1 Problem 1

Proof goes here.

2 Problem 4

Figure 1: Learning curves for small batch experiments.

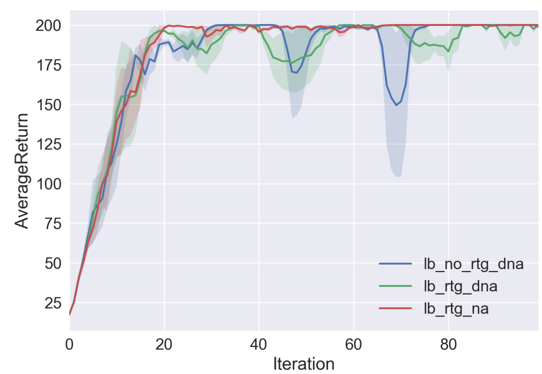

Figure 2: Learning curves for large batch experiments.

Answers:

1. Which gradient estimator has better performance without advantage-centering, the trajectory-centric one, or the one using reward-to-go?
   The one using reward-to-go have a better performance. From the learning curves for small batch experiments, we can see the green curve(reward-to-go) has a high average return than the blue curve(trajectory-centric).

2. Did advantage centering help?
   It helps. From the learning curves for small batch experiments, we can see the red curve(with advantage-centering) fluctuates less than the green curve(without advantage-centering).

3. Did the batch size make an impact?
   Yes, by comparing the learning curves between small batch experiments and large batch experiments, we find large batch experiments converge more quickly.

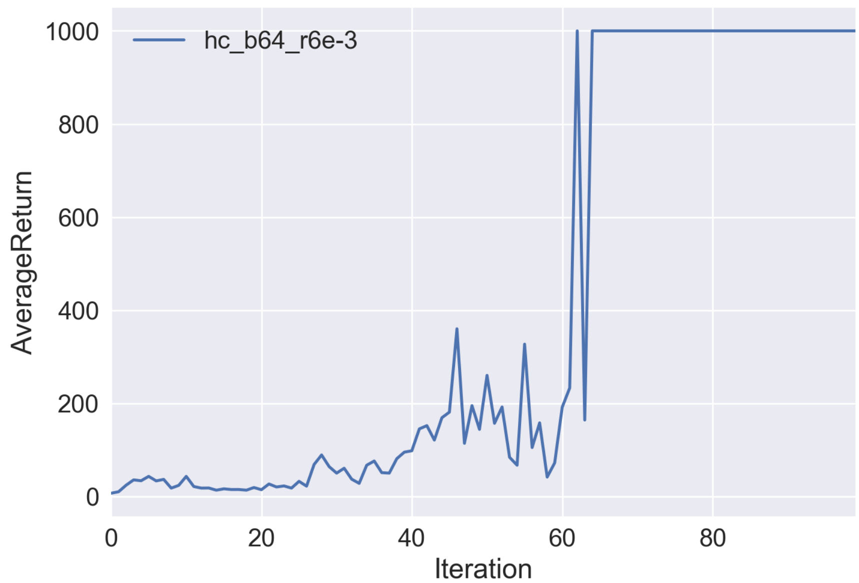

Figure 3: Learning curve with $b = 64$ and $lr = 0.006$. The policy gets to optimum at about iteration #65.

3 Problem 5

4 Problem 7

5 Problem 8

After a 3×3 grid search, the best parameter set is $b = 50000, r = 0.02$.
Answer: How did the batch size and learning rate affect the performance?
Large batch size will help the learning curve use less iterations to converge. Using a small learning rate can make sure not to miss any local minimum, but adjust the learning rate larger properly can help the performance improve more quickly.

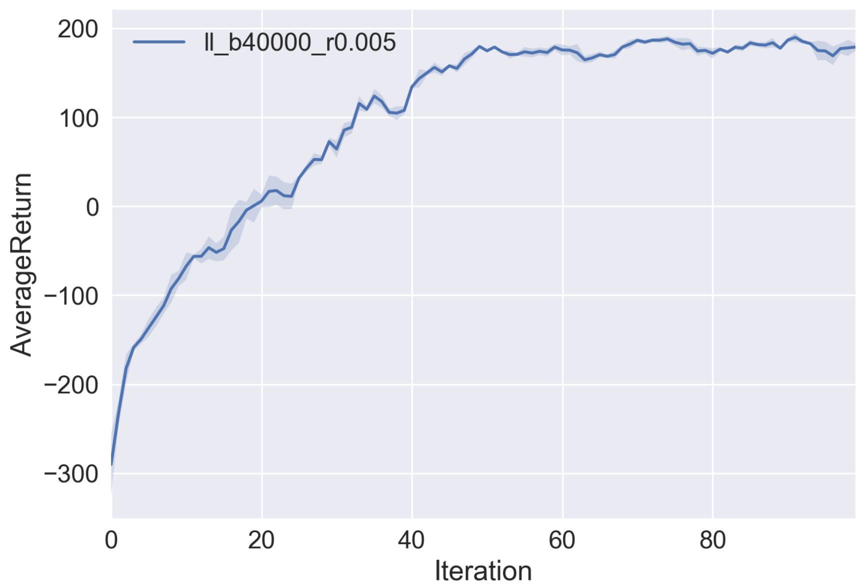

Figure 4: Learning curve for LunarLander. The policy finally achieved an average return of around 180.

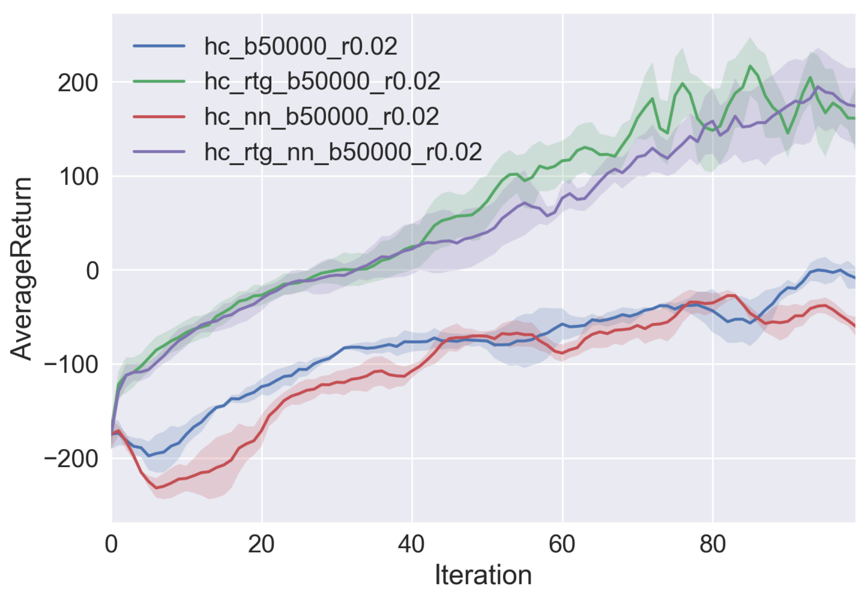

Figure 5: Learning curve for HalfCheetah with different parameters.