CS294–Fall 2018 — Homework 2Solutions

Jiajian Lu, SID 3033084290

2. Review

(a)

$$\begin{aligned}
\mathbb{E}_{\tau\sim p_\theta(\tau)}[\nabla_\theta \log \pi_\theta(a_t|s_t)b(s_t)] &= \mathbb{E}_{\tau\sim p_\theta(s_t,a_t)p_\theta(\tau/s_t,a_t|s_t,a_t)}[\nabla_\theta \log \pi_\theta(a_t|s_t)b(s_t)] \\
&= \mathbb{E}_{\tau/s_t,a_t\sim p_\theta(\tau/s_t,a_t|s_t,a_t)}[\mathbb{E}_{s_t,a_t\sim p_\theta(s_t,a_t)}[\nabla_\theta \log \pi_\theta(a_t|s_t)b(s_t)|s_t,a_t]] \\
&= \mathbb{E}_{\tau/s_t,a_t\sim p_\theta(\tau/s_t,a_t|s_t,a_t)}[b(s_t)\sum_{a_t}\sum_{s_t} p_\theta(s_t,a_t)\nabla_\theta \log \pi_\theta(a_t|s_t)] \\
&= \mathbb{E}_{\tau/s_t,a_t\sim p_\theta(\tau/s_t,a_t|s_t,a_t)}[b(s_t)\sum_{a_t} \pi_\theta(a_t|s_t)\nabla_\theta \log \pi_\theta(a_t|s_t)] \\
&= \mathbb{E}_{\tau/s_t,a_t\sim p_\theta(\tau/s_t,a_t|s_t,a_t)}[b(s_t)\sum_{a_t} \nabla_\theta \pi_\theta(a_t|s_t)] \quad \text{(convenient identity)} \\
&= \mathbb{E}_{\tau/s_t,a_t\sim p_\theta(\tau/s_t,a_t|s_t,a_t)}[b(s_t)0] \\
&= 0
\end{aligned}$$

So

$$\sum_{t=1}^{T} \mathbb{E}_{\tau\sim p_\theta(\tau)}[\nabla_\theta \log \pi_\theta(a_t|s_t)b(s_t)] = 0$$

(b) (a) Because the trajectory is a MDP and the probability of future trajectory only depends on the most recent state.

(b) $p(a_t, s_{t+1}, \cdots, a_{T-1}, s_T) = \pi_\theta(a_t|s_t)p(s_{t+1}|s_t,a_t), \cdots, p(s_T|a_{T-1}, s_{T-1})$

$$\begin{aligned}
&\mathbb{E}_{\tau\sim p_\theta(\tau)}[\nabla_\theta \log \pi_\theta(a_t|s_t)b(s_t)] \\
&= \mathbb{E}_{s_{1:t},a_{1:t-1}}[\mathbb{E}_{s_{t+1:T},a_{t:T}}\nabla_\theta \log \pi_\theta(a_t|s_t)b(s_t)] \\
&= \mathbb{E}_{s_{1:t},a_{1:t-1}}[\sum_{a_t}\cdots\sum_{s_T} p(a_t, s_{t+1}, \cdots, a_{T-1}, s_T|s_t)\nabla_\theta \log \pi_\theta(a_t|s_t)b(s_t)] \\
&= \mathbb{E}_{s_{1:t},a_{1:t-1}}[\sum_{a_t}\cdots\sum_{s_T} \pi_\theta(a_t|s_t)p(s_{t+1}|s_t,a_t), \cdots, p(s_T|a_{T-1}, s_{T-1})\nabla_\theta \log \pi_\theta(a_t|s_t)b(s_t)] \\
&= \mathbb{E}_{s_{1:t},a_{1:t-1}}[b(s_t)\sum_{a_t} \pi_\theta(a_t|s_t)\nabla_\theta \log \pi_\theta(a_t|s_t)] \quad \text{(other summations are all 1)} \\
&= \mathbb{E}_{s_{1:t},a_{1:t-1}}[b(s_t)0] \\
&= 0
\end{aligned}$$

So

$$\sum_{t=1}^{T} \mathbb{E}_{\tau\sim p_\theta(\tau)}[\nabla_\theta \log \pi_\theta(a_t|s_t)b(s_t)] = 0$$

Problem 4

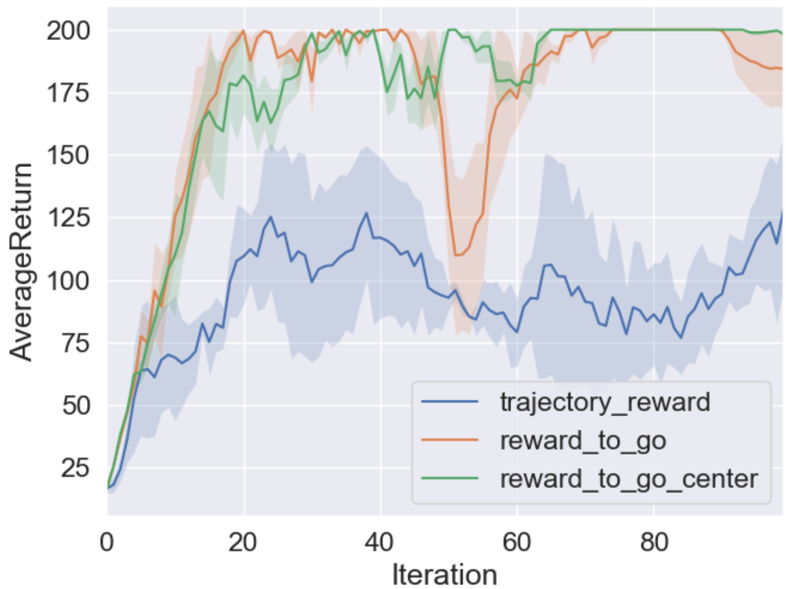

Figure 1: Small Batch size

(a)

Figure 2: Big Batch size

(b)

(c) Reward to go has better performance.

(d) Advantage centering does help.

(e) Yes.

(f)
```
python train_pg_f18.py CartPole-v0 -n 100 -b 1000 -e 3 -dna
--exp_name sb_no_rtg_dna

python train_pg_f18.py CartPole-v0 -n 100 -b 1000 -e 3 -rtg -dna
--exp_name sb_rtg_dna

python train_pg_f18.py CartPole-v0 -n 100 -b 1000 -e 3 -rtg
--exp_name sb_rtg_na

python train_pg_f18.py CartPole-v0 -n 100 -b 5000 -e 3 -dna
--exp_name lb_no_rtg_dna

python train_pg_f18.py CartPole-v0 -n 100 -b 5000 -e 3 -rtg -dna
--exp_name lb_rtg_dna

python train_pg_f18.py CartPole-v0 -n 100 -b 5000 -e 3 -rtg
--exp_name lb_rtg_na
```

Problem 5

(a) $b^* = 500$ and $r^* = 0.01$

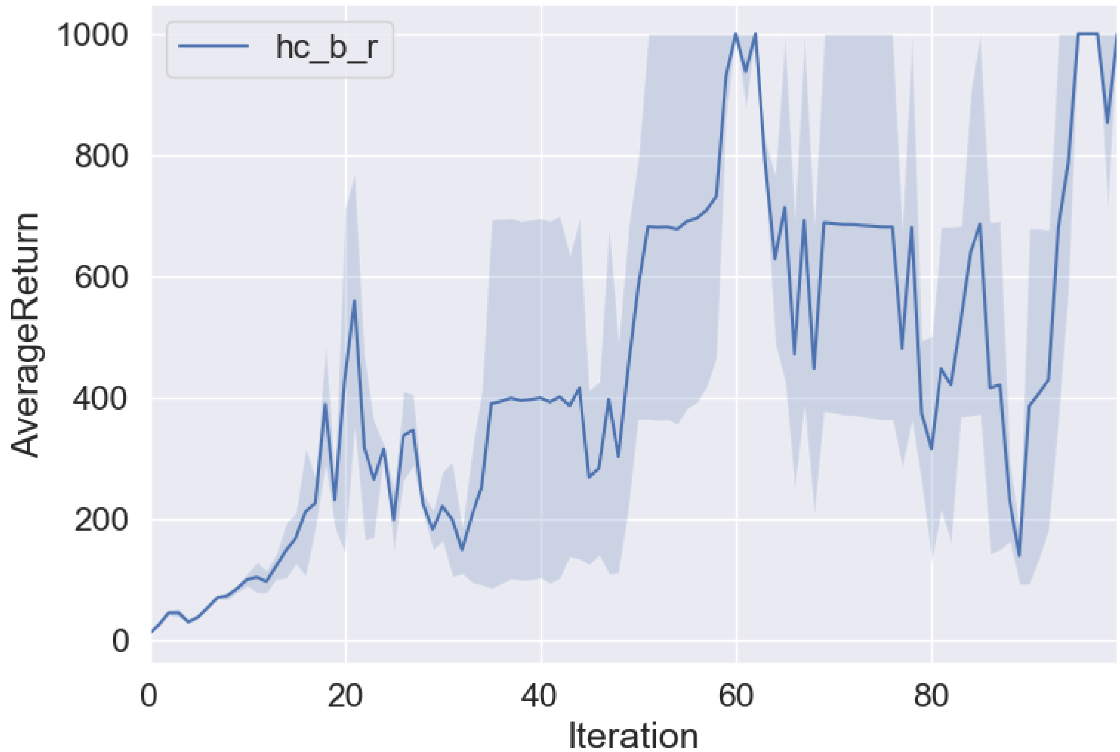

Figure 3: Inverted Pendulum

(b) `python train_pg_f18.py InvertedPendulum-v2 -ep 1000 --discount 0.9 -n 100 -e 3 -l 2 -s 64 -b 600 -lr 0.01 -rtg --exp_name hc_b_r`

7. Lunar Lander

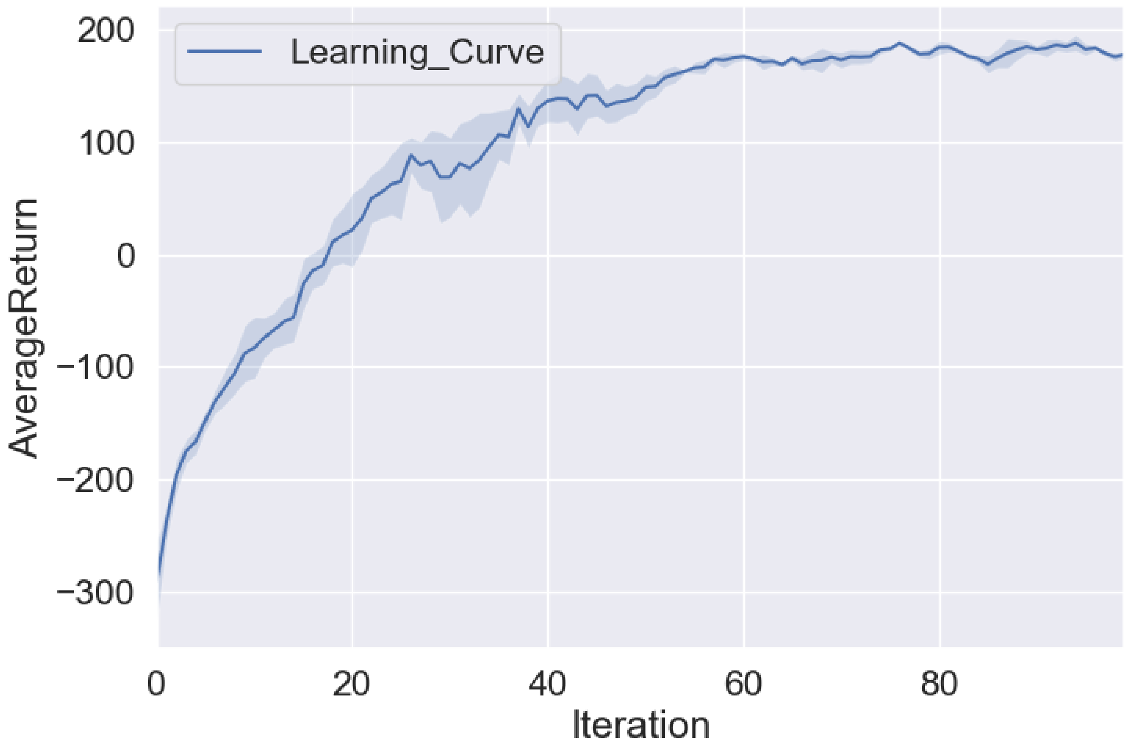

Figure 4: Inverted Pendulum

```
python train_pg_f18.py LunarLanderContinuous-v2 -ep 1000
--discount 0.99 -n 100 -e 3 -l 2 -s 64 -b 40000 -lr 0.005 -rtg
--nn_baseline --exp_name ll_b40000_r0.005
```

8. HalfCheetah

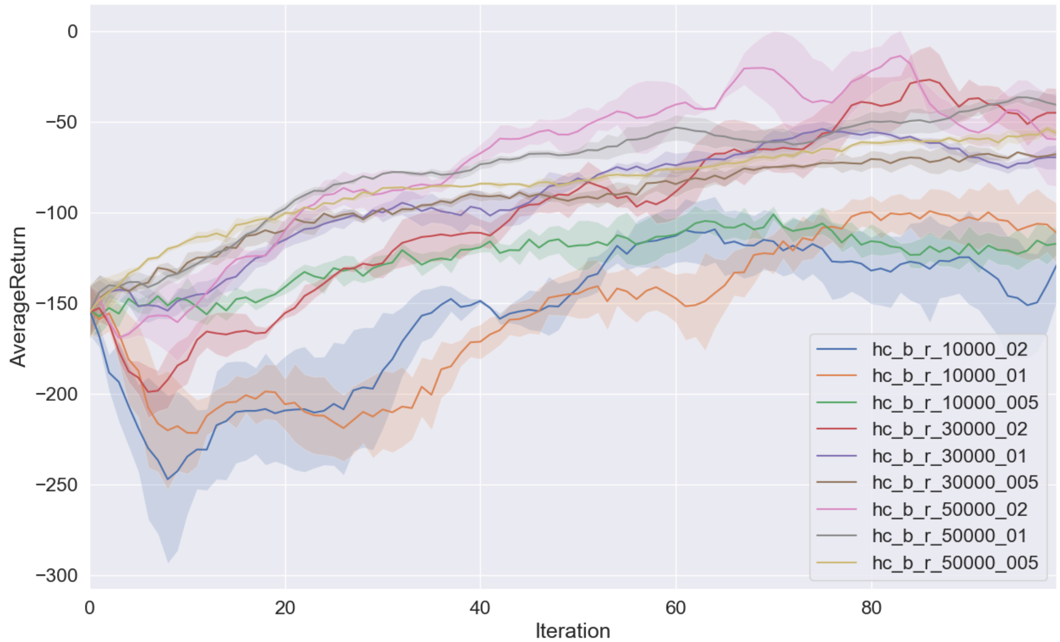

Figure 5: finding b* and lr*

(a) Larger batch size can have higher average return. Smaller learning rate can reduce the variance of the learning curve.

And I choose batch size = 50000 and learning rate = 0.01.

(b)
```
python train_pg_f18.py HalfCheetah-v2 -ep 150 --discount 0.95 -n 100
-e 3 -l 2 -s 32 -b 50000 -lr 0.01 --exp_name hc_b_r_095_50000_01

python train_pg_f18.py HalfCheetah-v2 -ep 150 --discount 0.95 -n 100
-e 3 -l 2 -s 32 -b 50000 -lr 0.01 -rtg --exp_name hc_b_r_095_50000_01_rtg

python train_pg_f18.py HalfCheetah-v2 -ep 150 --discount 0.95 -n 100
-e 3 -l 2 -s 32 -b 50000 -lr 0.01 --nn_baseline --exp_name hc_b_r_095_50000_01_nn

python train_pg_f18.py HalfCheetah-v2 -ep 150 --discount 0.95 -n 100
-e 3 -l 2 -s 32 -b 50000 -lr 0.01 -rtg --nn_baseline
--exp_name hc_b_r_095_50000_01_rtg_nn
```

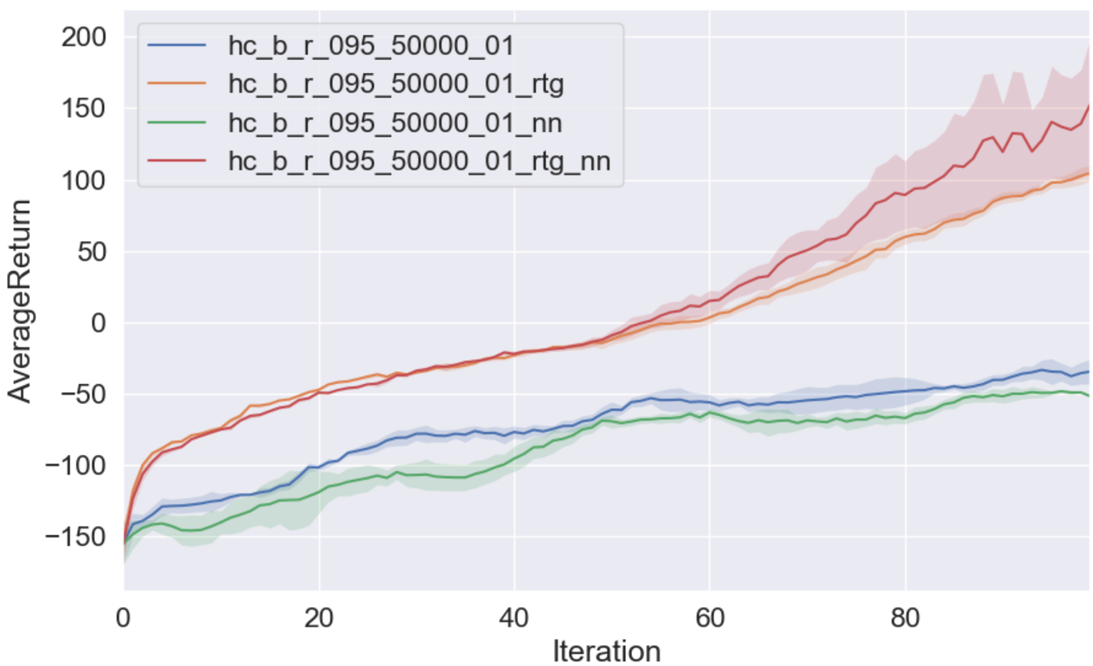

Figure 6: More Task