Exercise 1

Marco Di Francesco - 100632815
ELEC-E8125 - Reinforcement Learning

November 8, 2022

Task 1

Look at plot 1.

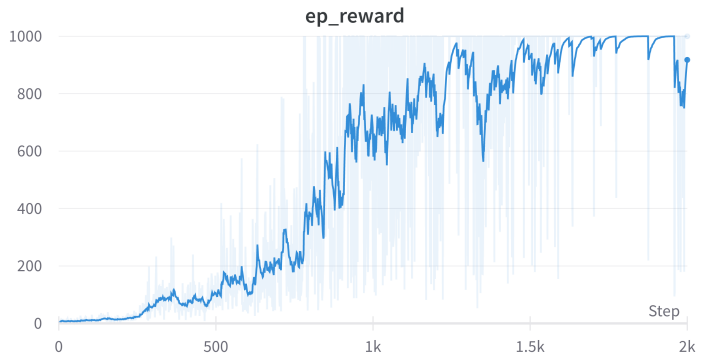

Figure 1: Training performance plot task 1

Question 1.1

The relationship between reinforce with baseline and actor-critic is that they both a policy and state-value function, but we can't consier reinforce with baseline as an actor critic because the state value function is used as a baseline and not as a critic.

Actor critic methods use bootstrapping, thus updating the value estimate based on subsequent states, while reinforce with baseline uses only a baseline for the state where the estimate is being updated.

Question 1.2

The value of advantage can be interpreted as the improvement an action has compared to the others at a given state, this means that the evaluation of an action is based not only on how good the action is, but also how much better it can be.

Question 1.3

In reinforce we do not introduce bias, thus making it overfitting to a local minima. This is because you have to wait to the end of each episode before applying updates.

In actor critic we introduce bias making it not overfit. This method introduces initial bias towards however the model was initialised. Moreover it generally converges faster than reinforce and reduces variance.

Question 1.4

One solution may be the use of n-step returns. In this approach we have an N-step bootstrapping that observes a longer chain of rewards before entering the estimated value state.

Task 2

Look at plot 2.

Figure 2: Training performance plot task 2

Question 2.1

It's not possible off-policy data for policy gradient methods because we are using cumulative reward in order to compute the gradient contribution by each trajectory. We can fix this by using importance sampling, and this is used in order to correct for the slight drift from being exactly on policy.

Question 2.2

One disadvantage of deterministic models is that we learn from a deterministic action instead of drawing it from a distribution. Drawing from a distribution allows the agent to explore.

Another disadvantage is that it may be computationally inefficient. In the publication by Heess et al. [1] they demonstrated a significant performance advantage to using deterministic policy gradients over stochastic policy gradients, especially for high dimensional tasks.

References

[1] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, "Deterministic policy gradient algorithms," p. 9.