1 Problem 1

(a)

$$
\begin{aligned}
&\nabla_\theta \mathbb{E}_{s_t, a_t \sim p(s_t, a_t)}[b_{(s_t)}] \\
=&\nabla_\theta \mathbb{E}_{s_t, a_t}[\mathbb{E}_{a_t \sim \pi_\theta(a_t|s_t)}[b(s_t)]] \\
=&\mathbb{E}[\nabla_\theta \int \pi_\theta(a_t|s_t) b(s_t) da_t] \\
=&\mathbb{E}_{s_t \sim p(s_t)}[b(s_t) \cdot \int \nabla_\theta \pi_\theta(a_t|s_t) da_t] \\
=&\mathbb{E}_{s_t \sim p(s_t)}[b(s_t) \cdot \nabla_\theta \cdot \underbrace{\int \pi_\theta(a_t|s_t) da_t]}_{=1} \\
=&\mathbb{E}_{s_t \sim p(s_t)}[b(s_t) \cdot \nabla_\theta] \\
=&0
\end{aligned} \tag{1}
$$

(b)

(1) Because this sequence is a markov decision process. Current state is only affected by last state. So conditioning on $(s_1, a_1, ..., a_{t^*-1}, s_{t^*})$ is equivalent to conditioning on s_{t^*}.

(2)

$$
\begin{aligned}
&\nabla_\theta \mathbb{E}_{\tau \sim \pi_\theta(\tau)}[b(s_t^*)] \\
=&\nabla_\theta \cdot \int p(a_{t^*}|s_{t^*} \cdot s_{t^*-1} \cdot a_{t^*-1} \cdot \cdots \cdot s_1) p(s_{t^* \cdot s_{t^*-1} \cdot \cdots \cdot s_1}) b(s_{t^*}) da_{t^*} ds_{t^*} \cdot \cdots \cdot ds_1 \\
=&\nabla_\theta [\int p_\theta(a_{t^*}|s_{t^*}) \cdot p(s_{t^*}|...) b(s_{t^*}) ds_{t^*} da_t] \underbrace{[\int p(s_{t^*-1} \cdot \cdots \cdot s_1) ds_{t^*-1} \cdot \cdots \cdot ds_1]}_{=1} \\
=&\underbrace{\nabla_\theta \int p_\theta(a_{t^*}|s_{t^*}) da_{t^*}}_{=0} \cdot \mathbb{E}_{s_{t^*} \sim p(s_{t^*})}[b(s_{t^*})]]) \cdot 1 \\
=&0
\end{aligned} \tag{2}
$$

2 Problem 4

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

Figure 1: Learning curves for small batch experiments.

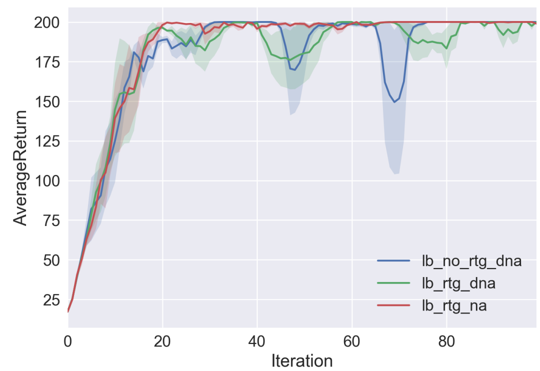

Figure 2: Learning curves for large batch experiments.

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.

3 Problem 5

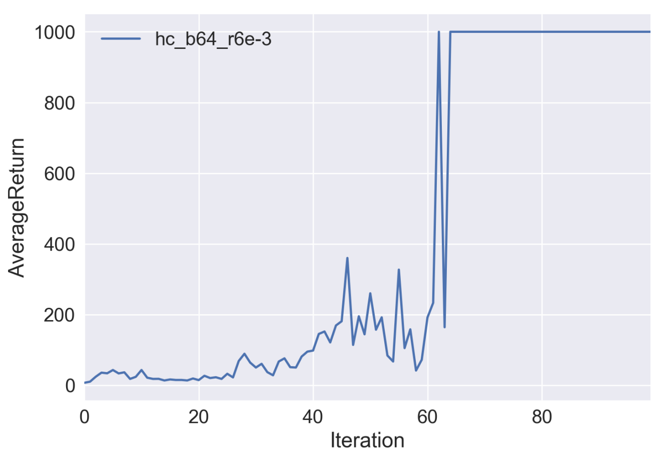

Figure 3: Learning curve with $b = 64$ and $lr = 0.006$. The policy gets to optimum at about iteration #65.

4 Problem 7

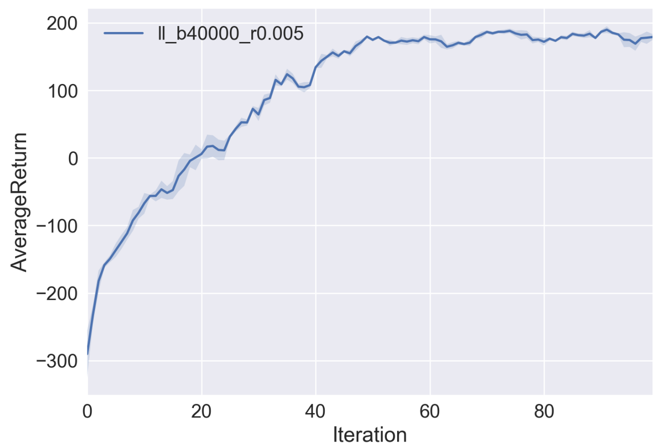

Figure 4: Learning curve for LunarLander. The policy finally achieved an average return of around 180.

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.

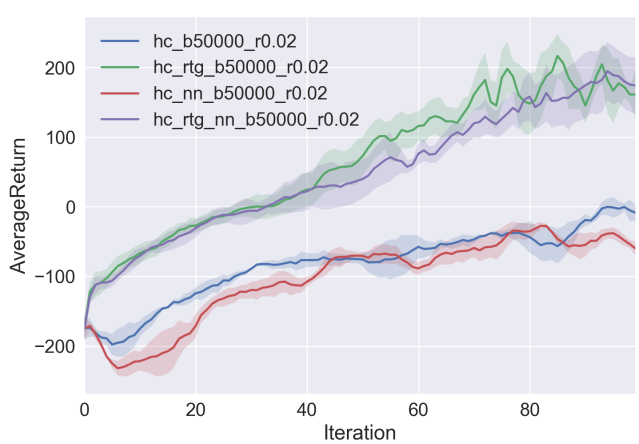

Figure 5: Learning curve for HalfCheetah with different parameters.