Home assignment 4

Question 1

Based on back propagation formular, the returns affect the gradient. The normalization on returns will help backpropagation **NOT to yield extreme values on network weights.**

Question 2

3 executions were used to compare the scenario. In case of **decay variance**, it took **~ 500, 600, 500 episodes**. Meanwhile, in the case of **fixed variance**, it took **~ 6000, 11000, 7000** episodes.

Decay variance were better. Fixed variance took longer to train.

The network output get better over time, which means to train efficiently, we should trust the output of the network more over time. Decaying variance is about just that, because as variance decays, the sample of the distribution is closer to the mean of the distribution (which is the network output). Fixed variance, on the others, keeps the algorithm explorative eternally.

Question 3

3 executions were used to compare the scenario (policy gradient result from task 1 is reused).
For **policy gradient** (with decaying variance), it took **~ 500, 600, 500 episodes**
For **actor-critic** scenario, it took **~ 350, 450, 450 episodes**

Conclusion: actor-critic learns faster

Benefits of actor-critic algorithm: the actor-critic algorithm general converges faster because each **sample is evaluated more efficiently** using (along with sampled return) the critic's score on the actor.

Question 5

Similar to the answer in question 3, actor-critic algorithm with extra critic factor improves the quality of gradient of sample, while policy gradient can often get stuck at a local minima.

(continue on next page)

Question 6

Policy gradient's advantages over Q-learning:

- It can handle well problem with large discrete or continuous state space, usually with an approximator. Q-learning with value based methods can be computationally inefficient in this case

- It can handles well problem with a stochastic policy as the optimal policy

I would use Q-learning with value-based methods for problem domain with small size, and also requires good learning speed. Policy gradients for the other cases.