Assignment 2: Policy Gradient

Andrew ID: `lzaceria`
Collaborators: `Write the Andrew IDs of your collaborators here (if any).`
NOTE: Please do **NOT** change the sizes of the answer blocks or plots.

5 Small-Scale Experiments

5.1 Experiment 1 (Cartpole) – [5 points total]

5.1.1 Configurations

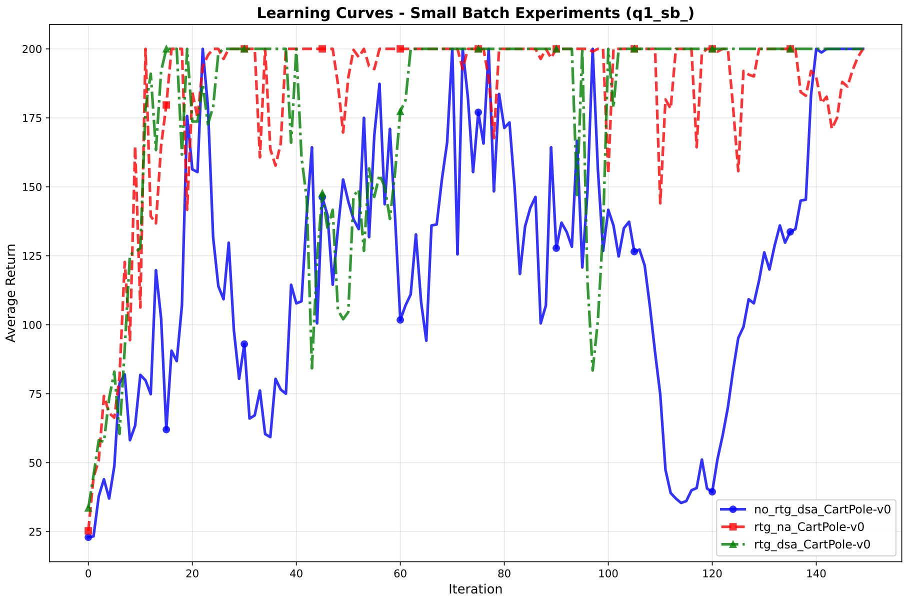
Q5.1.1

```
python rob831/scripts/run_hw2.py --env_name CartPole-v0 -n 150 -b 1500 \
    -dsa --exp_name q1_sb_no_rtg_dsa

python rob831/scripts/run_hw2.py --env_name CartPole-v0 -n 150 -b 1500 \
    -rtg -dsa --exp_name q1_sb_rtg_dsa

python rob831/scripts/run_hw2.py --env_name CartPole-v0 -n 150 -b 1500 \
    -rtg --exp_name q1_sb_rtg_na

python rob831/scripts/run_hw2.py --env_name CartPole-v0 -n 150 -b 6000 \
    -dsa --exp_name q1_lb_no_rtg_dsa

python rob831/scripts/run_hw2.py --env_name CartPole-v0 -n 150 -b 6000 \
    -rtg -dsa --exp_name q1_lb_rtg_dsa

python rob831/scripts/run_hw2.py --env_name CartPole-v0 -n 150 -b 6000 \
    -rtg --exp_name q1_lb_rtg_na
```

5.1.2 Plots

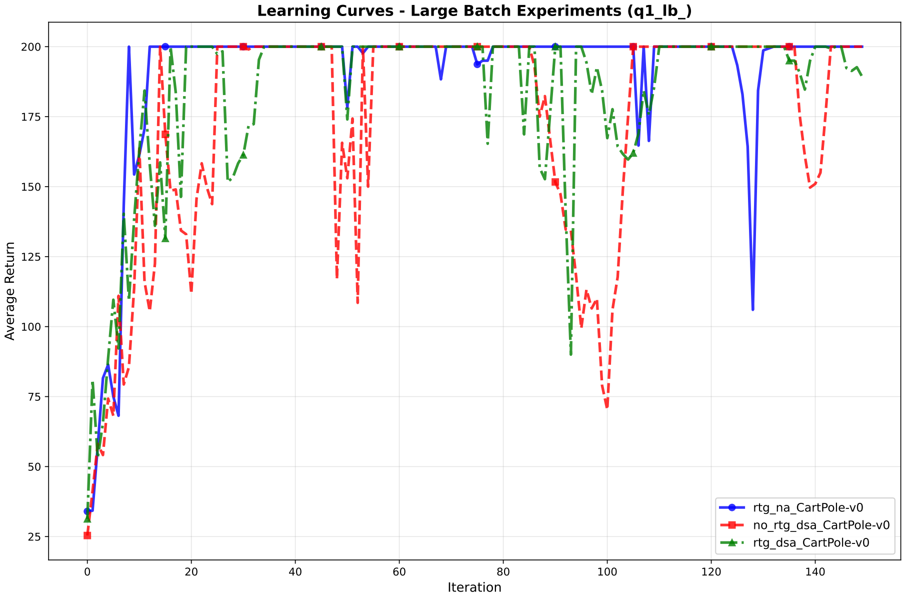
5.1.2.1 Small batch – [1 points]

Q5.1.2.1



5.1.2.2 Large batch – [1 points]

Q5.1.2.2



5.1.3 Analysis

5.1.3.1 Value estimator – [1 points]

Q5.1.3.1

Reward-to-go performs better than trajectory-centric.

5.1.3.2 Advantage standardization – [1 points]

Q5.1.3.2

Yes, advantage standardization helped.

5.1.3.3 Batch size – [1 points]

Q5.1.3.3

Yes, the batch size made an impact. A bigger batch size generally improves performance.

5.2 Experiment 2 (InvertedPendulum) – [4 points total]

5.2.1 Configurations – [1.5 points]

Q5.2.1

```
python rob831/scripts/run_hw2.py \
    --env_name InvertedPendulum-v4 \
    --ep_len 1000 \
    --discount 0.92 \
    -n 100 \
    -l 2 \
    -s 64 \
    -b 50 \
    -lr 0.02 \
    -rtg \
    --exp_name q2_b50_r0.02_test
python rob831/scripts/run_hw2.py \
    --env_name InvertedPendulum-v4 \
    --ep_len 1000 \
    --discount 0.92 \
    -n 100 \
    -l 2 \
    -s 64 \
    -b 40 \
    -lr 0.02 \
    -rtg \
    --exp_name q2_b40_r0.02_test
python rob831/scripts/run_hw2.py \
    --env_name InvertedPendulum-v4 \
    --ep_len 1000 \
    --discount 0.92 \
    -n 100 \
    -l 2 \
    -s 64 \
    -b 50 \
    -lr 0.01 \
    -rtg \
    --exp_name q2_b50_r0.01_test
```

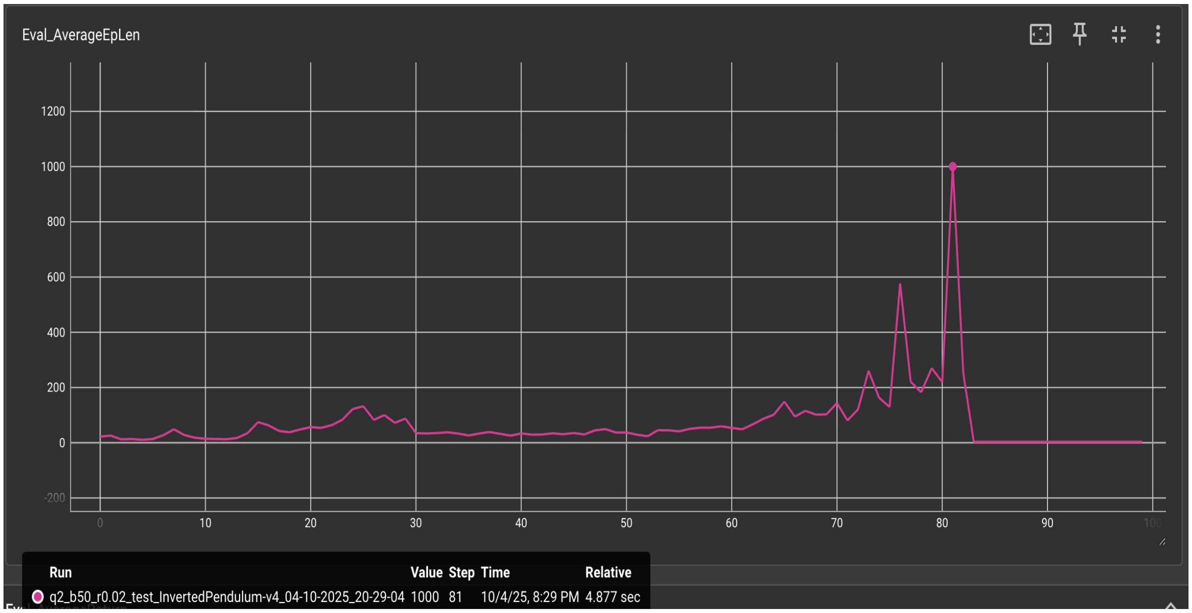
5.2.2 smallest b* and largest r* (same run) – [1.5 points]

Q5.2.2

b*=50, r*=0.02

5.2.3 Plot – [1 points]

Q5.2.3



Run	Value	Step	Time	Relative
● q2_b50_r0.02_test_InvertedPendulum-v4_04-10-2025_20-29-04	1000	81	10/4/25, 8:29 PM	4.877 sec

7 More Complex Experiments

7.1 Experiment 3 (LunarLander) – [1 points total]

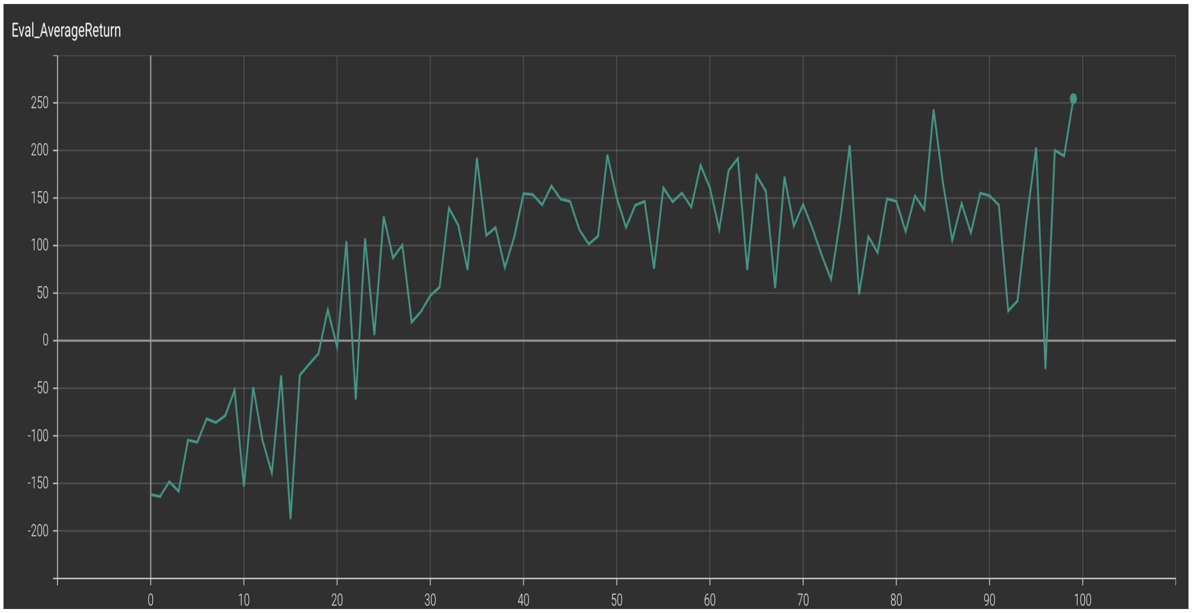
7.1.1 Configurations

Q7.1.1

```
Had to switch to v3
python rob831/scripts/run_hw2.py \
    --env_name LunarLanderContinuous-v3 --ep_len 1000
    --discount 0.99 -n 100 -l 2 -s 64 -b 10000 -lr 0.005 \
    --reward_to_go --nn_baseline --exp_name q3_b10000_r0.005
```

7.1.2 Plot – [1 points]

Q7.1.2



7.2 Experiment 4 (HalfCheetah) – [1 points]

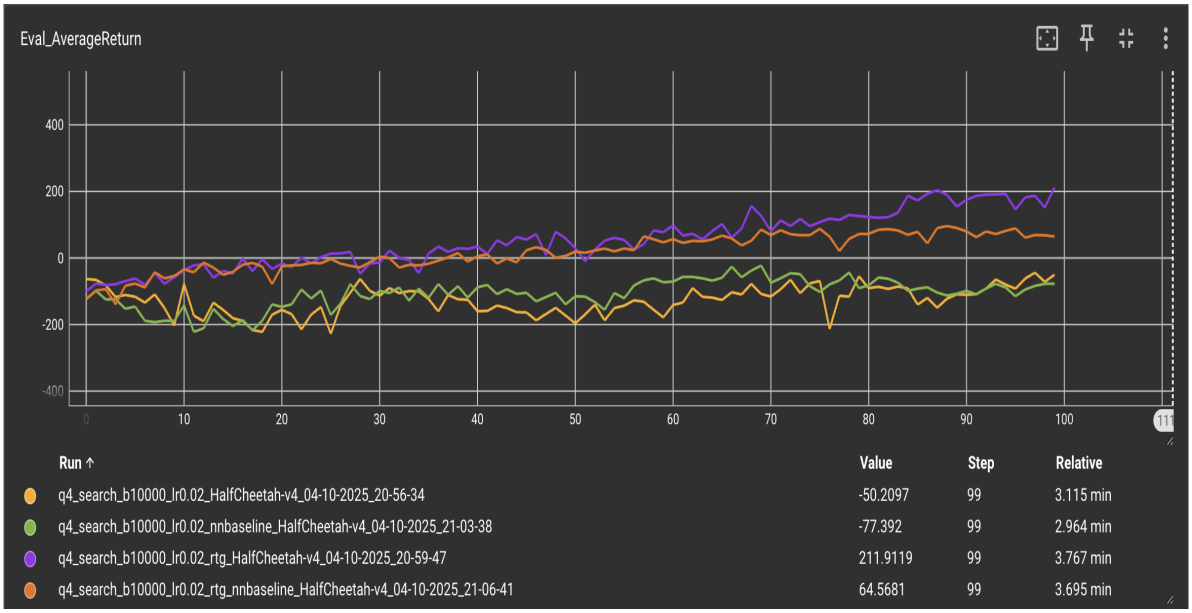
7.2.1 Configurations

Q7.2.1

```
python rob831/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 \
    --exp_name q4_search_b10000_lr0.02
python rob831/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 \
    --exp_name q4_search_b10000_lr0.02_rtg
python rob831/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 --nn_baseline \
    --exp_name q4_search_b10000_lr0.02_nnbaseline
python rob831/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
```

7.2.2 Plot – [1 points]

Q7.2.2



Run ↑	Value	Step	Relative
● q4_search_b10000_lr0.02_HalfCheetah-v4_04-10-2025_20-56-34	-50.2097	99	3.115 min
● q4_search_b10000_lr0.02_nnbaseline_HalfCheetah-v4_04-10-2025_21-03-38	-77.392	99	2.964 min
● q4_search_b10000_lr0.02_rtg_HalfCheetah-v4_04-10-2025_20-59-47	211.9119	99	3.767 min
● q4_search_b10000_lr0.02_rtg_nnbaseline_HalfCheetah-v4_04-10-2025_21-06-41	64.5681	99	3.695 min

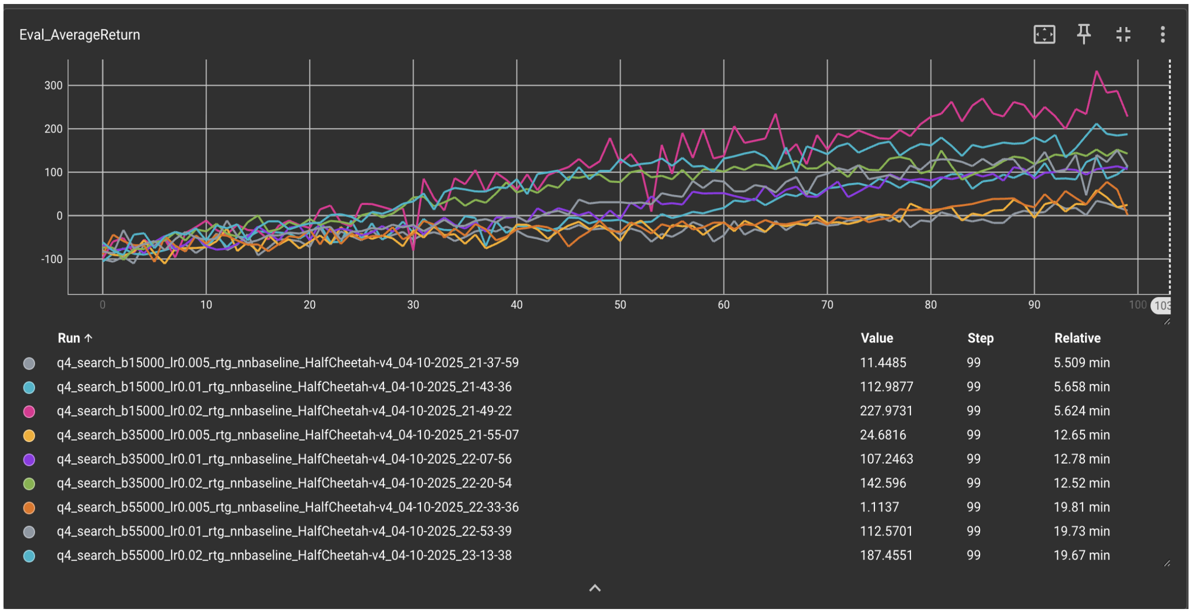
7.2.3 Optimal b* and r* – [0.5 points]

Q7.2.3

b15000, r0.02

7.2.4 Plot – [0.5 points]

Q7.2.4



	Run ↑	Value	Step	Relative
●	q4_search_b15000_lr0.005_rtg_nnbaseline_HalfCheetah-v4_04-10-2025_21-37-59	11.4485	99	5.509 min
●	q4_search_b15000_lr0.01_rtg_nnbaseline_HalfCheetah-v4_04-10-2025_21-43-36	112.9877	99	5.658 min
●	q4_search_b15000_lr0.02_rtg_nnbaseline_HalfCheetah-v4_04-10-2025_21-49-22	227.9731	99	5.624 min
●	q4_search_b35000_lr0.005_rtg_nnbaseline_HalfCheetah-v4_04-10-2025_21-55-07	24.6816	99	12.65 min
●	q4_search_b35000_lr0.01_rtg_nnbaseline_HalfCheetah-v4_04-10-2025_22-07-56	107.2463	99	12.78 min
●	q4_search_b35000_lr0.02_rtg_nnbaseline_HalfCheetah-v4_04-10-2025_22-20-54	142.596	99	12.52 min
●	q4_search_b55000_lr0.005_rtg_nnbaseline_HalfCheetah-v4_04-10-2025_22-33-36	1.1137	99	19.81 min
●	q4_search_b55000_lr0.01_rtg_nnbaseline_HalfCheetah-v4_04-10-2025_22-53-39	112.5701	99	19.73 min
●	q4_search_b55000_lr0.02_rtg_nnbaseline_HalfCheetah-v4_04-10-2025_23-13-38	187.4551	99	19.67 min

7.2.5 Describe how b* and r* affect task performance – [0.5 points]

Q7.2.5

A higher learning rate r* generally improves performance, while b* does not change performance much if comparing with the same r*.

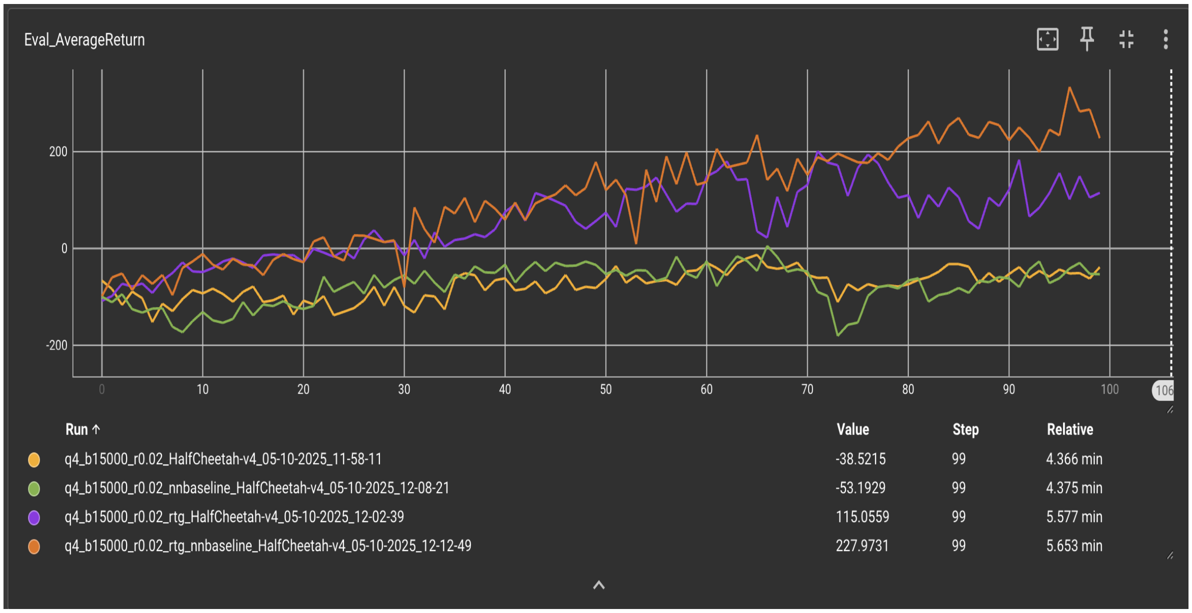
7.2.6 Configurations with optimal b* and r* – [0.5 points]

Q7.2.6

```
python rob831/scripts/run_hw2.py --env_name HalfCheetah-v4 --ep_len 150 \
    --discount 0.95 -n 100 -l 2 -s 32 -b b15000 -lr 0.02 \
    --exp_name q4_b15000_r0.02

python rob831/scripts/run_hw2.py --env_name HalfCheetah-v4 --ep_len 150 \
    --discount 0.95 -n 100 -l 2 -s 32 -b b15000 -lr 0.02 -rtg \
    --exp_name q4_b15000_r0.02_rtg

python rob831/scripts/run_hw2.py --env_name HalfCheetah-v4 --ep_len 150 \
    --discount 0.95 -n 100 -l 2 -s 32 -b b15000 -lr 0.02 --nn_baseline \
    --exp_name q4_b15000_r0.02_nnbaseline

python rob831/scripts/run_hw2.py --env_name HalfCheetah-v4 --ep_len 150 \
    --discount 0.95 -n 100 -l 2 -s 32 -b b15000 -lr 0.02 -rtg --nn_baseline \
    --exp_name q4_b15000_r0.02_rtg_nnbaseline
```

7.2.7 Plot for four runs with optimal b* and r* – [0.5 points]

Q7.2.7



	Run ↑	Value	Step	Relative
●	q4_b15000_r0.02_HalfCheetah-v4_05-10-2025_11-58-11	-38.5215	99	4.366 min
●	q4_b15000_r0.02_nnbaseline_HalfCheetah-v4_05-10-2025_12-08-21	-53.1929	99	4.375 min
●	q4_b15000_r0.02_rtg_HalfCheetah-v4_05-10-2025_12-02-39	115.0559	99	5.577 min
●	q4_b15000_r0.02_rtg_nnbaseline_HalfCheetah-v4_05-10-2025_12-12-49	227.9731	99	5.653 min

8 Implementing Generalized Advantage Estimation

8.1 Experiment 5 (Hopper) – [4 points]

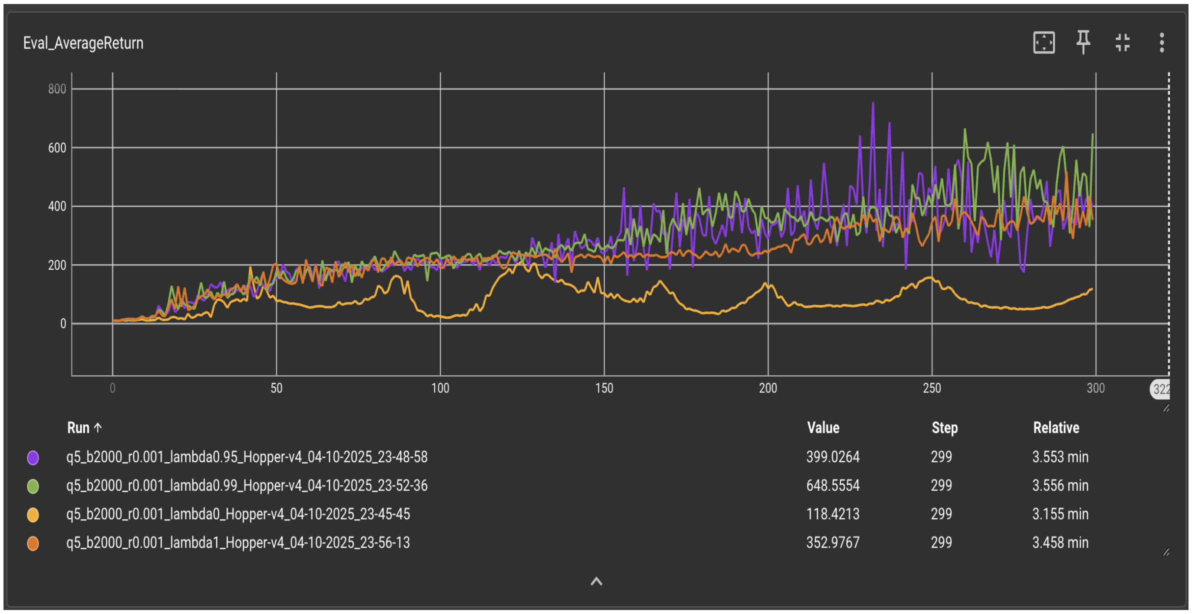
8.1.1 Configurations

Q8.1.1

```
# λ ∈ [0, 0.95, 0.99, 1]
python rob831/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 <λ> \
    --exp_name q5_b2000_r0.001_lambda<λ>
```

8.1.2 Plot – [2 points]

Q8.1.2



	Run ↑	Value	Step	Relative
●	q5_b2000_r0.001_lambda0.95_Hopper-v4_04-10-2025_23-48-58	399.0264	299	3.553 min
●	q5_b2000_r0.001_lambda0.99_Hopper-v4_04-10-2025_23-52-36	648.5554	299	3.556 min
●	q5_b2000_r0.001_lambda0_Hopper-v4_04-10-2025_23-45-45	118.4213	299	3.155 min
●	q5_b2000_r0.001_lambda1_Hopper-v4_04-10-2025_23-56-13	352.9767	299	3.458 min

8.1.3 Describe how λ affects task performance – [2 points]

Q8.1.3

Generally, a higher λ leads to a higher return and better performance.

9 More Bonus!

9.1 Parallelization – [1.5 points]

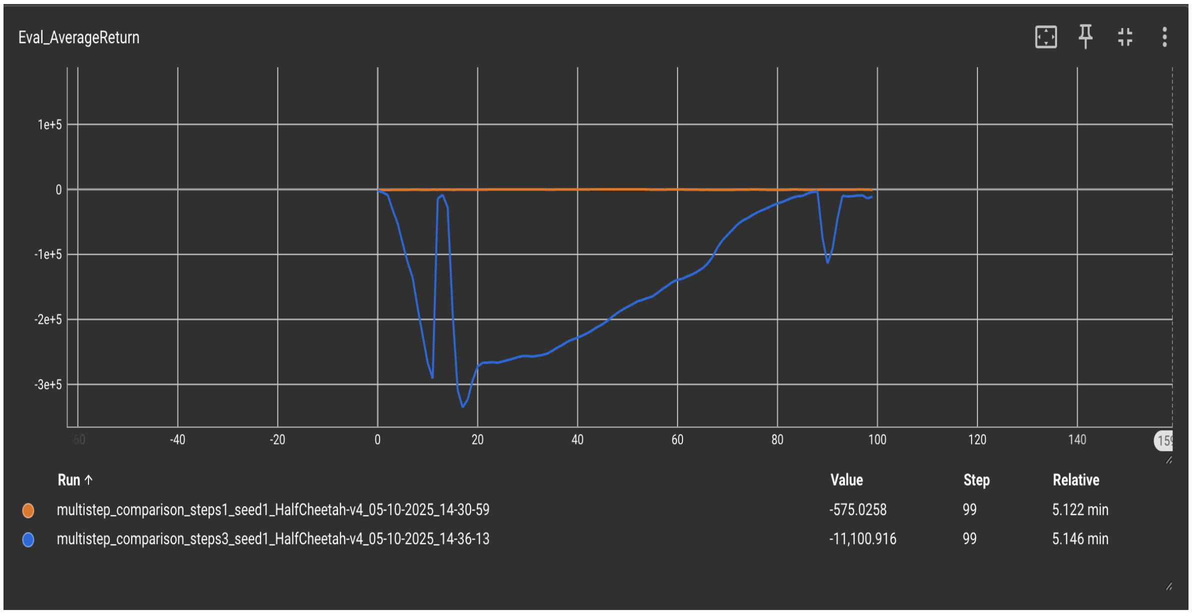
Q9.1

Difference in training time: 7.3 seconds

```
python rob831/scripts/run_hw2.py --env_name HalfCheetah-v4 \
-n 10 -b 10000 -lr 0.02 --exp_name test_cheetah_no_parallel

python rob831/scripts/run_hw2.py --env_name HalfCheetah-v4 \
--num_workers 4 -n 10 -b 10000 -lr 0.02 --exp_name test_cheetah_parallel
```

9.2 Multiple gradient steps – [1 points]

Q9.1



	Run ↑	Value	Step	Relative
●	multistep_comparison_steps1_seed1_HalfCheetah-v4_05-10-2025_14-30-59	-575.0258	99	5.122 min
●	multistep_comparison_steps3_seed1_HalfCheetah-v4_05-10-2025_14-36-13	-11,100.916	99	5.146 min

```
# Steps=1
python rob831/scripts/run_hw2.py --env_name "HalfCheetah-v4" --exp_name "multistep_comparison_steps1_seed1"
--n_iter 100 --batch_size 5000 --eval_batch_size 5000 --learning_rate 0.02 --discount 0.95 --n_layers 2 --size 64
↪  --seed 1
--num_policy_gradient_steps_per_batch 1 --reward_to_go --nn_baseline --scalar_log_freq 1

# Steps=3
python rob831/scripts/run_hw2.py --env_name "HalfCheetah-v4" --exp_name "multistep_comparison_steps3_seed1"
--n_iter 100 --batch_size 5000 --eval_batch_size 5000 --learning_rate 0.02 --discount 0.95 --n_layers 2 --size 64
↪  --seed 1
--num_policy_gradient_steps_per_batch 3 --reward_to_go --nn_baseline --scalar_log_freq 1
```