Assignment 2: Policy Gradients

1 Vanilla Policy Gradients

1.1 The learning curves

From Figure 1 we can see that in the short batch experiments, using reward-to-go and standardizing the advantages can improve the performance of the model.

From Figure 2 we can see that in the long batch experiments, using reward-to-go helps reduce the variance of our model.

1.2 Questions

1. The one using reward-to-go has better performance without advantage-standardization.

2. Yes, advantage standardization helps reduce the variance, making the learning curves smoother.

3. Yes, small batch size leads to slow convergence.

1.3 Command line configuration

```
python cs285/scripts/run_hw2_policy_gradient.py −−env_name CartPole−v0 −n
    100 −b 1000 −dsa −−exp_name sb_no_rtg_dsa
python cs285/scripts/run_hw2_policy_gradient.py −−env_name CartPole−v0 −n
    100 −b 1000 −rtg −dsa −−exp_name sb_rtg_dsa
python cs285/scripts/run_hw2_policy_gradient.py −−env_name CartPole−v0 −n
    100 −b 1000 −rtg −−exp_name sb_rtg_na
python cs285/scripts/run_hw2_policy_gradient.py −−env_name CartPole−v0 −n
    100 −b 5000 −dsa −−exp_name lb_no_rtg_dsa
python cs285/scripts/run_hw2_policy_gradient.py −−env_name CartPole−v0 −n
    100 −b 5000 −rtg − dsa −−exp_name lb_rtg_dsa
python cs285/scripts/run_hw2_policy_gradient.py −−env_name CartPole−v0 −n
    100 −b 5000 −rtg −−exp_name lb_rtg_na
```

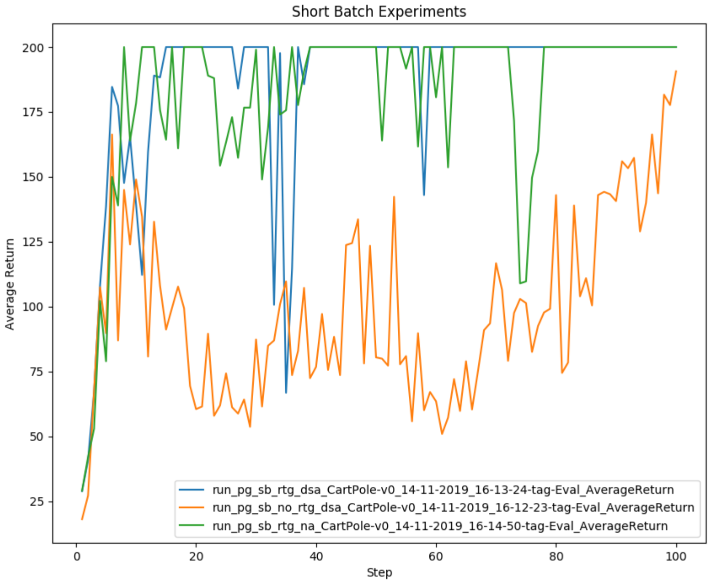

Figure. 1: The learning curves of the small batch experiments

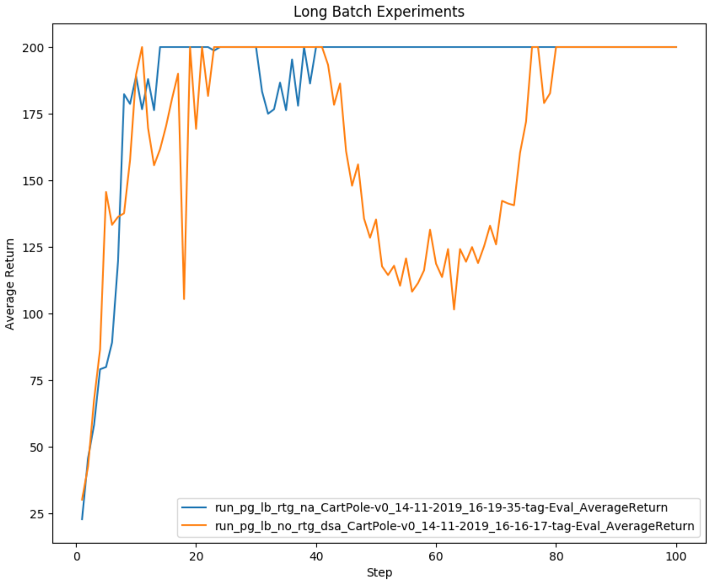

Figure. 2: The learning curves of the long batch experiments