Assignment 2

Guy Hadad - 316508126

Tomer Shaked -

Elad Sofer –

Section 1

- The advantage estimate term reflects the advantage of taking a particular action  *at*​ in state compared to the baseline expectation.

It is better to follow the gradient computed with the advantage estimate instead of just the return itself because it reduces the variance (without adding bias) of the gradient estimates. This method in more stable and efficient learning.

- The prerequisite condition for this equation to be true is that the baseline is independent of the action chosen but only on the state.

The proof:

We used the identity: =

Section 2:

It is the same because of the relationship of these two terms:

As you can see, the TD-error includes the difference between the value function and the advantage estimate. Therefore, updating the policy network parameters using either the TD-error or the advantage estimate would lead to similar adjustments in the policy.

In the Actor-Critic model, the actor is the policy network, which decides which actions to take based on the current state. Its role is to select actions that maximize expected future rewards. The critic, on the other hand, is the value network, which evaluates the state or state-action pairs and provides feedback to the actor about the goodness of its actions. Its role is to estimate the value of being in a certain state or taking a certain action in a given state, guiding the actor towards better actions.