CSE 599B: Homework 2 Submission

Alexey Yermakov

May 6, 2024

1 Policy Gradient

- The mean rewards during training with default parameters for Policy Gradient is shown below as well as Policy Gradient without normalization:

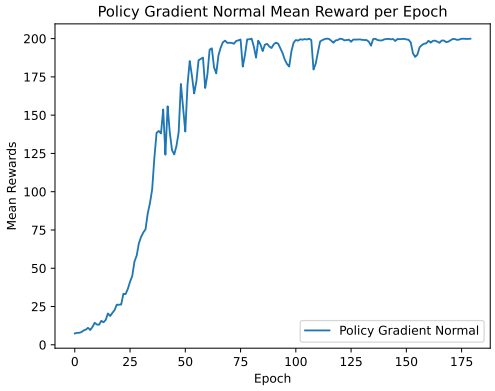

Figure 1: Policy Gradient trained with normal hyperparameters.

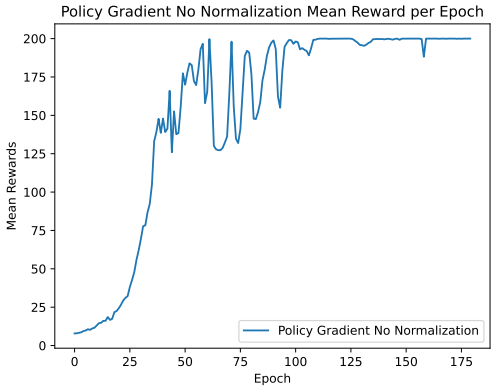

Figure 2: Policy Gradient trained with no normalization.

- The success rate and average reward for Policy Gradient with normal parameters is:

```
Success rate:  0.97
Average reward (success only):  200.0
Average reward (all):  199.58
```

- The success rate and average reward for Policy Gradient with no normalization is:

```
Success rate:  1.0
Average reward (success only):  200.0
Average reward (all):  200.0
```

- For policy gradient for the inverted pendulum problem, it appears that the average reward is unaffected when removing the reward normalization.

2 Actor Critic

- The mean rewards during training with default parameters for Actor Critic is shown below as well as Actor Critic without a separate target Q function:

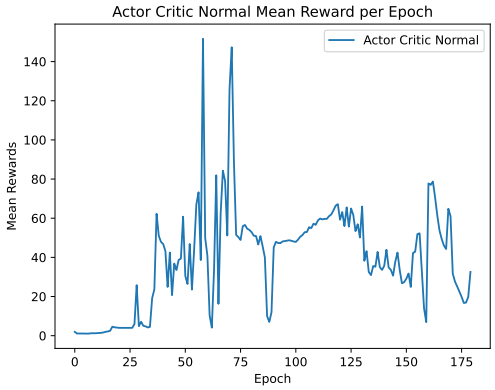

Figure 3: Actor Critic trained with normal hyperparameters.

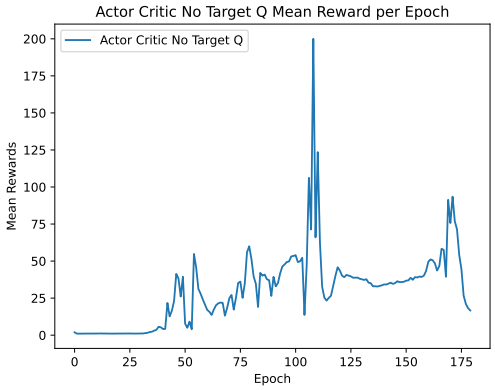

Figure 4: Actor Critic trained without a separate target Q function.

- The success rate and average reward for Actor Critic with normal parameters is:

```
Success rate:  1.0
Average reward (success only):  200.0
Average reward (all):  200.0
```

- The success rate and average reward for Actor Critic without a separate target Q function is:

```
Success rate:  0.0
Average reward (success only):  0.0
Average reward (all):  16.09
```

- For Actor Critic for the inverted pendulum problem, it appears that the average reward is completely stunted when not using a separate target Q function. This isn't shown in the average rewards plot, but is shown in the testing average reward, where the Actor Critic policy with no separate target Q function doesn't learn anything.

3 Discussion

- The role of the value function in actor-critic methods is that it enables actor-critic methods to learn off-policy and with lower variance. The off-policy learning comes from using two Q functions: one whose parameters we're updating, and the other whose parameters are fixed (and are usually set to the previous parameter values of the Q function being updated). The lower variance is a result of the gradient updates use the entire set of all previous roll-outs. Thus, actor critic methods are much more sample efficient and stable due to being able to use old information.

- Policy gradient, however, has to rollout new paths every time a gradient step is taken. It cannot utilize previous rollouts like actor-critic since the gradient step requires information from the current policy. Thus, policy gradient is on-policy and much less sample efficient. This also makes it significantly less stable, since it has a significantly smaller amount of data it can use due to not being able to use old information.

- From the above plots, its clear that the mean reward per epoch is highly variable. This is likely due to the learned Q function being only an approximation to the sum of rewards, whereas policy gradient doesn't have this approximation and instead uses monte-carlo to obtain this value.

- Lastly, from the above plots it appears that policy gradient seems to converge pretty well given enough epochs.