Project work

Jari Mattila - 35260T
ELEC-E8125 - Reinforcement Learning

November 25, 2021

Introduction

To cite works, put them in the template.bib file and use [1].

Describe the chosen method

To embed code snippets in the report, you can use the `pycode` environment.

Part I

Answers to the questions in part 1.

Algorithm 1: Twin Delayed Deep Deterministic Policy Gradient (TD3)

Question 1

What are the problems with the deep deterministic policy gradient (DDPG) algorithm? How does TD3 solve these problems?

Question 2

For policy gradient methods seen in Exercise 5, we update the agent using only on-policy data, while in TD3 we can use off-policy data. Why is this the case?

Question 3

Finish the implementation of the TD3 algorithm and train the agent with both InvertedPendulumBulletEnvv0 and HalfCheetahBulletEnv-v0 environments.

Train your agents with three random seeds for both environments.

Include the training plots in the project report and attach the agents' weights to your submission, with filenames ending with td3.pth containing the run ID (1, 2, and 3, each with a different random seed) and the environment name.

If you add a figure, you can refer to it using Figure. 1.

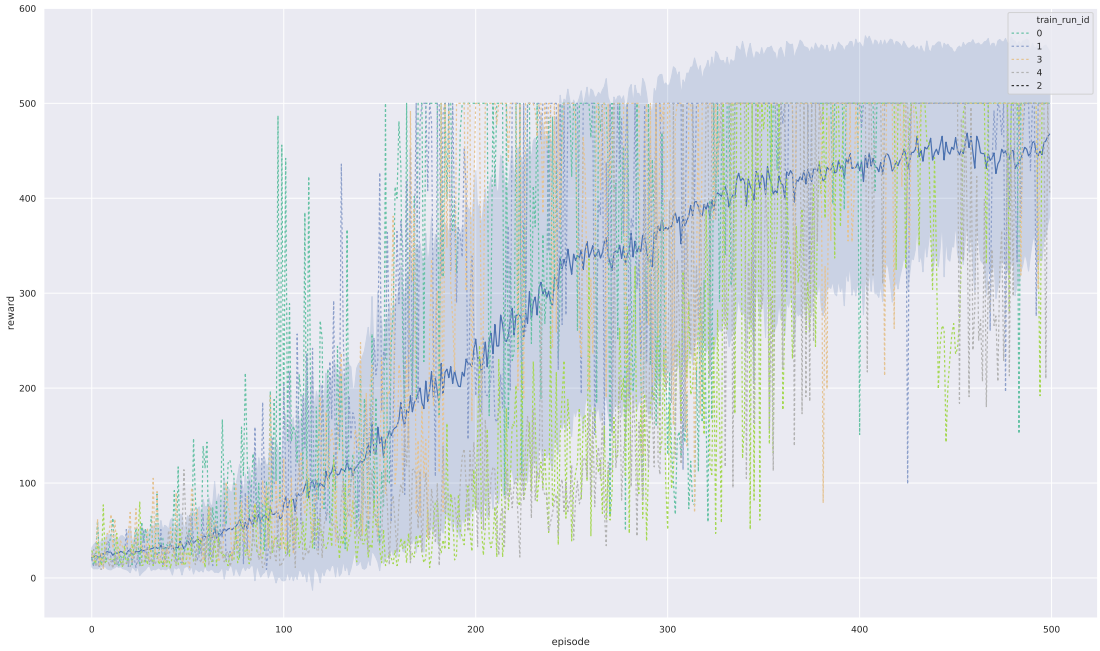

Figure 1: This is a sample figure.

Question 4

Now let's analyze the sensitivity of TD3 to hyperparameters. Choose one hyperparameter, e.g., target action noise, exploration noise, or policy update frequency, that you think heavily influence the training and explain why. Then, train your agent with the modified hyperparameter on both InvertedPendulumBulletEnv-v0 and HalfCheetahBulletEnv-v0 environments (3 random seeds). Show the training plots and submit the trained model of HalfCheetah with name "td3_q4.pth".

Question 5

After playing with the TD3 algorithm, could you find any aspect that could be improved? Please list three of them. Also, please propose a potential solution to one of the problems you listed. You can answer this question by providing a paper link and explaining in your own words how the proposed approach solves/mitigates the problem.

Algorithm 2: Proximal Policy Optimization Algorithms (PPO)

Question 1

Why does clipping the $\frac{\pi_\theta(a|s)}{\pi_{old}(a|s)}$ ratio stabilize the training? What is the relationship between TRPO [7] and PPO?

Question 2

Please finish the implementation of the PPO algorithm.

Similar to Question 3, train the agent on both the InvertedPendulumBulletEnv-v0 and the HalfCheetahBulletEnv-v0 environments.

Train your agents with three random seeds for both environments.

Include the training plots in the project report and attach the agents' weights to your submission, with filenames ending with ppo.pth containing the run ID (1, 2, and 3, each with a different random seed) and the environment name.

For the training curve, you can reference 2.

Question 3

In PPO, the target value is calculated by generalized advantage estimation (GAE) [8], as shown in the second equation. Explain the relationship between n-step advantage and GAE. Why is GAE better than n-step advantage?

Part 2

Answers to the questions in part 2.

Question 1

Please correctly implement your algorithm and show the training plots against the TD3/PPO with three random seeds.

The plots should include both InvertedPendulum and HalfCheetah environments.

Also, you need to describe the network structure and training procedure as well as hyperparemeters. A clear way to show the hyperparameters is using a table.

Question 2

Let's analysis your algorithm by performing an ablation study. You could modify one design option that you expect to influence the training performance.

Then train the agent using the modified code and compare the results to the original algorithm. Training your agent on HalfCheetah environment is enough but with three random seeds.

Conclusions

oekdwpodkwpod

References

[1] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.