# DDA4230 Homework 4

Name: Xiang Fei
Student ID: 120090414

# Problem 1

- **Question 1**

REINFORCE directly learns the policy by estimating the gradient of the expected return with respect to the policy parameters. It operates in an episodic manner, where it collects trajectories by following the policy and then updates the policy based on the complete returns obtained. REINFORCE tends to have higher variance due to the use of full trajectories for policy updates. Besides, It might require more samples to converge due to the high variance, making it less sample-efficient. In addition, REINFORCE does not estimate the value function separately. It directly learns the policy.

Actor-Critic method combine both value-based and policy-based approaches. In this case, there are two components - the actor (policy) and the critic (value function). The actor decides the action based on the policy, and the critic evaluates the state or state-action pairs. Actor-Critic methods typically have lower variance compared to REINFORCE due to the inclusion of a value function. The critic (value function) provides an estimation of the advantage, which helps to reduce variance in policy updates. Besides, Combining value estimation with policy learning tends to lead to more stable and faster convergence. Also, Actor-Critic is often more sample-efficient compared to REINFORCE due to lower variance.

- **Question 2**

Parameters for REINFORCE:

1. EPISODES = 2000: Given the nature of REINFORCE, which relies on complete episodes, a higher number of episodes allows for better policy learning.
2. STEPS = 500: This parameter might be relevant for environments where episodes have a fixed length. For Pendulum, this step count might affect the granularity of learning. Also, enough steps are necessary for done a experiment.
3. GAMMA = 0.98: This value of 0.98 indicates a higher emphasis on future rewards without making them too distant, which could be suitable for the Pendulum environment.
4. learning_rate = 0.001: It's a common starting point that often works well for simple problems.

5. hidden_size = 32: This parameter might affect the network's capacity to learn complex policies. Given the simplicity of the Pendulum environment, a smaller size like 32 neurons might be sufficient.

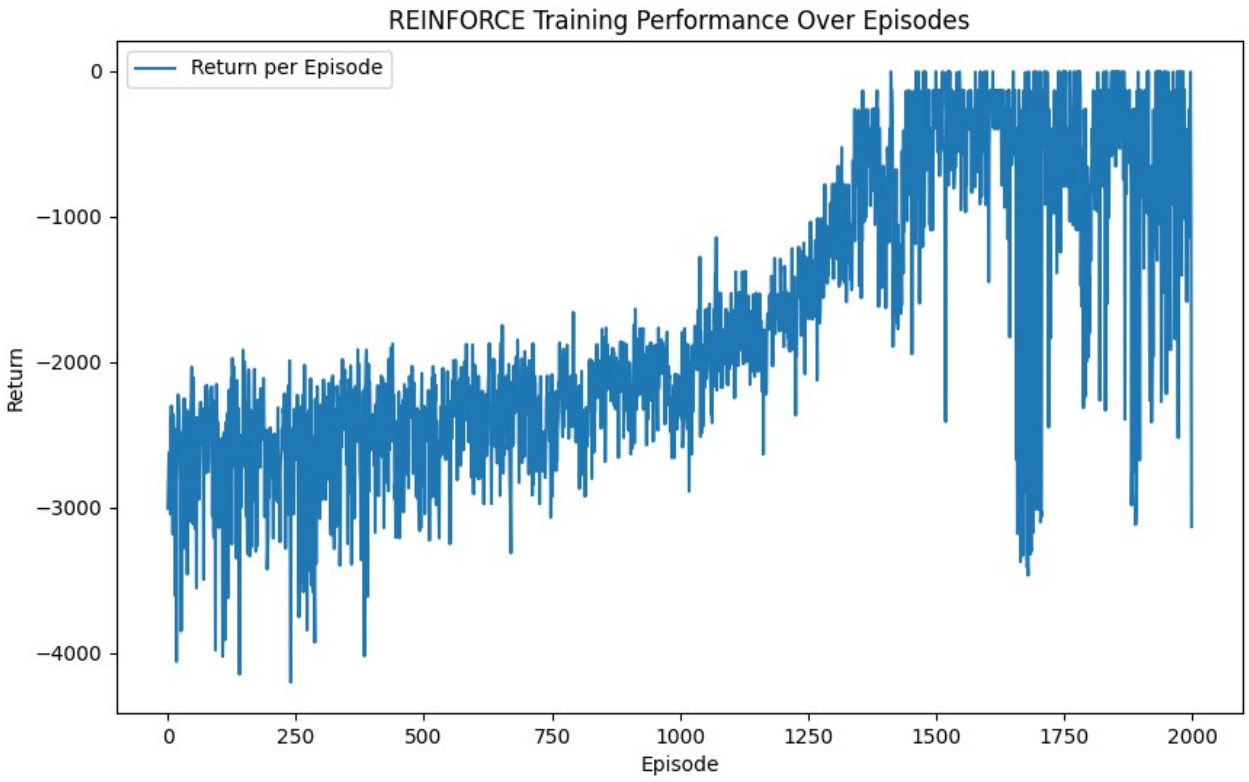
Parameters for Actor-Critic:

1. The parameters for EPISODES, STEPS, and GAMMA remain the same, as these hyperparameters are more related to the nature of the problem and learning dynamics rather than the specific algorithm.
2. learning_rate = 0.001: This learning rate is a reasonable starting point. Actor-Critic methods may benefit from moderate learning rates as they combine both value-based and policy-based updates.
3. hidden_size = 32: This size may be adequate for capturing the environment's dynamics and policy in the Actor network while estimating value functions in the Critic network.

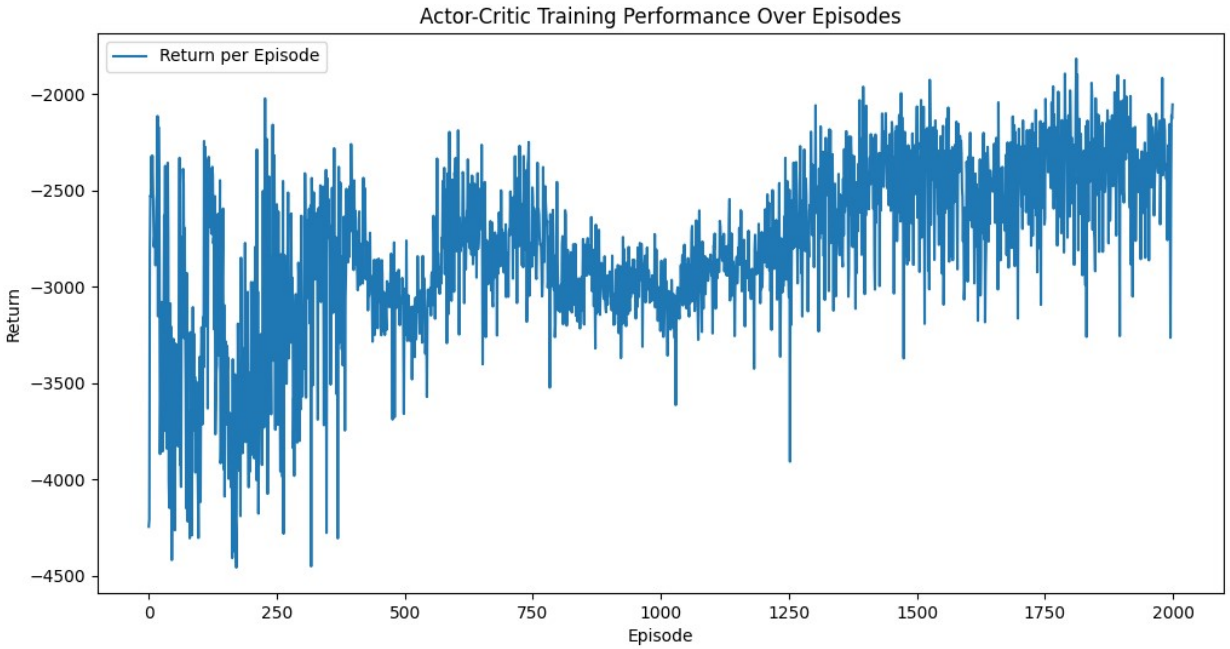
Experimentally, for both REINFORCE and Actor-Critic, the parameters I chose are reasonable.

- **Question 3**

Training graph for REINFORCE:




Training graph for Actor-Critic:




- **Question 4**

It seems that for my training process, REINFORCE is better since the return improvement is larger under the limited episode number. However, REINFORCE is less stable compared to Actor-Critic method. Also, the influence of parameters is definitely huge.

Typically, It's relatively straightforward to implement and understand REINFORCE. Besides, no value function estimation can sometimes be an advantage in certain scenarios where value estimation might be challenging or unnecessary. However, REINFORCE tends to have higher variance due to the use of full trajectories, making it less sample-efficient.

In conclusion, from my perspective, I think:

1. For simpler problems or environments where the state-action space is relatively straightforward and the variance isn't a significant concern, REINFORCE might perform reasonably well and could be easier to implement.
2. However, in more complex environments with high-dimensional state spaces, Actor-Critic methods tend to perform better due to their lower variance, faster convergence, and the advantage of having both policy and value function estimations.