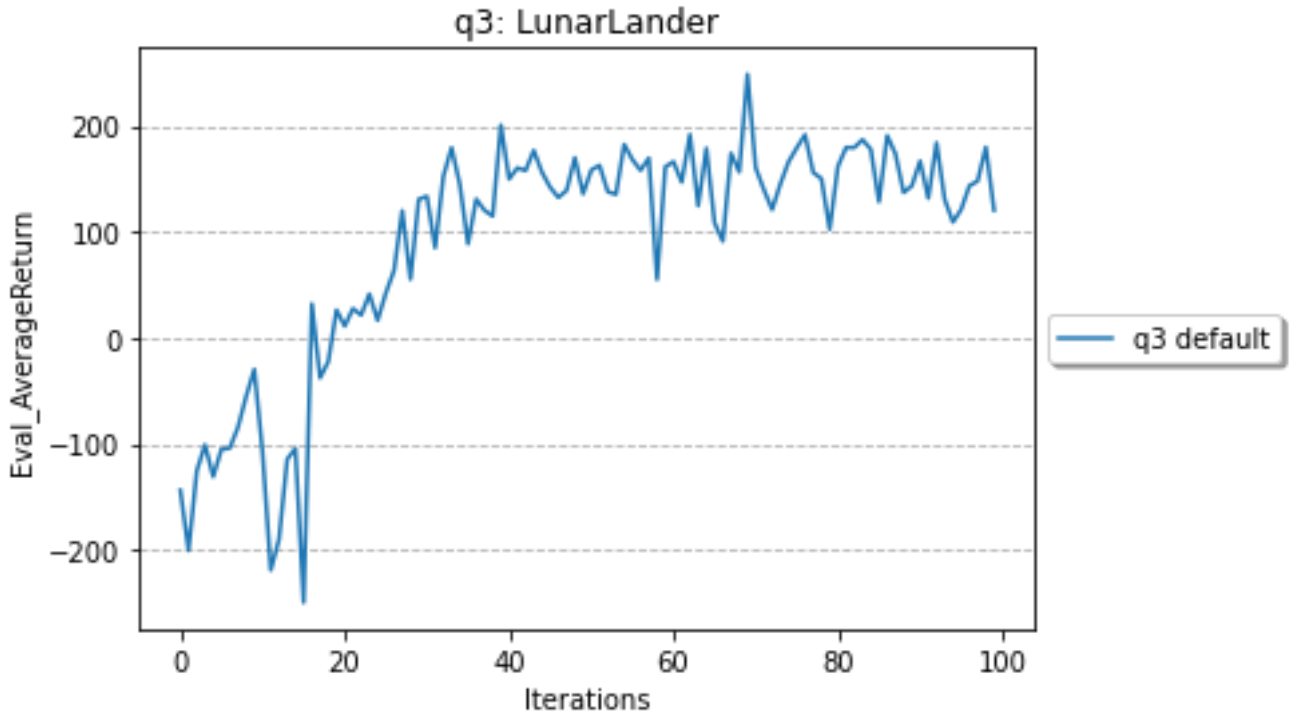


I used $b^* = 5000, r^* = 0.01$.

The command line is

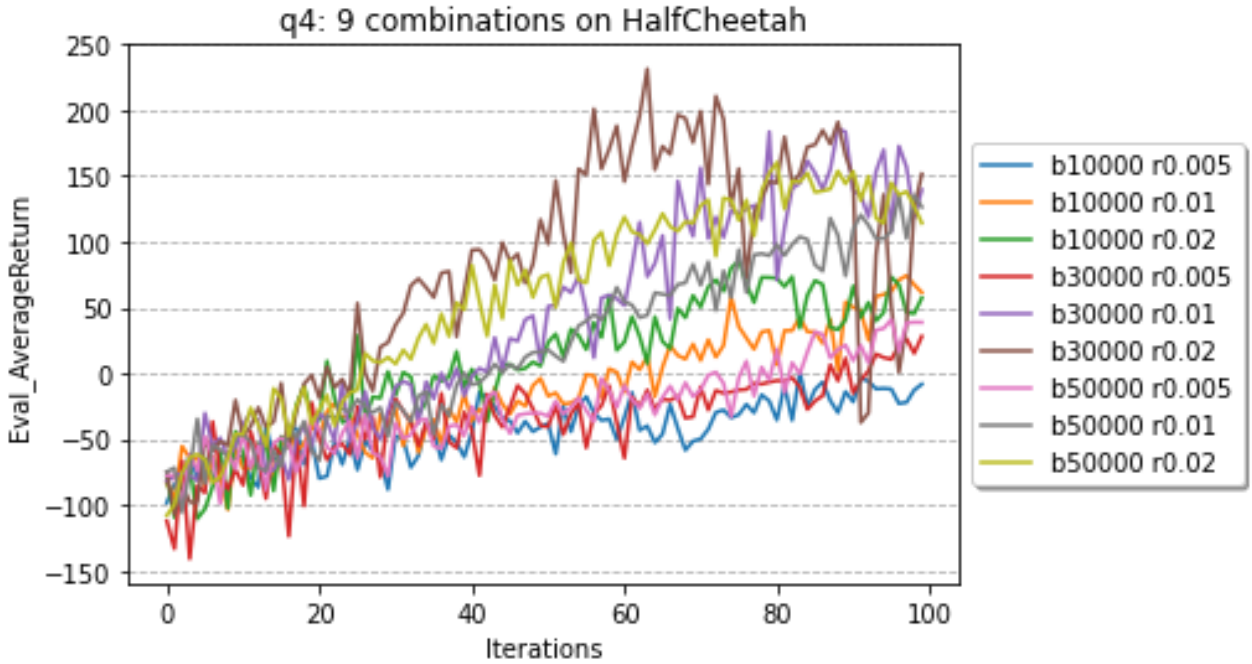
```
python cs285/scripts/run_hw2.py --env_name InvertedPendulum-v4 \
--ep_len 1000 --discount 0.9 -n 100 -l 2 -s 64 -b 5000 -lr 0.01 -rtg \
--exp_name q2_b5000_r0.01
```

Experiment 3



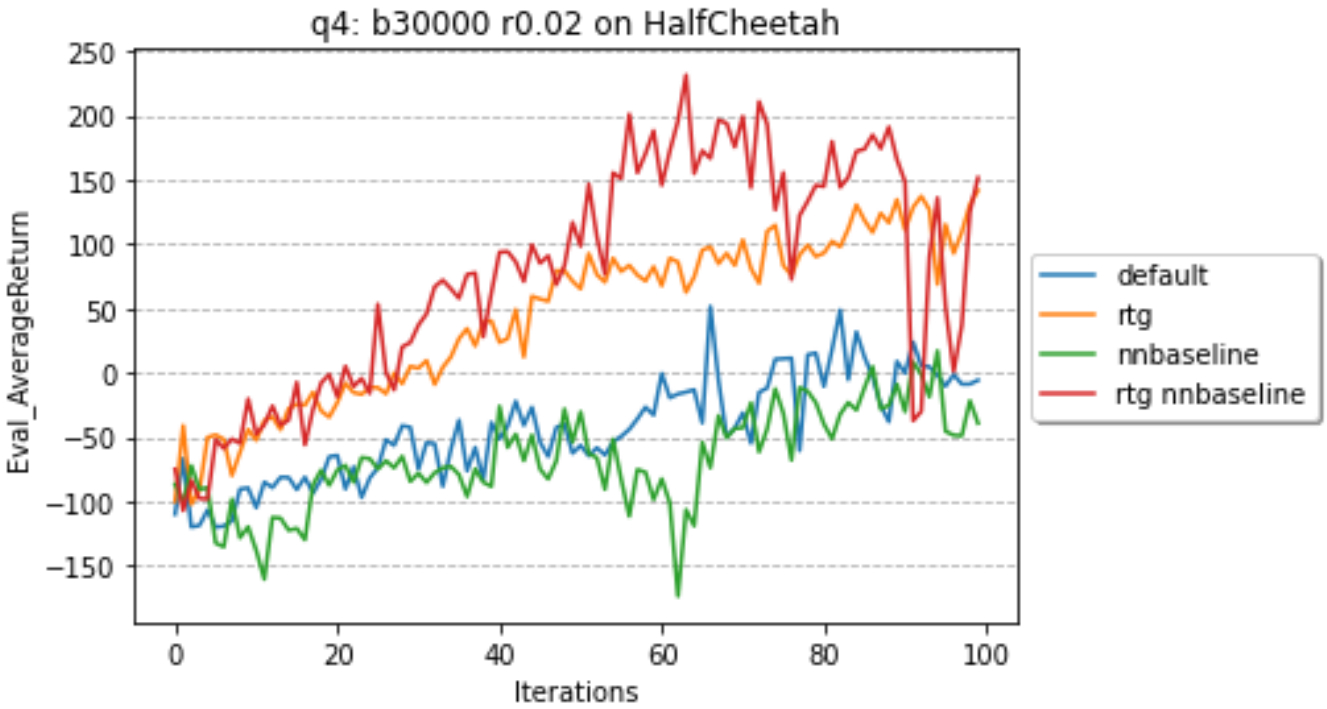
It reaches more than 200 at the highest point.

Experiement 4

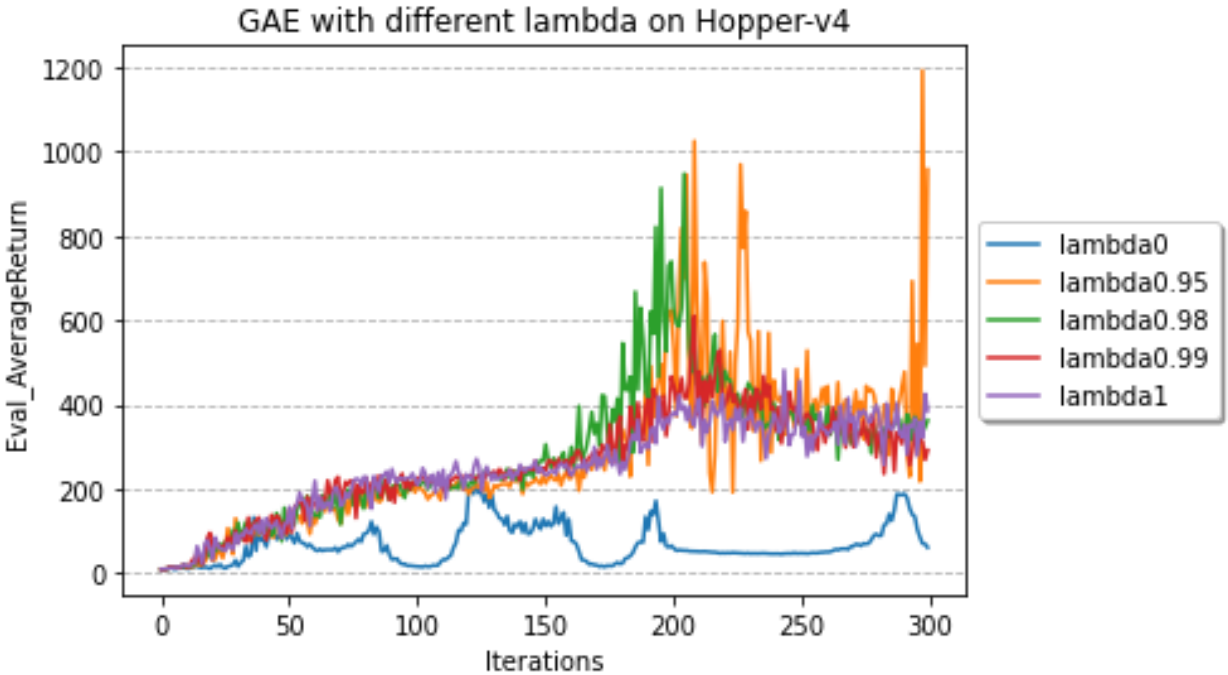


The $b = 30000, r = 0.02$ gets the best result.

A higher learning rate produces a faster increase in early time. A larger batch size can increase the best reward to some extent.



Experiment 5



The best λ is 0.95. The λ can influence the highest score largely, while influcing the convergence not apprently. A good λ can produce a very high best score.

Appendix

I run all the experiments through a Makefile file. The commands I used are:

```
# Makefile

submit:
    -rm data.zip run_logs.zip
    zip cs285.zip -r cs285
    zip run_logs.zip -r data

exp1:
    python cs285/scripts/run_hw2.py --env_name CartPole-v0 -n 100 -b 1000 \
    -dsa --exp_name q1_sb_no_rtg_dsa
    python cs285/scripts/run_hw2.py --env_name CartPole-v0 -n 100 -b 1000 \
    -rtg -dsa --exp_name q1_sb_rtg_dsa
    python cs285/scripts/run_hw2.py --env_name CartPole-v0 -n 100 -b 1000 \
    -rtg --exp_name q1_sb_rtg_na
    python cs285/scripts/run_hw2.py --env_name CartPole-v0 -n 100 -b 5000 \
    -dsa --exp_name q1_lb_no_rtg_dsa
    python cs285/scripts/run_hw2.py --env_name CartPole-v0 -n 100 -b 5000 \
    -rtg -dsa --exp_name q1_lb_rtg_dsa
    python cs285/scripts/run_hw2.py --env_name CartPole-v0 -n 100 -b 5000 \
    -rtg --exp_name q1_lb_rtg_na

batch_size = 5000
learning_rate = 0.01

exp2:
    python cs285/scripts/run_hw2.py --env_name InvertedPendulum-v4 \
    --ep_len 1000 --discount 0.9 -n 100 -l 2 -s 64 -b $(batch_size) -lr
$(learning_rate) -rtg \
    --exp_name q2_b$(batch_size)_r$(learning_rate)

exp3:
    python cs285/scripts/run_hw2.py \
    --env_name LunarLanderContinuous-v2 --ep_len 1000 \
    --discount 0.99 -n 100 -l 2 -s 64 -b 40000 -lr 0.005 \
    --reward_to_go --nn_baseline --exp_name q3_b40000_r0.005

exp4-1:
    python cs285/scripts/run_hw2.py --env_name HalfCheetah-v4 --ep_len 150 \
    --discount 0.95 -n 100 -l 2 -s 32 -b 10000 -lr 0.005 -rtg --nn_baseline \
    --exp_name q4_search_b10000_lr0.005_rtg_nnbaseline
    python cs285/scripts/run_hw2.py --env_name HalfCheetah-v4 --ep_len 150 \
    --discount 0.95 -n 100 -l 2 -s 32 -b 10000 -lr 0.01 -rtg --nn_baseline \
    --exp_name q4_search_b10000_lr0.01_rtg_nnbaseline
    python cs285/scripts/run_hw2.py --env_name HalfCheetah-v4 --ep_len 150 \
    --discount 0.95 -n 100 -l 2 -s 32 -b 10000 -lr 0.02 -rtg --nn_baseline \
    --exp_name q4_search_b10000_lr0.02_rtg_nnbaseline
    python cs285/scripts/run_hw2.py --env_name HalfCheetah-v4 --ep_len 150 \
    --discount 0.95 -n 100 -l 2 -s 32 -b 30000 -lr 0.005 -rtg --nn_baseline \
    --exp_name q4_search_b30000_lr0.005_rtg_nnbaseline
    python cs285/scripts/run_hw2.py --env_name HalfCheetah-v4 --ep_len 150 \
```

```
    --discount 0.95 -n 100 -l 2 -s 32 -b 30000 -lr 0.01 -rtg --nn_baseline \
    --exp_name q4_search_b30000_lr0.01_rtg_nnbaseline
    python cs285/scripts/run_hw2.py --env_name HalfCheetah-v4 --ep_len 150 \
    --discount 0.95 -n 100 -l 2 -s 32 -b 30000 -lr 0.02 -rtg --nn_baseline \
    --exp_name q4_search_b30000_lr0.02_rtg_nnbaseline
    python cs285/scripts/run_hw2.py --env_name HalfCheetah-v4 --ep_len 150 \
    --discount 0.95 -n 100 -l 2 -s 32 -b 50000 -lr 0.005 -rtg --nn_baseline \
    --exp_name q4_search_b50000_lr0.005_rtg_nnbaseline
    python cs285/scripts/run_hw2.py --env_name HalfCheetah-v4 --ep_len 150 \
    --discount 0.95 -n 100 -l 2 -s 32 -b 50000 -lr 0.01 -rtg --nn_baseline \
    --exp_name q4_search_b50000_lr0.01_rtg_nnbaseline
    python cs285/scripts/run_hw2.py --env_name HalfCheetah-v4 --ep_len 150 \
    --discount 0.95 -n 100 -l 2 -s 32 -b 50000 -lr 0.02 -rtg --nn_baseline \
    --exp_name q4_search_b50000_lr0.02_rtg_nnbaseline

exp4-2:
    python cs285/scripts/run_hw2.py --env_name HalfCheetah-v4 --ep_len 150 \
    --discount 0.95 -n 100 -l 2 -s 32 -b 50000 -lr 0.02 \
    --exp_name q4_b50000_r0.02
    python cs285/scripts/run_hw2.py --env_name HalfCheetah-v4 --ep_len 150 \
    --discount 0.95 -n 100 -l 2 -s 32 -b 50000 -lr 0.02 -rtg \
    --exp_name q4_b50000_r0.02_rtg
    python cs285/scripts/run_hw2.py --env_name HalfCheetah-v4 --ep_len 150 \
    --discount 0.95 -n 100 -l 2 -s 32 -b 50000 -lr 0.02 --nn_baseline \
    --exp_name q4_b50000_r0.02_nnbaseline

exp5:
    python cs285/scripts/run_hw2.py \
    --env_name Hopper-v4 --ep_len 1000 \
    --discount 0.99 -n 300 -l 2 -s 32 -b 2000 -lr 0.001 \
    --reward_to_go --nn_baseline --action_noise_std 0.5 --gae_lambda 0 \
    --exp_name q5_b2000_r0.001_lambda0
    python cs285/scripts/run_hw2.py \
    --env_name Hopper-v4 --ep_len 1000 \
    --discount 0.99 -n 300 -l 2 -s 32 -b 2000 -lr 0.001 \
    --reward_to_go --nn_baseline --action_noise_std 0.5 --gae_lambda 0.95 \
    --exp_name q5_b2000_r0.001_lambda0.95
    python cs285/scripts/run_hw2.py \
    --env_name Hopper-v4 --ep_len 1000 \
    --discount 0.99 -n 300 -l 2 -s 32 -b 2000 -lr 0.001 \
    --reward_to_go --nn_baseline --action_noise_std 0.5 --gae_lambda 0.98 \
    --exp_name q5_b2000_r0.001_lambda0.98
    python cs285/scripts/run_hw2.py \
    --env_name Hopper-v4 --ep_len 1000 \
    --discount 0.99 -n 300 -l 2 -s 32 -b 2000 -lr 0.001 \
    --reward_to_go --nn_baseline --action_noise_std 0.5 --gae_lambda 0.99 \
    --exp_name q5_b2000_r0.001_lambda0.99
    python cs285/scripts/run_hw2.py \
    --env_name Hopper-v4 --ep_len 1000 \
    --discount 0.99 -n 300 -l 2 -s 32 -b 2000 -lr 0.001 \
```

```
    --reward_to_go --nn_baseline --action_noise_std 0.5 --gae_lambda 1 \
    --exp_name q5_b2000_r0.001_lambda1
```