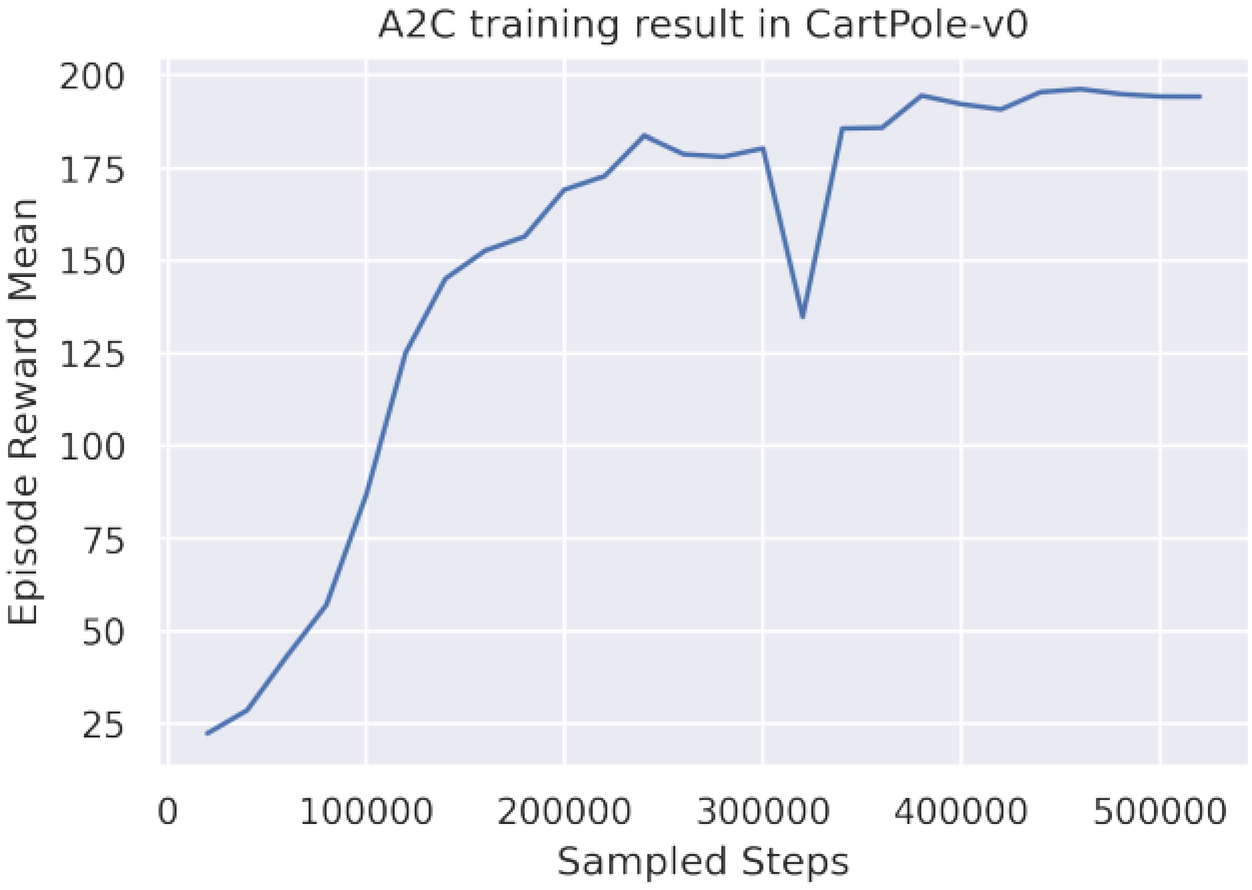
Result of Assignment 4 of IERG5350

- 30 points is already assigned to your implementation of BaseTrainer.
- 20 points is already assigned to your A2CTrainer, PPOTrainer and TD3Trainer, respectively.
- The learning curves should use the sampled timesteps as the X-coordinate and the episodic reward or the average success rate (you choose one) as the Y-coordinate.
- The MetaDrive Easy environment is `MetaDrive-Tut-Easy-v0` and the Hard environment is `MetaDrive-Tut-Hard-v0` , which is identical to the `MetaDrive-Tut-Test-v0` used in the generalization experiment.

Learning curves of A2C

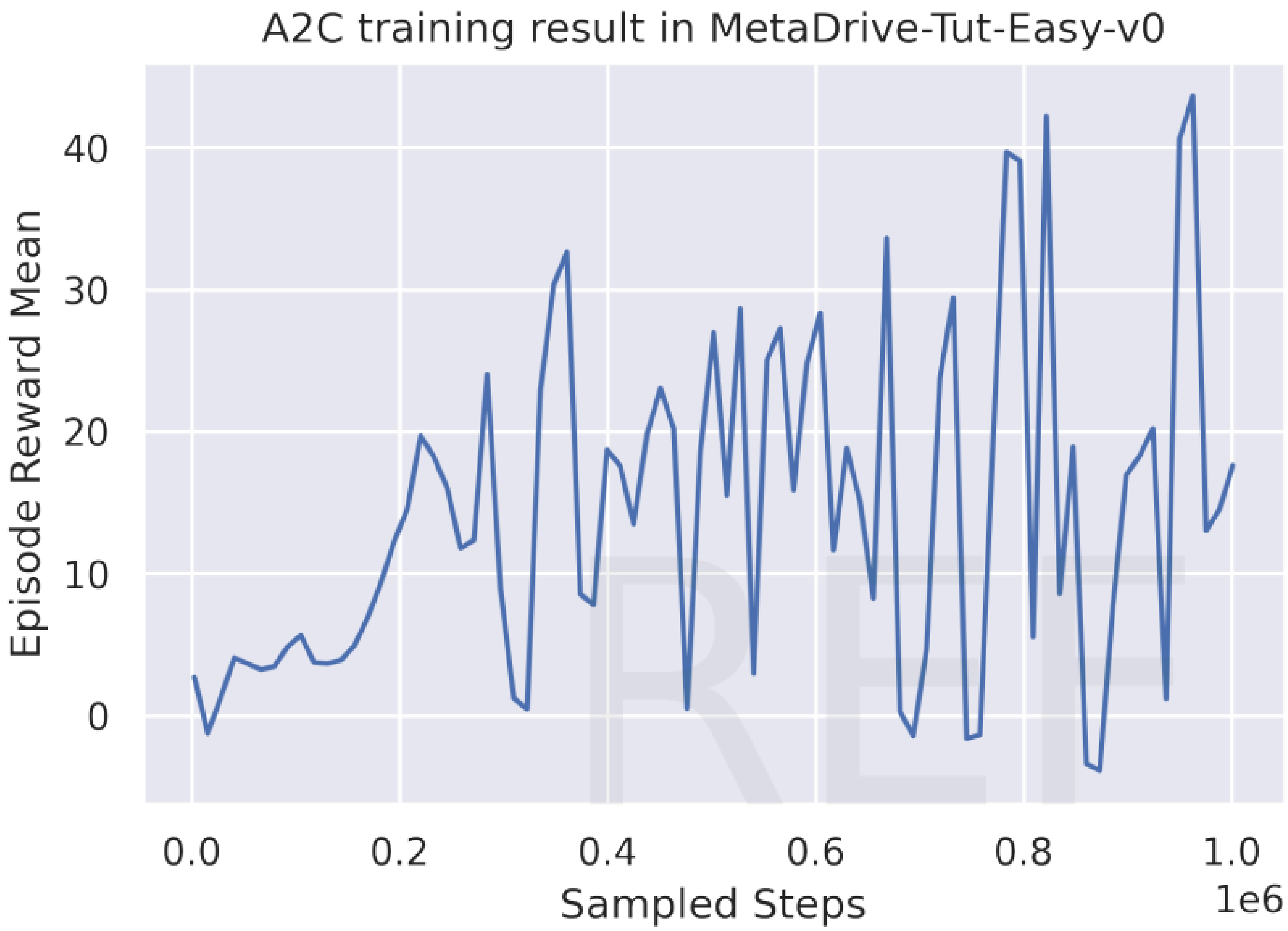
A2C in CartPole

(5 points)



A2C in MetaDrive Easy

(5 points)



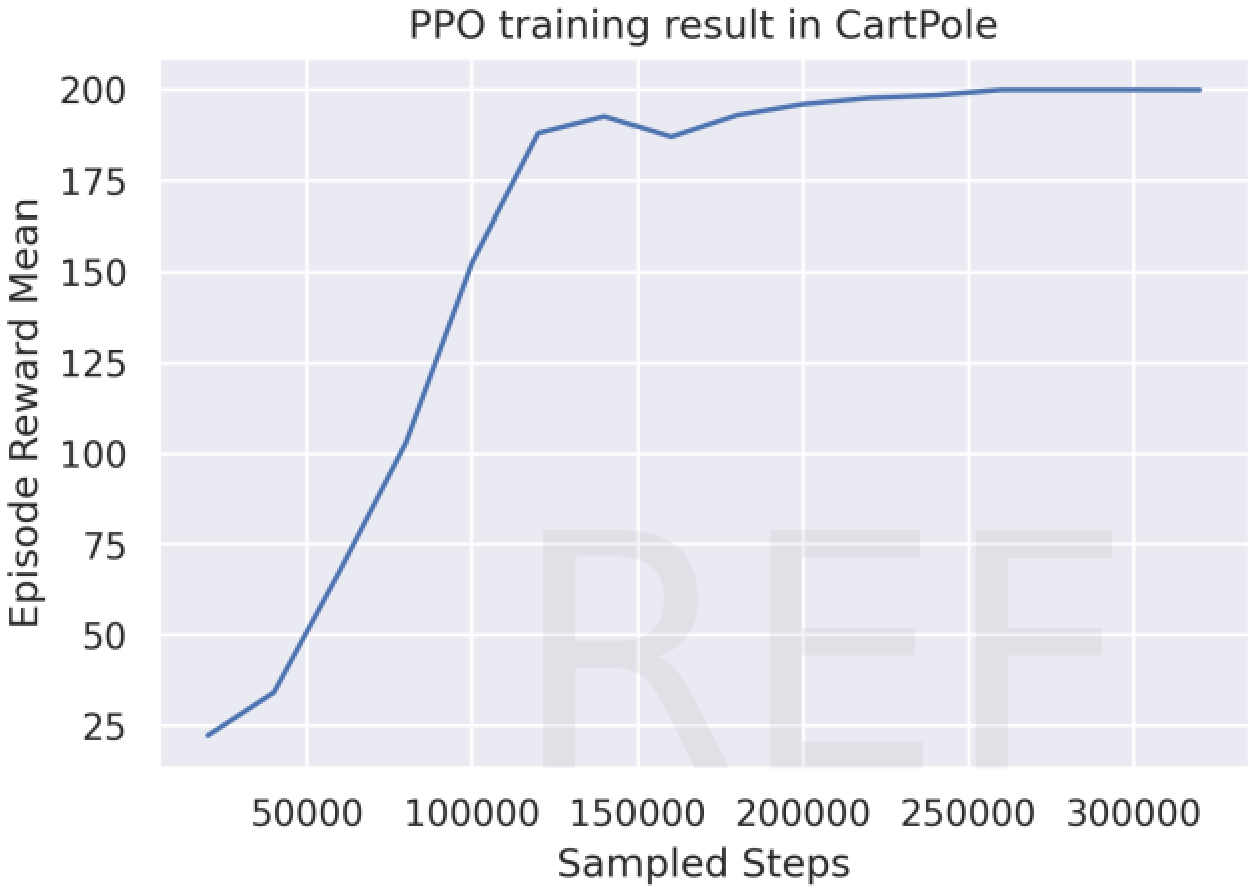
REF

Learning curves of PPO

We require at least 10M steps to train PPO.

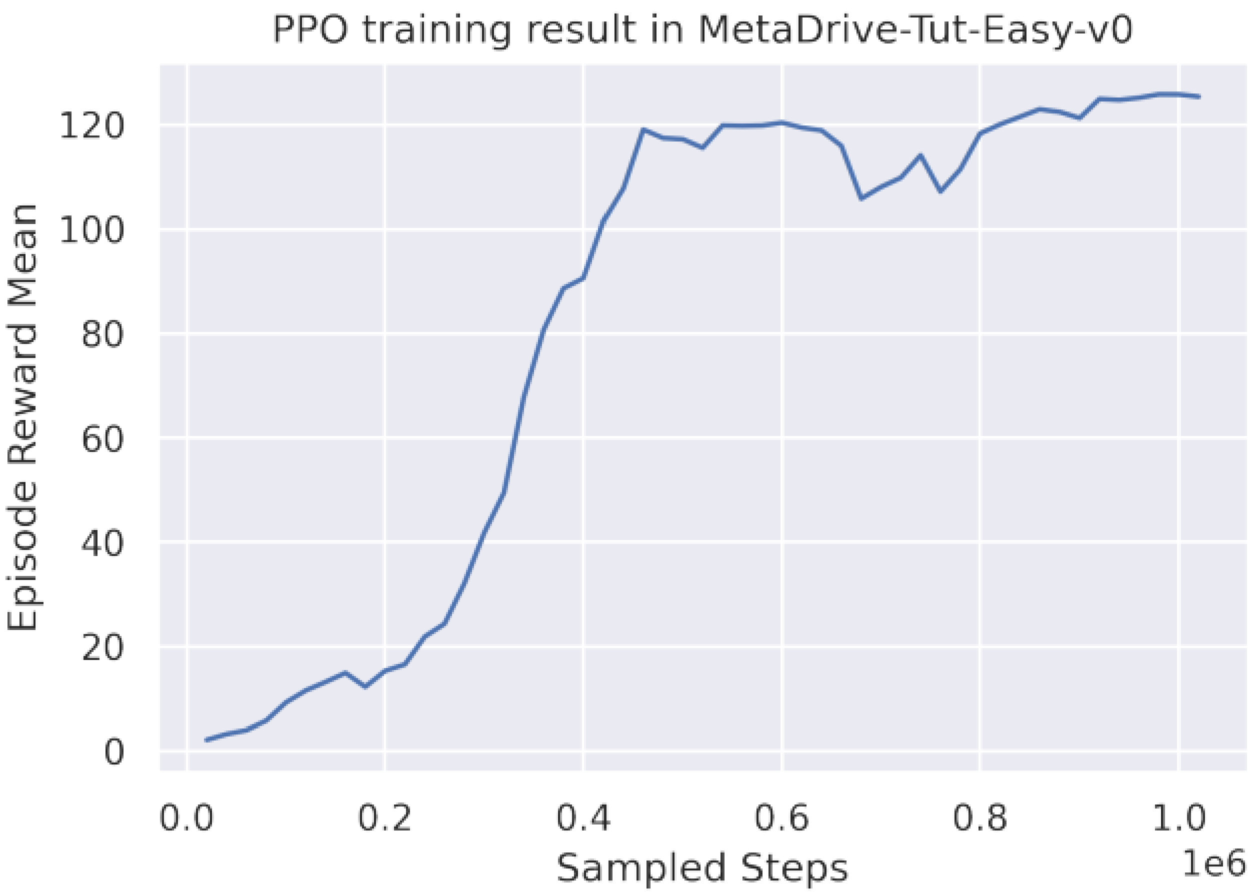
PPO in CartPole

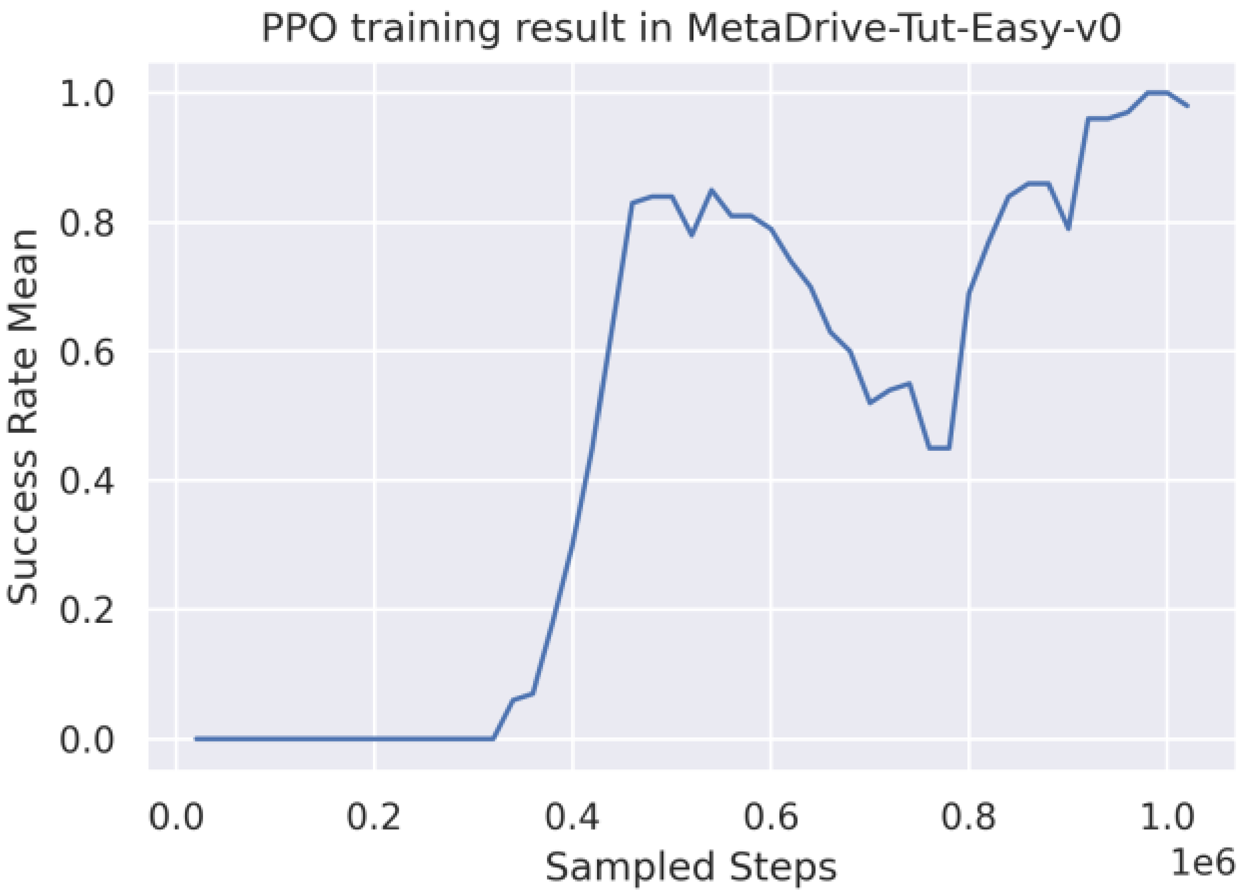
(5 points)



PPO in MetaDrive Easy

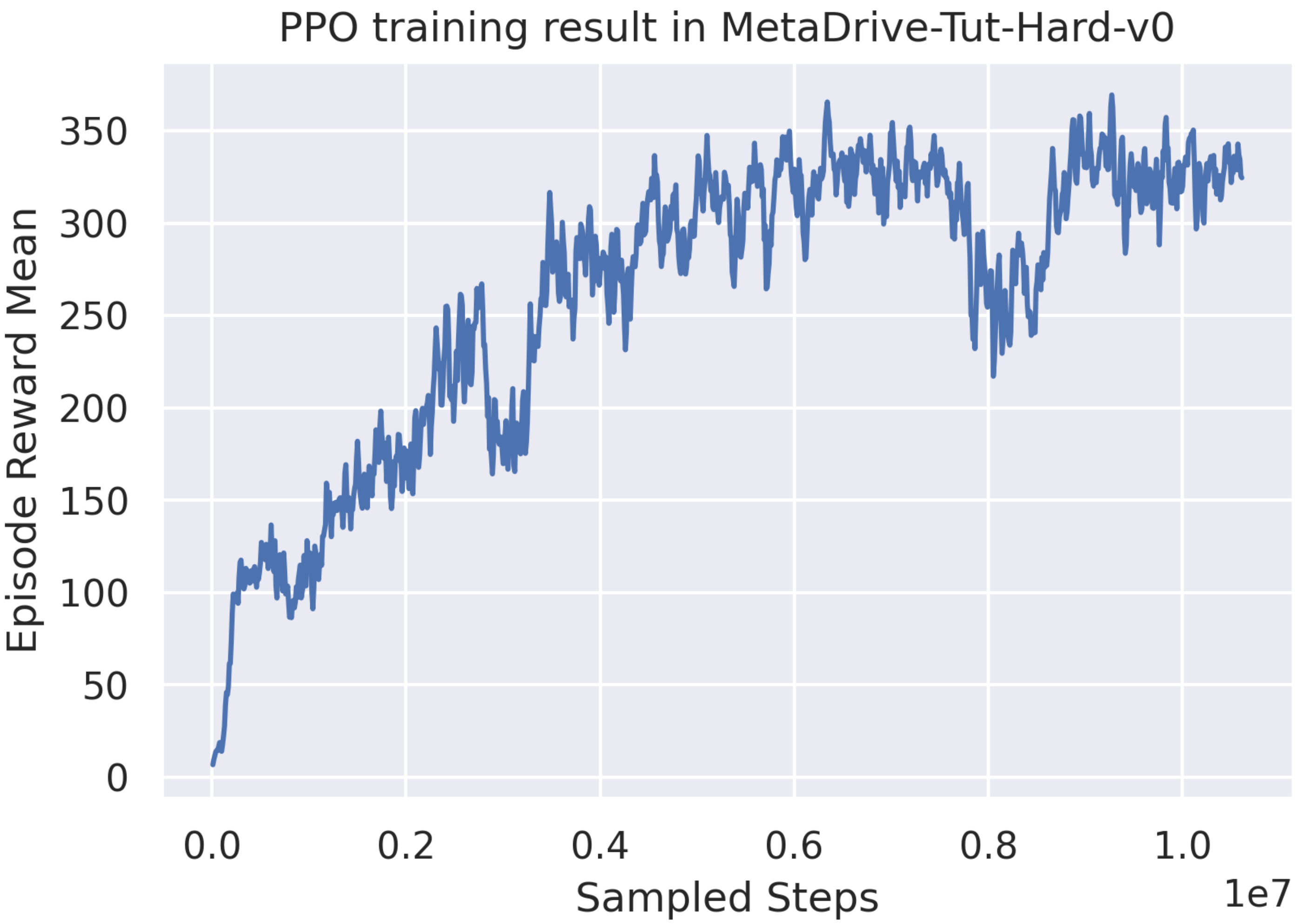
(5 points)

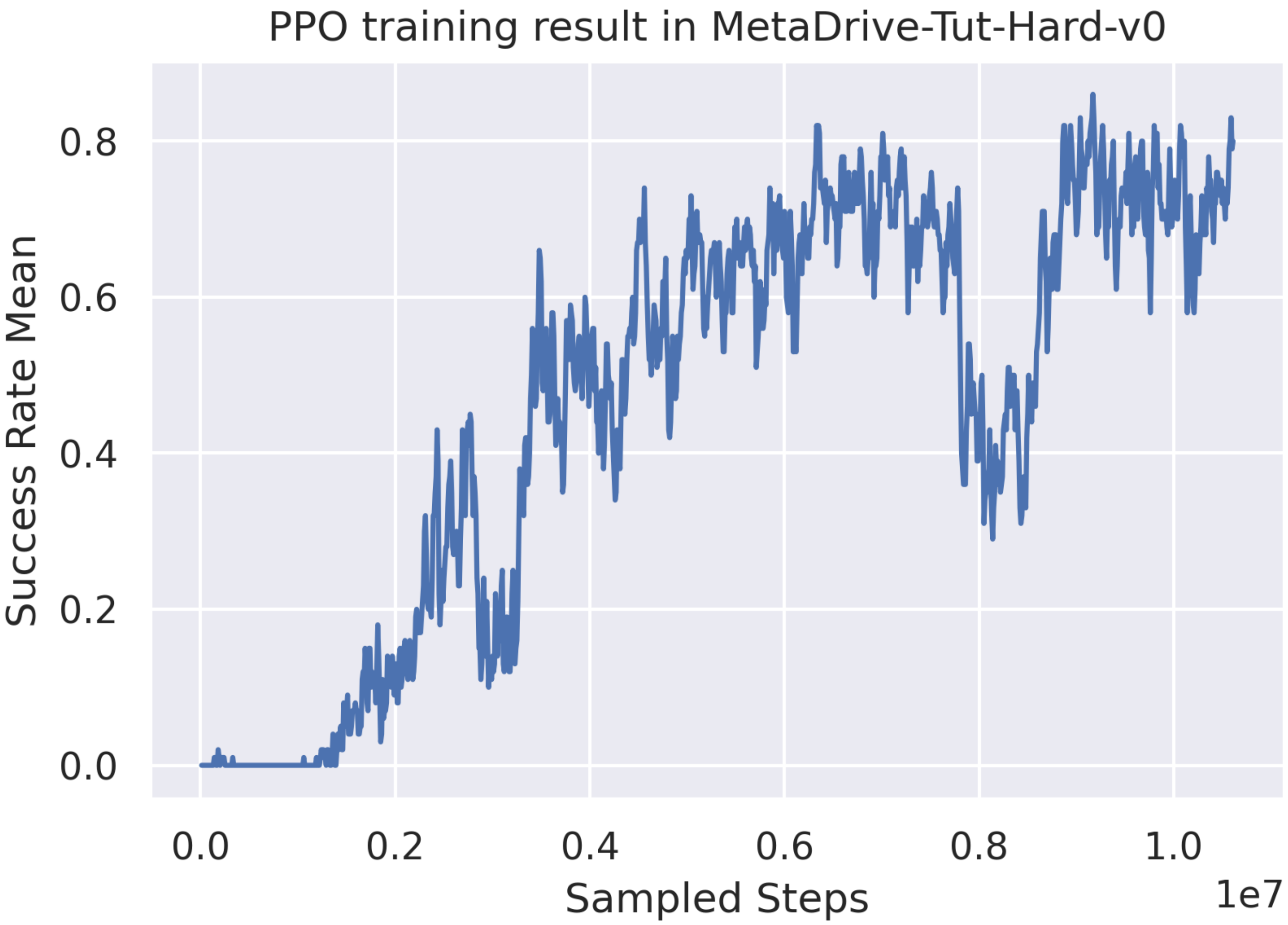

PPO training result in MetaDrive-Tut-Easy-v0
Episode Reward Mean
0
20
40
60
80
100
120
0.0
0.2
0.4
0.6
0.8
1.0
1e6
Sampled Steps


PPO training result in MetaDrive-Tut-Easy-v0
Success Rate Mean
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
1e6
Sampled Steps

PPO in MetaDrive Hard

(10 points)



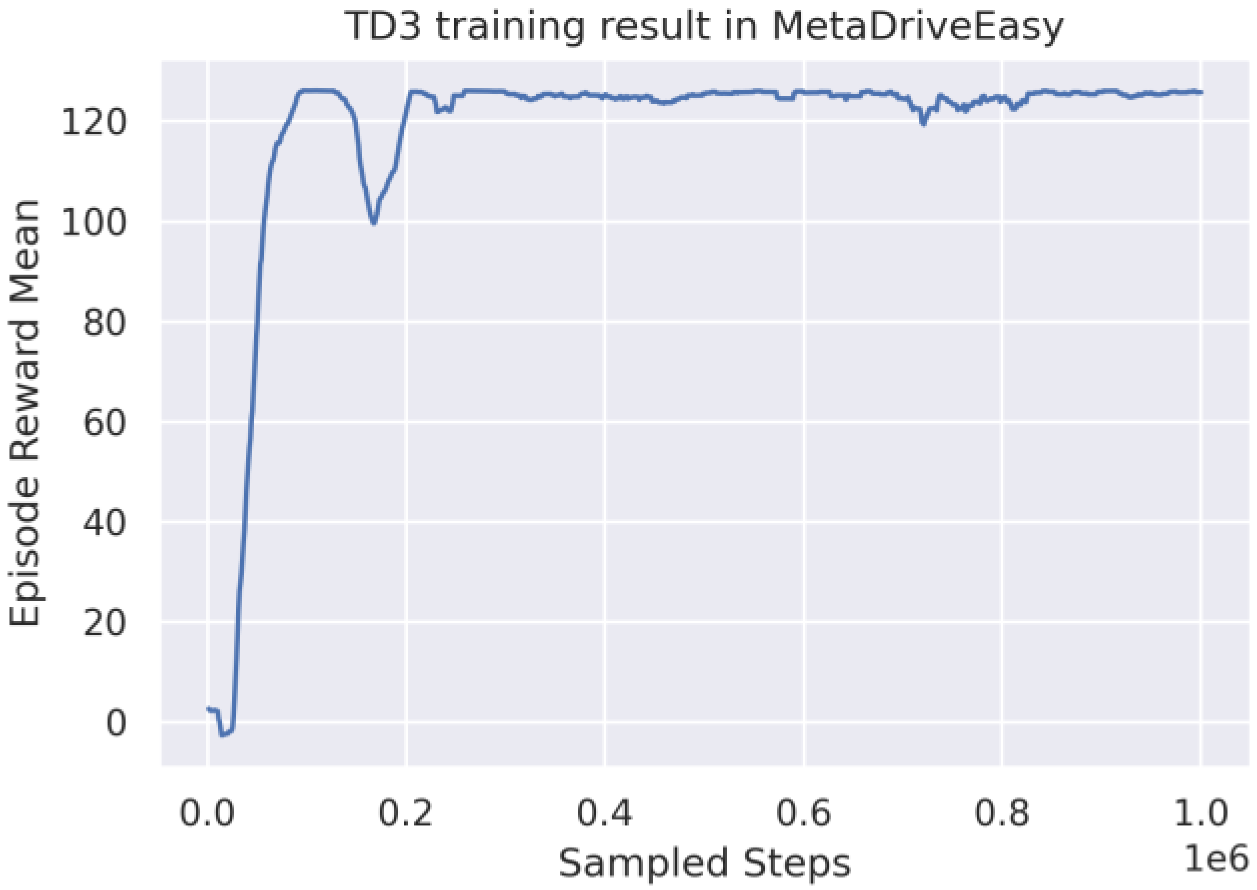


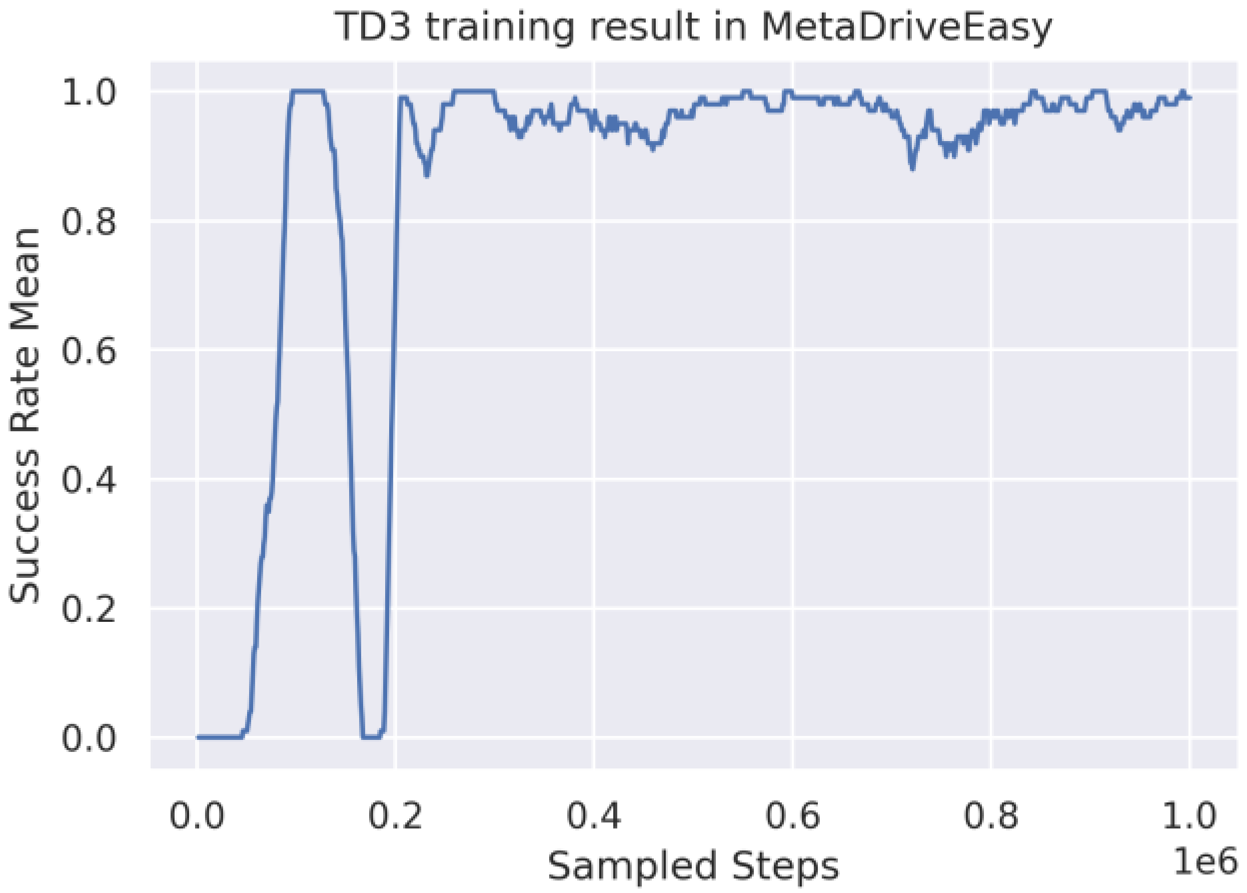
Learning curves of TD3

We require at least 1M steps to train TD3.

TD3 in MetaDrive Easy

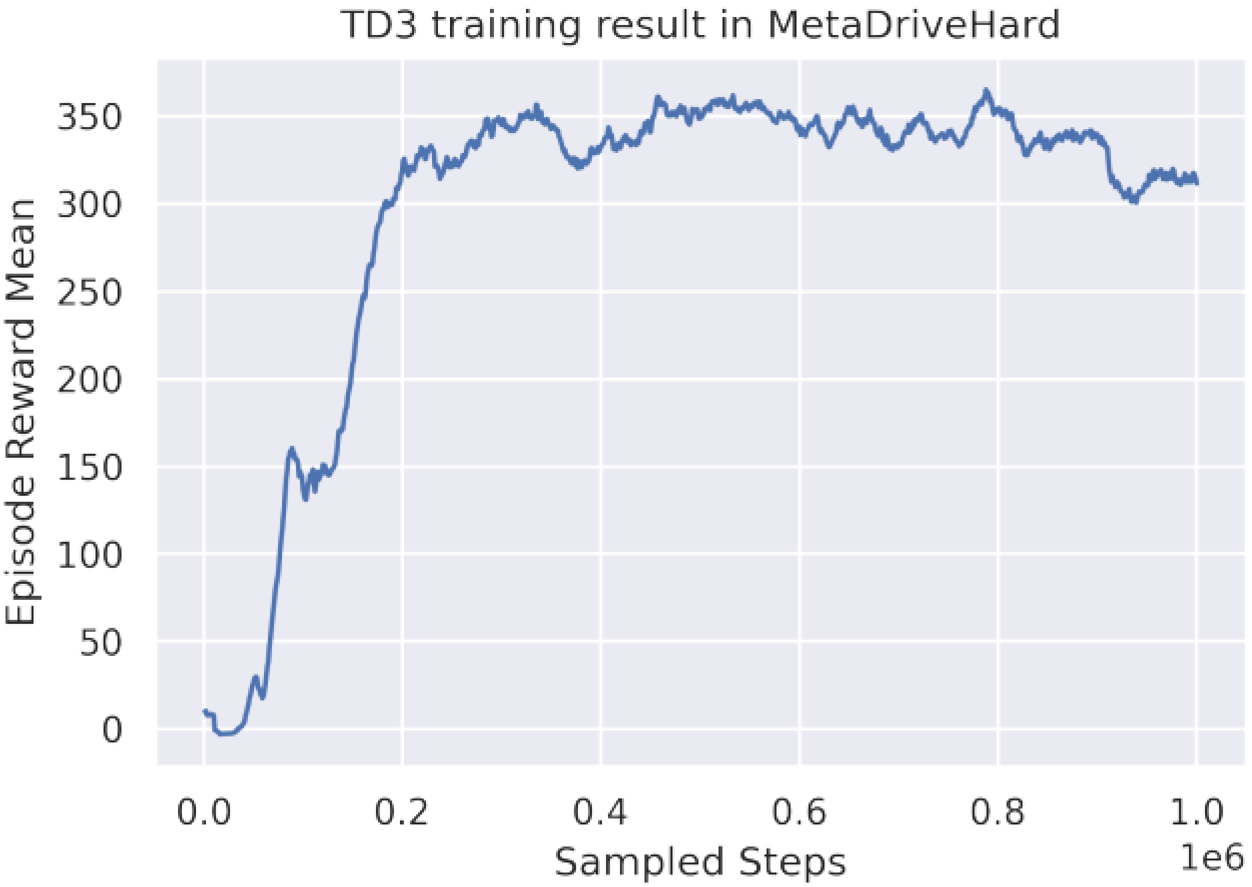
(10 points)

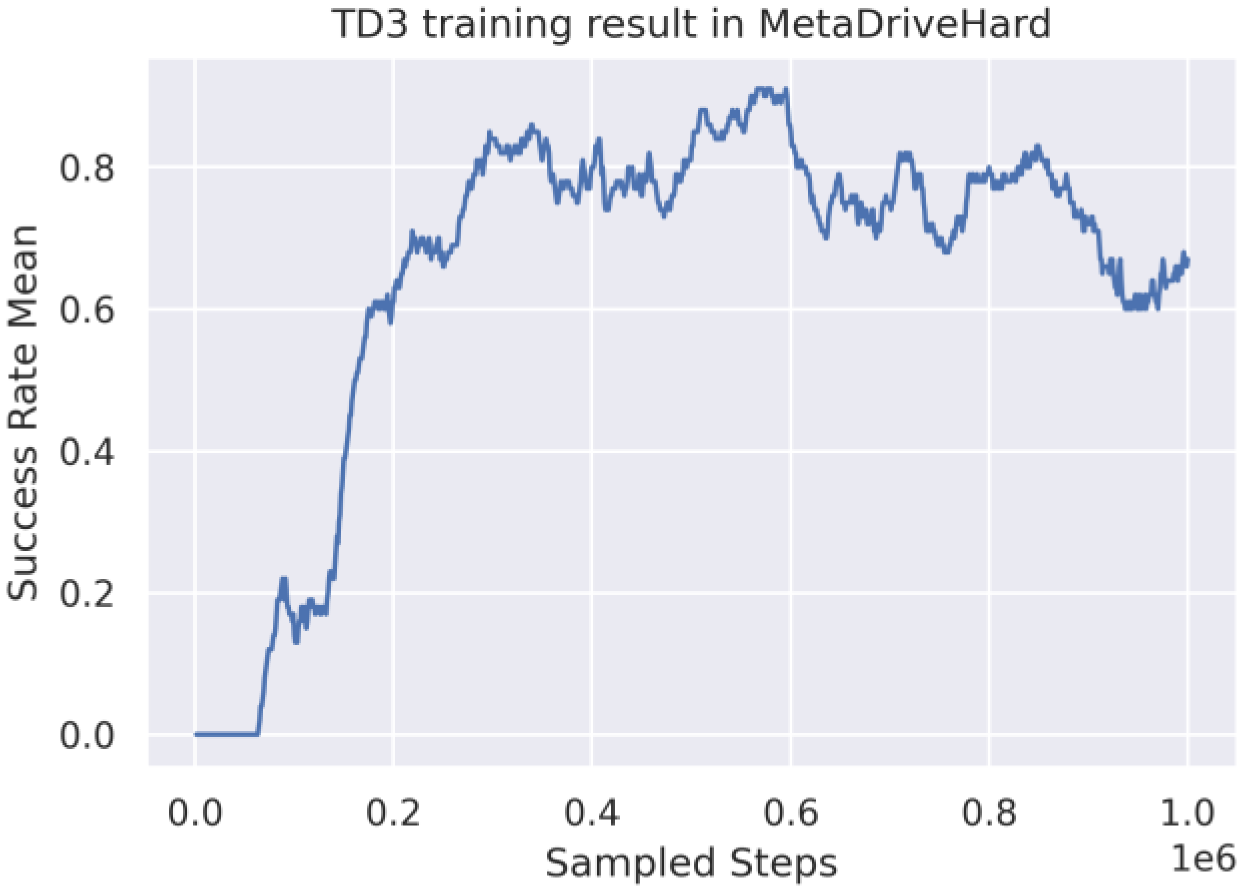




TD3 in MetaDrive Hard

(10 points)





Generalization Curves

In this section, we need two figures whose X-coordinate represents "the number of training scenes" and the Y-coordinate
represents "the episodic reward" or "the success rate" (choose one).
We expect two curves in each figure, showing the final training performance and
the test performance varying with the number of training scenes.

You can refer to the Figure 5 of the paper of MetaDrive paper
to see the expected curves. ProcGen paper is also highly relevant.

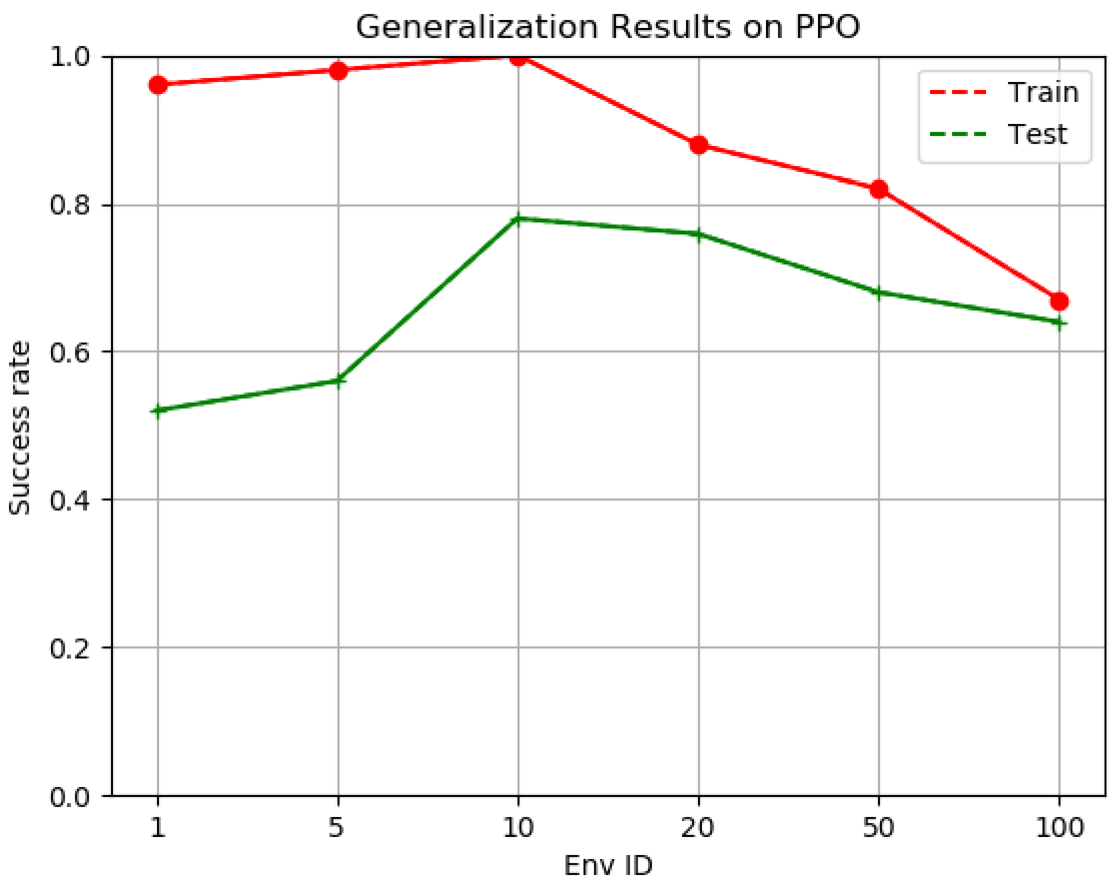
You should train PPO and TD3 agents in `MetaDrive-Tut-[1,5,10,20,50,100]Env-v0` environments and test all agents in `MetaDrive-Tut-Test-v0` . Therefore you should conduct 6 PPO experiments and 6 TD3 experiments. If you indeed can't access enough computing resource, you can omit the training in `MetaDrive-Tut-50Env-v0` and `MetaDrive-Tut-100Env-v0` .
You can refer to the `vis_and_eval.py` script to learn how to write script to evaluate your agents.

In each experiment, we will store many checkpoints. Please find the checkpoint with **best training performance** through observing training performance (since the training performance might not be increase monotonically) and test the checkpoint in the test environment. **You should implement the script to run each checkpoint in the test environment with at least 50 episodes.**

Sicne we don't run adequate repeated experiments, the curves might be chaotic and counterintuitive. Please discuss the reason if in this case.

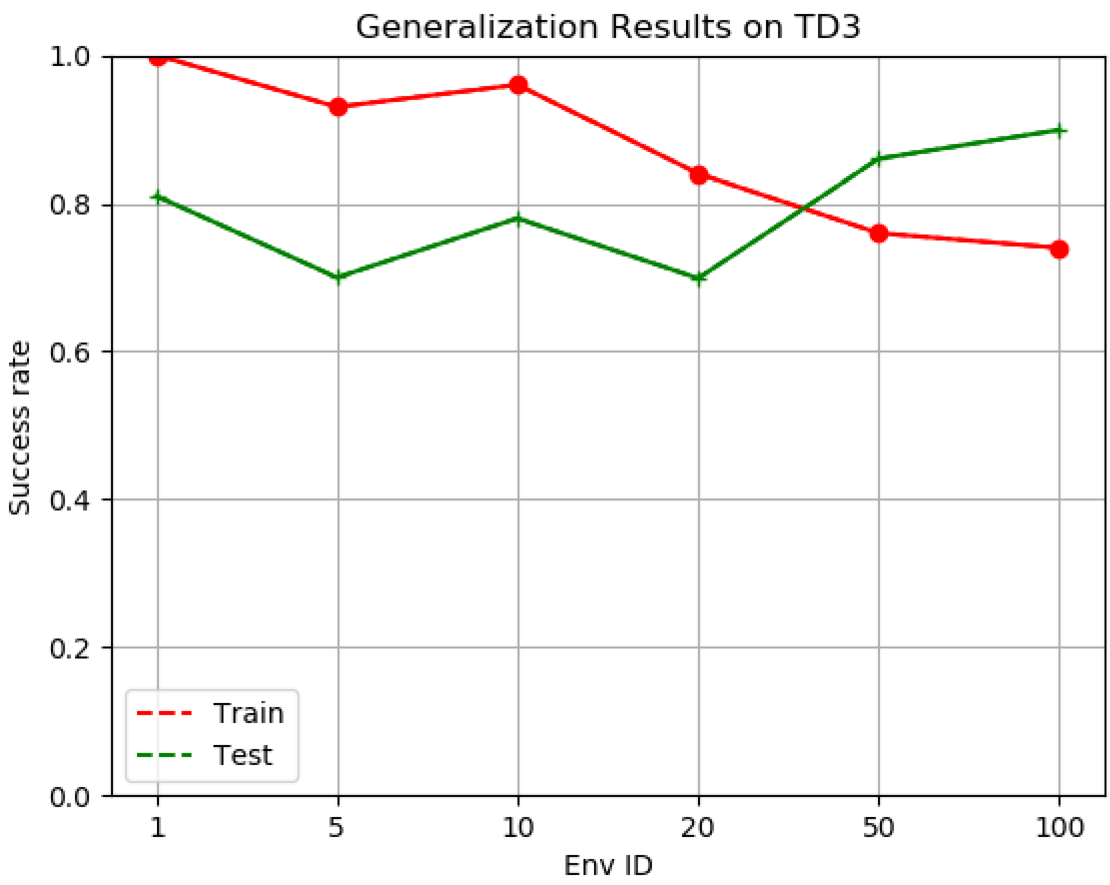
PPO curves

(40 points)



TD3 curves

(20 points)



Discussion

In this section, you are free to write down your ideas and discussion. Reasonable discussion can earn considerable bonus points! You can also leave this section blank, it is optional.

Do you introduce any new technique?

E.g. I introduce the value clipping trick in PPO, here is the ablation result: ...

Do you discover any interesting conclusion?

E.g. I find that the value network should be completely isolated from the policy network, here is the learning curves of this ablation study: ...

E.g. I done a huge batch of hyper-parameter searching experiments, here are what I find: ...

Why do some experiments fail?

Anything else?