CS285 - HW2

Ho Nam Nguyen

September 26, 2022

1 Experiment 1

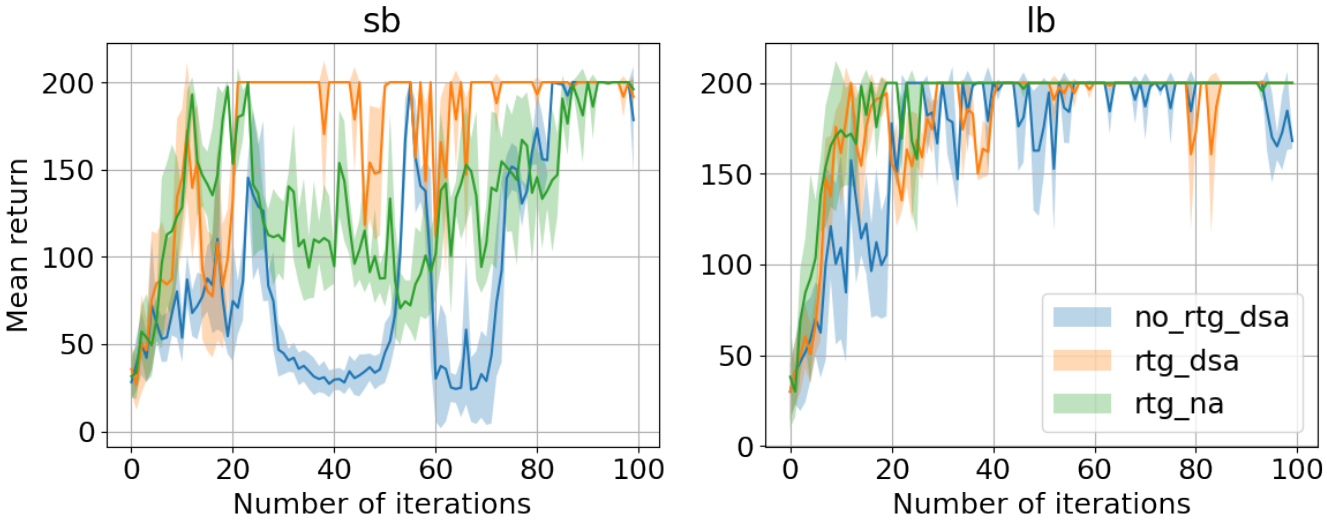

Figure 1

Answers

- Without advantage-standardization: reward-to-go (orange curve) performs better.
- Advantage-standardization doesn't help (green worse than orange)
- Batchsize helps training faster and more stable

Command lines

```
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
```

2 Experiment 2

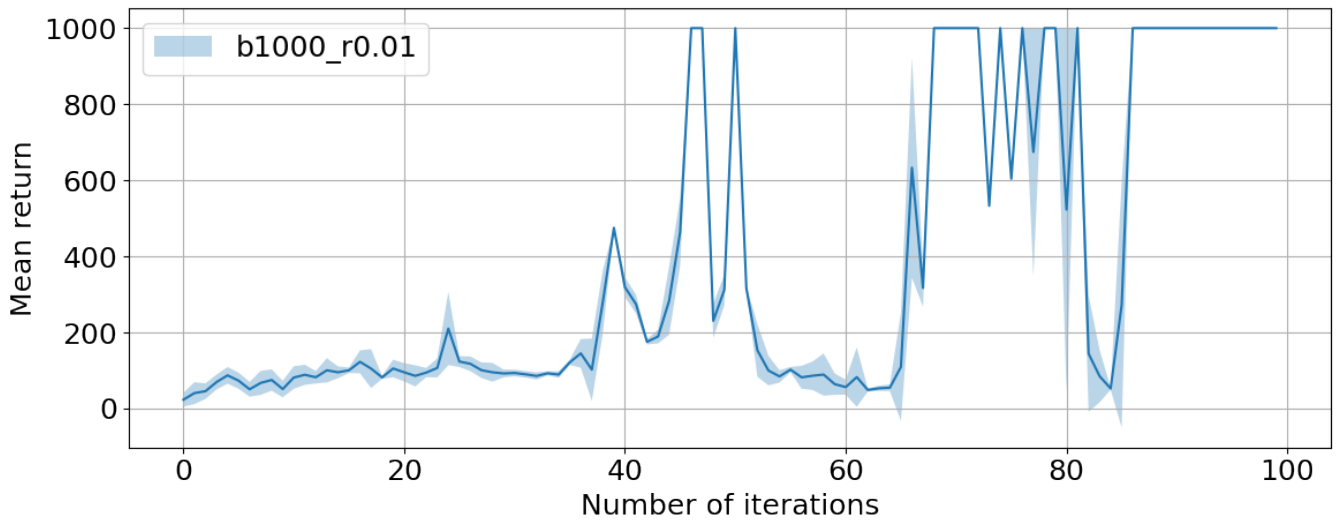

Figure 2

Command lines

python cs285/scripts/run_hw2.py --env_name InvertedPendulum-v4 --ep_len 1000 --discount 0.9 -n 100 -l 2 -s 64 -b 1000 -lr 0.01 -rtg --exp_name q2_b1000_r0.01

3 Experiment 3

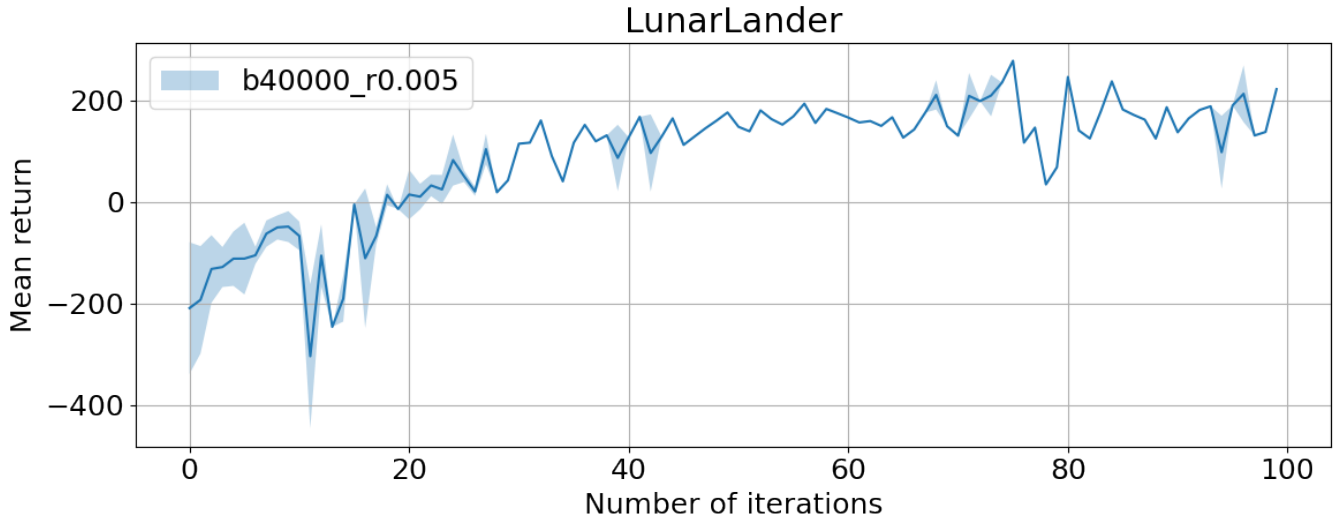

Figure 3

Command lines

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

4 Experiment 4

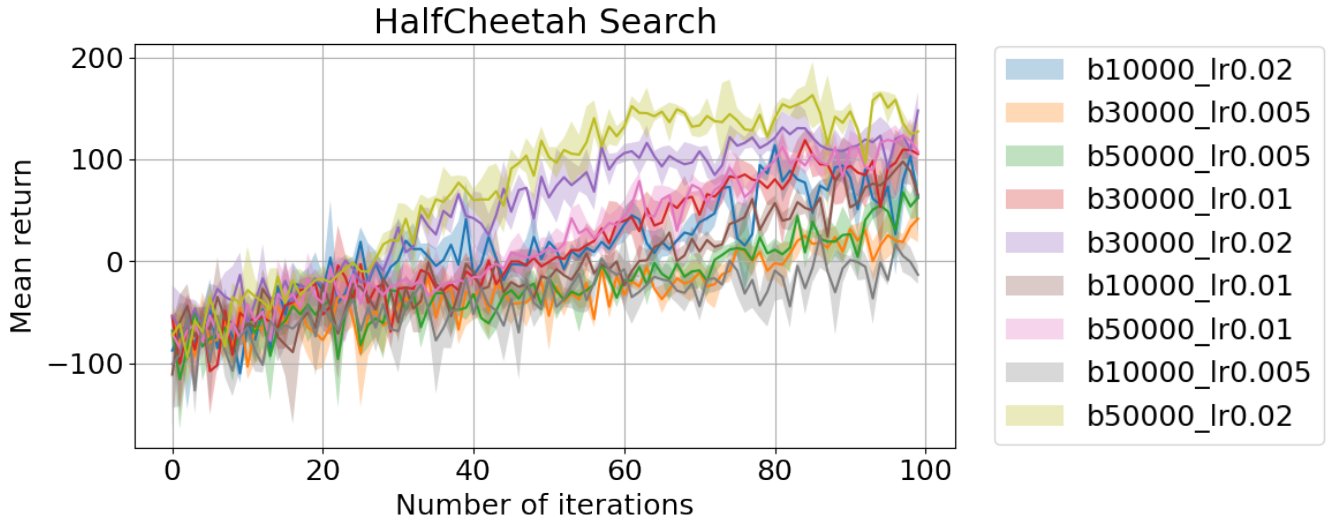

Figure 4

Answer

Larger batch size and larger learning rate generally perform better in this case. Best case is with batchsize=50000 and lr=0.02.

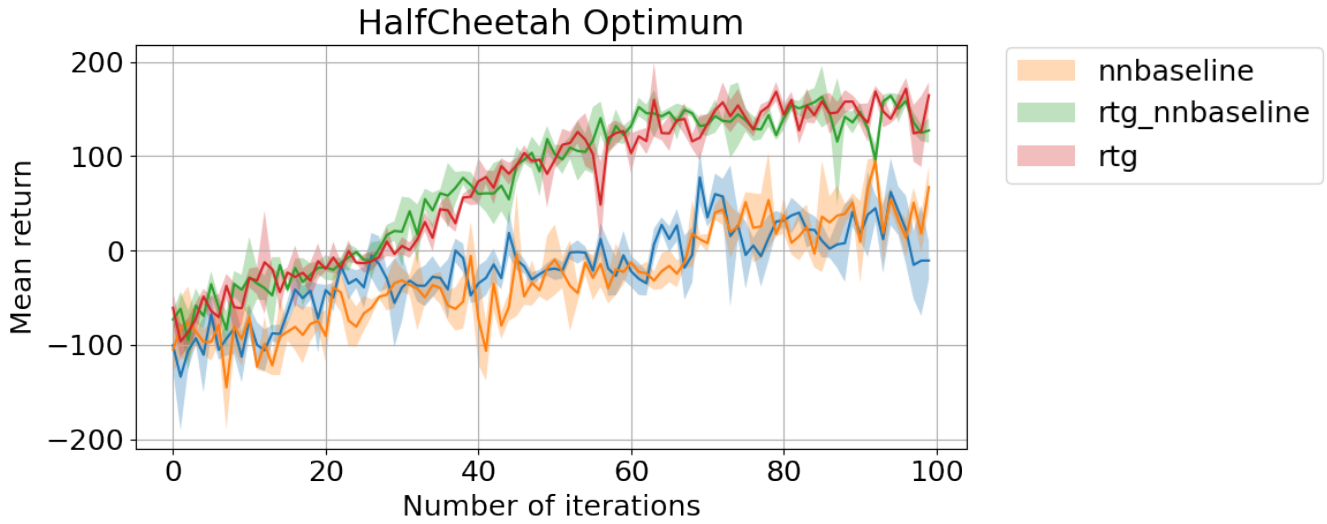

Figure 5

5 Experiment 5

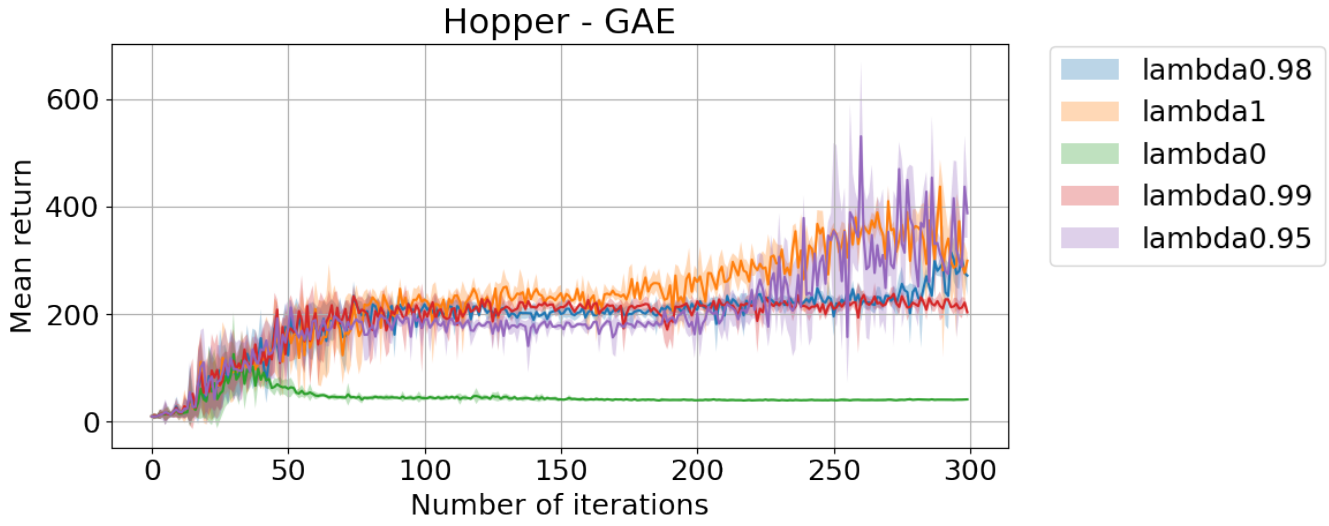

Figure 6

Answer

$\lambda = 0$ performs the worst whereas $\lambda = 0.95, 1$ perform the best. Increasing λ seems to stabilize the performance, i.e. smaller variance in the return.