Part 3

1.Average return for cartpole experiments

Figure 1: Experiments run with cartpole

Figure 2: Experiments run with large batch cartpole

Did advantage normalization help?
- Yes, normalizing advantages allows the advantages to stabilize, keeping the average return stable once it peaks around 250.

Which value estimator has better performance without advantage normalization: the trajectory centric one, or the one using reward-to-go?
- reward to go performed better without advantage normalization

Did batch Size make an impact
it made the reward converge faster, so it helps with making the average return more reliable.

Part 4.2

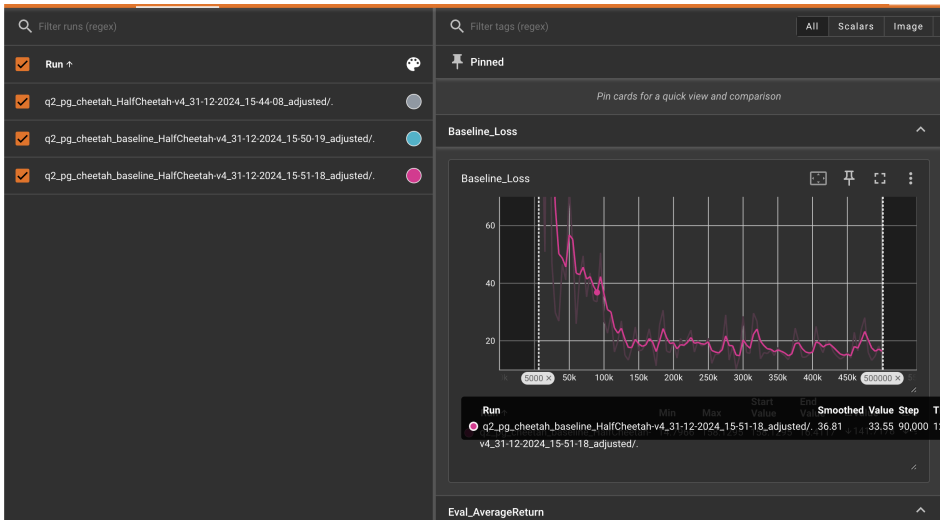


Figure 3: Baseline Loss for Cheetah With and without baseline

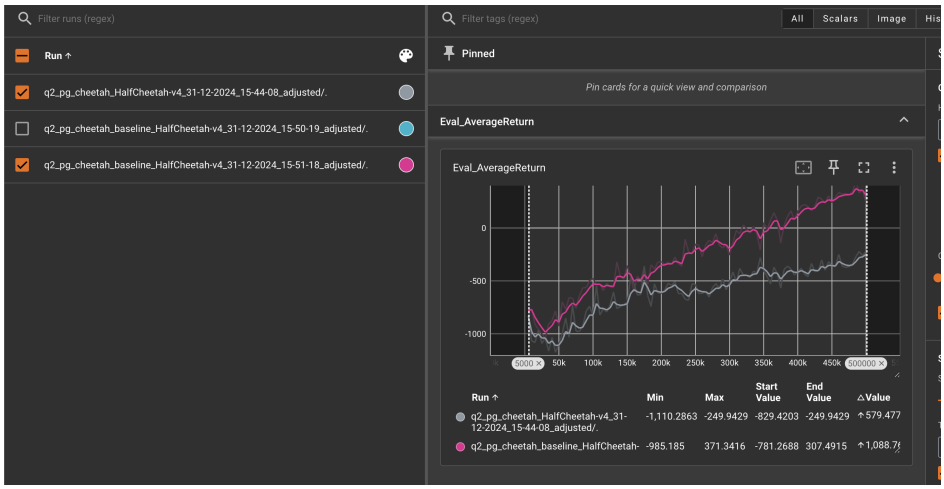


Figure 4: Return for Cheetah with and without baseline

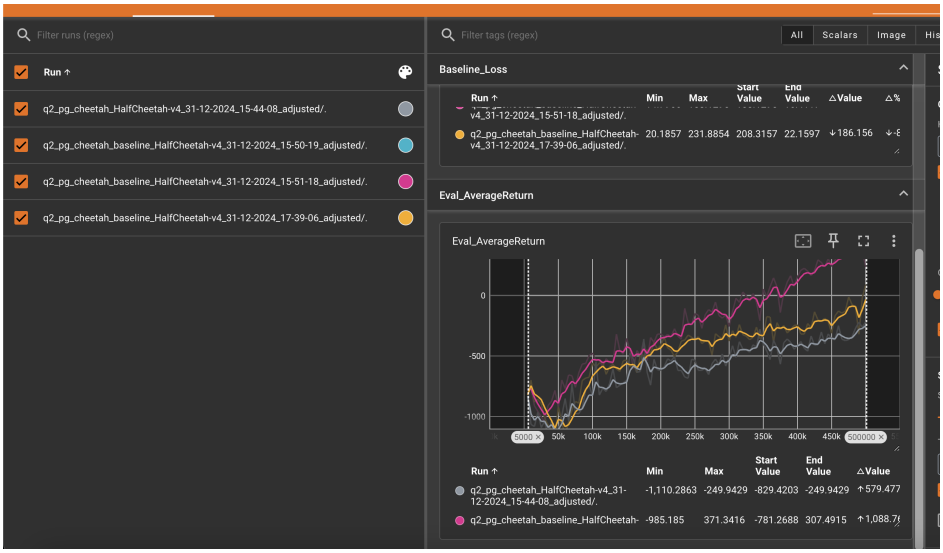


Figure 5: Return for Cheetah with higher blr and bgs (purple)

yellow line is case using baseline, lower blr and fewer bgs, as you can see it has lower average return compared to the baseline with higher blr and bgs (purple). Grey is no baseline

Part 5: Implementing GAE

```
\python cs285/scripts/run_hw2.py \
--env_name LunarLander-v2 --ep_len 1000 \
--discount 0.99 -n 300 -l 3 -s 128 -b 2000 -lr 0.001 \
--use_reward_to_go --use_baseline --gae_lambda 0 \
--exp_name lunar_lander_lambda0

python cs285/scripts/run_hw2.py \
--env_name LunarLander-v2 --ep_len 1000 \
--discount 0.99 -n 300 -l 3 -s 128 -b 2000 -lr 0.001 \
--use_reward_to_go --use_baseline --gae_lambda 0.95 \
--exp_name lunar_lander_lambda0.95

python cs285/scripts/run_hw2.py \
--env_name LunarLander-v2 --ep_len 1000 \
--discount 0.99 -n 300 -l 3 -s 128 -b 2000 -lr 0.001 \
--use_reward_to_go --use_baseline --gae_lambda 0.98 \
--exp_name lunar_lander_lambda0.98

python cs285/scripts/run_hw2.py \
--env_name LunarLander-v2 --ep_len 1000 \
--discount 0.99 -n 300 -l 3 -s 128 -b 2000 -lr 0.001 \
--use_reward_to_go --use_baseline --gae_lambda 0.99 \
--exp_name lunar_lander_lambda0.99

python cs285/scripts/run_hw2.py \
--env_name LunarLander-v2 --ep_len 1000 \
--discount 0.99 -n 300 -l 3 -s 128 -b 2000 -lr 0.001 \
--use_reward_to_go --use_baseline --gae_lambda 1 \
```

```
--exp_name lunar_lander_lambda1
```

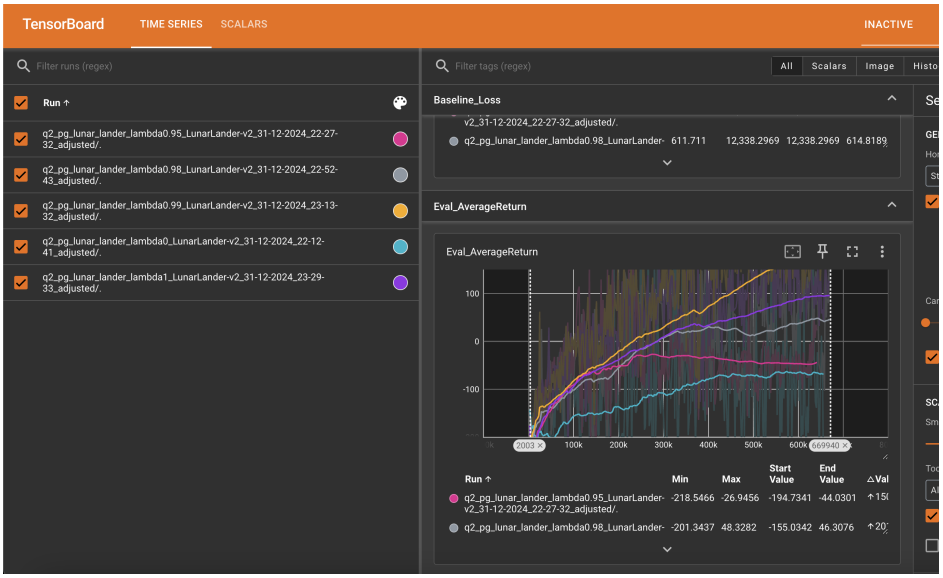


Figure 6: Effect of λ on LunarLander EvalReturn. Orange is $\lambda = 0.99$, blue is $\lambda = 0$

higher lambda represents weighing the reward of future actions in calculating the advantage the same as the current reward. if lambda is 0 that means we only consider the current action and its reward set whereas if lambda is 1 we weigh future rewards at a timestep the same as the current reward

Part 6: Hyperparameters and Sample Efficiency

Default, yellow on the graph

```
python cs285/scripts/run_hw2.py --env_name InvertedPendulum-v4 -n 100 \
--exp_name pendulum_default_s5 \
-rtg --use_baseline -na \
--batch_size 5000 \
--seed 5
```

lr 0.02, grey on the graph

```
python cs285/scripts/run_hw2.py \
--env_name InvertedPendulum-v4 --exp_name pendulum_0.02lr \
-n 100 --batch_size 5000 --seed 5 \
-lr 0.02 --discount 0.98 \
-rtg --use_baseline -na
```

lr 0.04, blue on the graph

```
python cs285/scripts/run_hw2.py \
--env_name InvertedPendulum-v4 --exp_name pendulum_0.04lr \
-n 100 --batch_size 6000 --seed 5 \
-lr 0.04 --discount 0.97 \
-rtg --use_baseline -na
```

lr 0.03, red on the graph

```
python cs285/scripts/run_hw2.py \
--env_name InvertedPendulum-v4 --exp_name pendulum_0.03lr \
-n 100 --batch_size 6000 --seed 5 \
-lr 0.03 --discount 0.98 \
-rtg --use_baseline -na
```

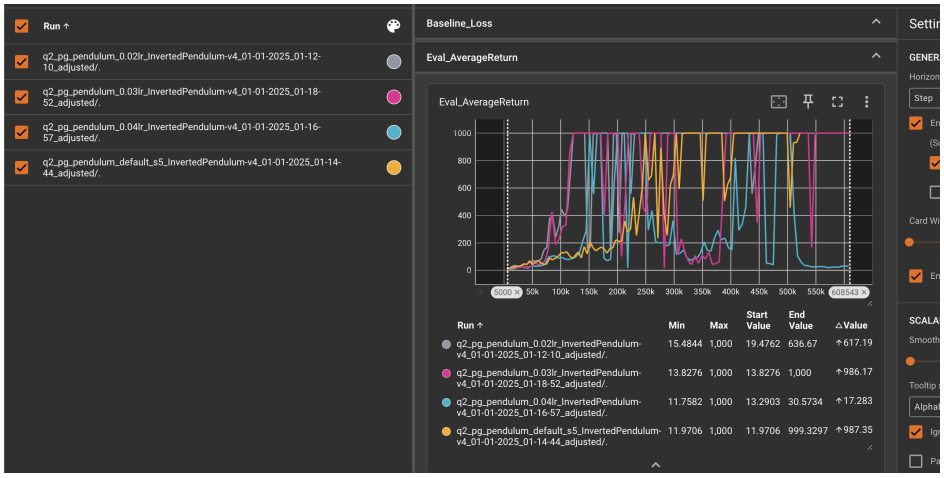
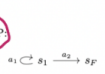


Figure 7: Parameters for optimal performance on Inverted Pendulum

8 Analysis

Consider the following infinite-horizon MDP:



At each step, the agent stays in state s_1 and receives reward 1 if it takes action a_1, and receives reward 0 and terminates the episode otherwise. Parametrize the policy as stationary (not dependent on time) with a single parameter:

$$\pi_\theta(a_1|s_1) = \theta, \pi_\theta(a_2|s_1) = 1 - \theta$$

1. Applying policy gradients
   (a) Use policy gradients to compute the gradient of the expected return $J(\theta) = \mathbb{E}_{\pi_\theta} R(\tau)$ with respect to the parameter θ. **Do not use discounting.**

   Hint: to compute $\sum_{k=1}^{\infty} k\alpha^{k-1}$, you can write:

$$\sum_{k=1}^{\infty} k\alpha^{k-1} = \sum_{k=1}^{\infty} \frac{d}{d\alpha}\alpha^k = \frac{d}{d\alpha}\sum_{k=1}^{\infty}\alpha^k$$

$\nabla_\theta J(\theta) = \mathbb{E}_{\tau\sim p_\theta(\tau)}\left[\nabla_\theta \log p_\theta(\tau)\cdot R(\tau)\right]$

$= \int p_\theta(\tau)\cdot\nabla_\theta \log p_\theta(\tau)\cdot R(\tau)\, d\tau$

$= \int \nabla_\theta p_\theta(\tau)\cdot R(\tau)\, d\tau$ (∂: differentiate over τ; $\to p_\theta(s_1, a_1, s_F)$)

$= \sum_{k=1}^{\infty} \frac{d}{d\theta}(\theta^{k-1})(1-\theta)\cdot(k-1)$

$= \frac{d}{d\theta}\sum_{k=1}^{\infty}\theta^{k-1}(1-\theta)\cdot(k-1)$

$= \frac{d}{d\theta}\left(\frac{-\theta}{\theta-1}\right) = \frac{d}{d\theta}\left(\frac{\theta}{1-\theta}\right) = \frac{1}{(\theta-1)^2}$

Figure 8: Enter Caption

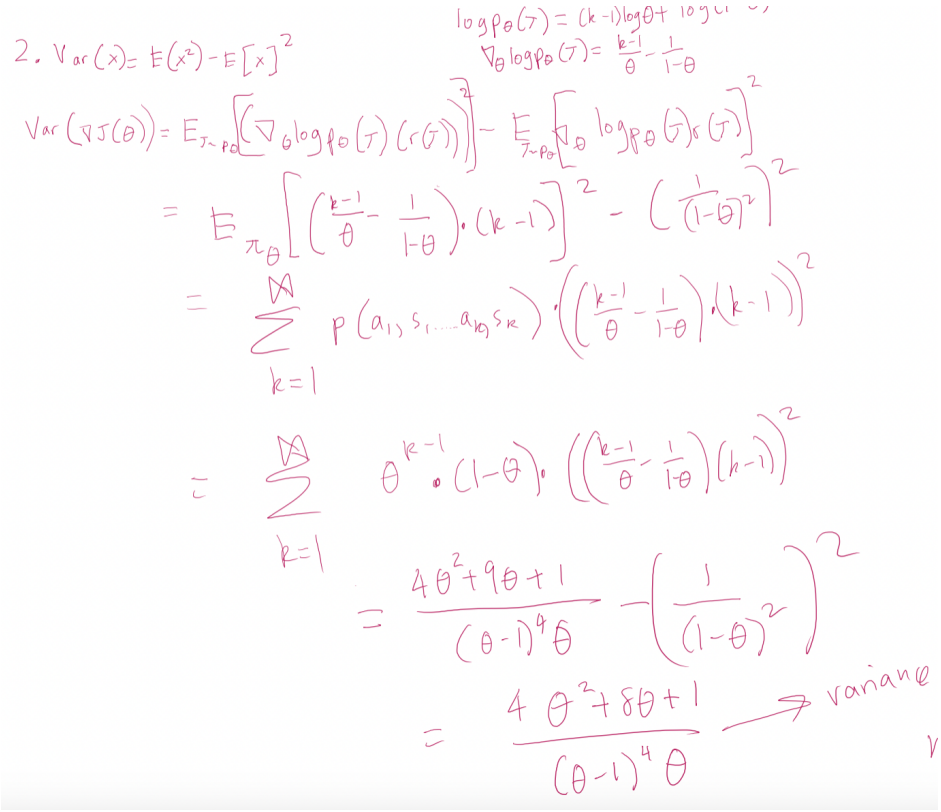


Figure 10: Enter Caption

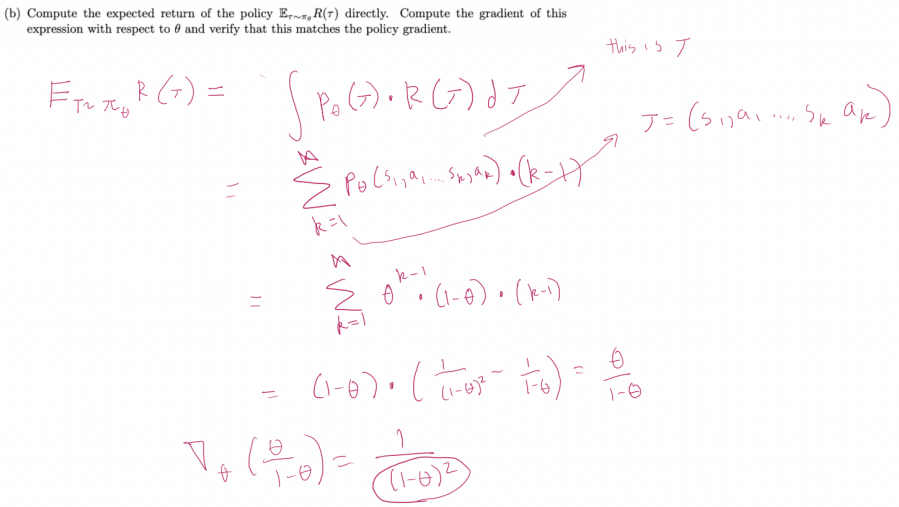


Figure 9: Enter Caption