Assignment 2

Guy Hadad -

Tomer Shaked -

Elad Sofer –

Section 1

- The advantage estimate term reflects the advantage of taking a particular action a_t *at* in state s_t 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.

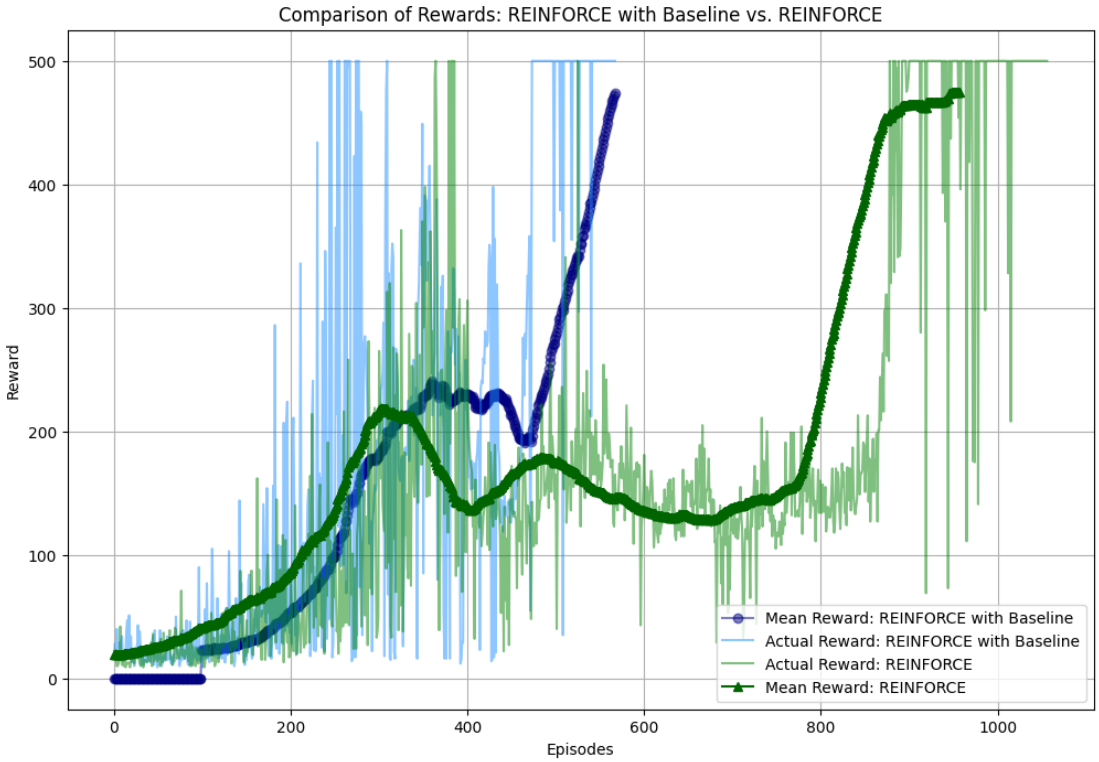
The proof:

$$\mathrm{E}_{\pi_\theta}[\nabla log\pi_\theta(a_t|s_t)b(s_t)] = \sum_{a_t} \pi_\theta(a_t \mid s_t)\nabla\log\pi_\theta(a_t \mid s_t)b(s_t)$$

$$= \sum_{a_t} \nabla\pi(a_t \mid s_t)b(s_t)$$

$$= b(s_t)\sum_{a_t} \nabla\pi(a_t \mid s_t) = b(s_t)\nabla\sum_{a_t} \pi(a_t \mid s_t) = b(s_t)\nabla 1 = 0$$

We used the identity: $\pi_\theta(a_t \mid s_t)\nabla\log\pi_\theta(a_t \mid s_t) = \nabla\pi_\theta(a_t \mid s_t)$



The convergence time for REINFORCE was observed after 1,056 episodes, whereas for REINFORCE with baseline, it occurred after 568 episodes.

We implemented the algorithm as shown in the class, employing adaptive and different learning rates for the two networks: 0.0007 for the value network and 0.0005 for the policy network. The learning rates are reduced by a factor of 0.6 whenever the average reward over the last four episodes exceeds 450, with a lower bound set at 7e-10 for the learning rates to prevent them from becoming too small. Updates to the value network are withheld if the last 10 episodes yield perfect results, and similarly, updates to the policy network are paused if perfection is achieved over 5 episodes. Various methods that proved ineffective include gradient clipping, dropout, layer normalization, batch normalization, L2 regularization, exploring different network architectures, and testing various activation functions. We utilized a discount factor (gamma) of 0.99.

Section 2:

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

$$\hat{\delta}_t = r_{t+1} + \gamma \hat{V}_v(s_{t+1}) - \hat{V}_v(s_t)$$

$$\hat{A}_t = \hat{Q}_w(s_t, a_t) - \hat{V}_v(s_t)$$

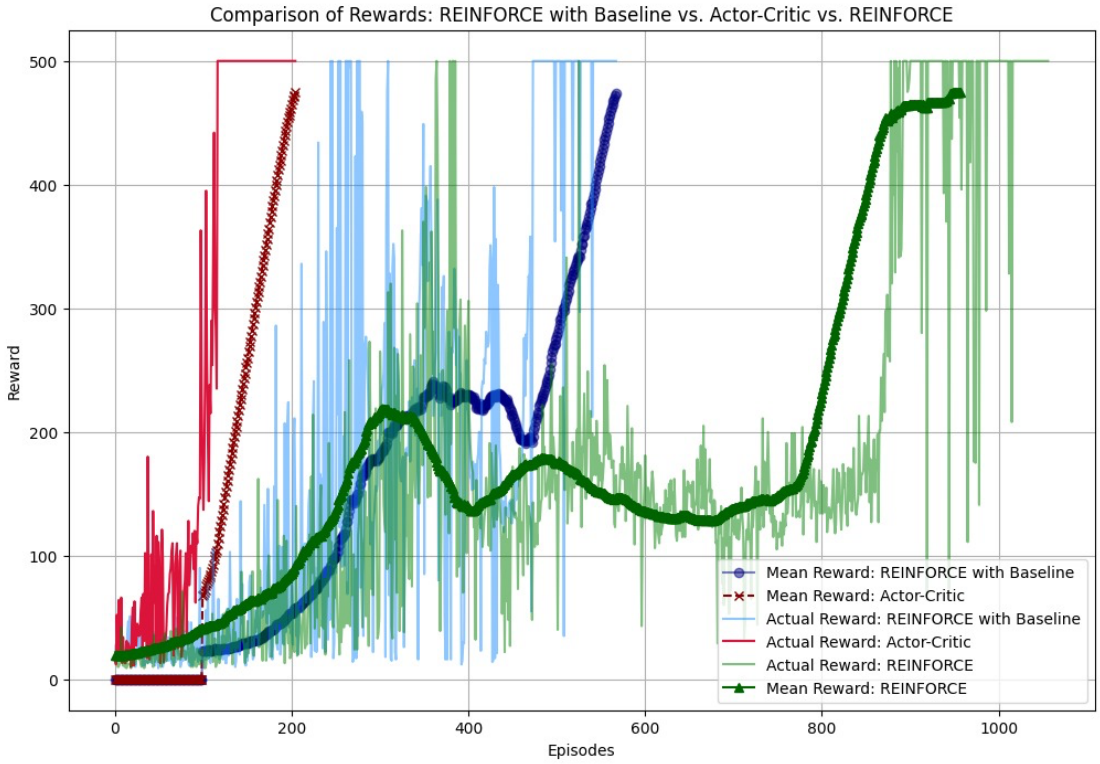
$$\hat{Q}_w(s_t, a_t) = r_{t+1} + \gamma \hat{V}_v(s_{t+1})$$

$$\hat{A}_t = r_{t+1} + \gamma \hat{V}_v(s_{t+1}) - \hat{V}_v(s_t)$$

$$\hat{\delta}_t = \hat{A}_t$$

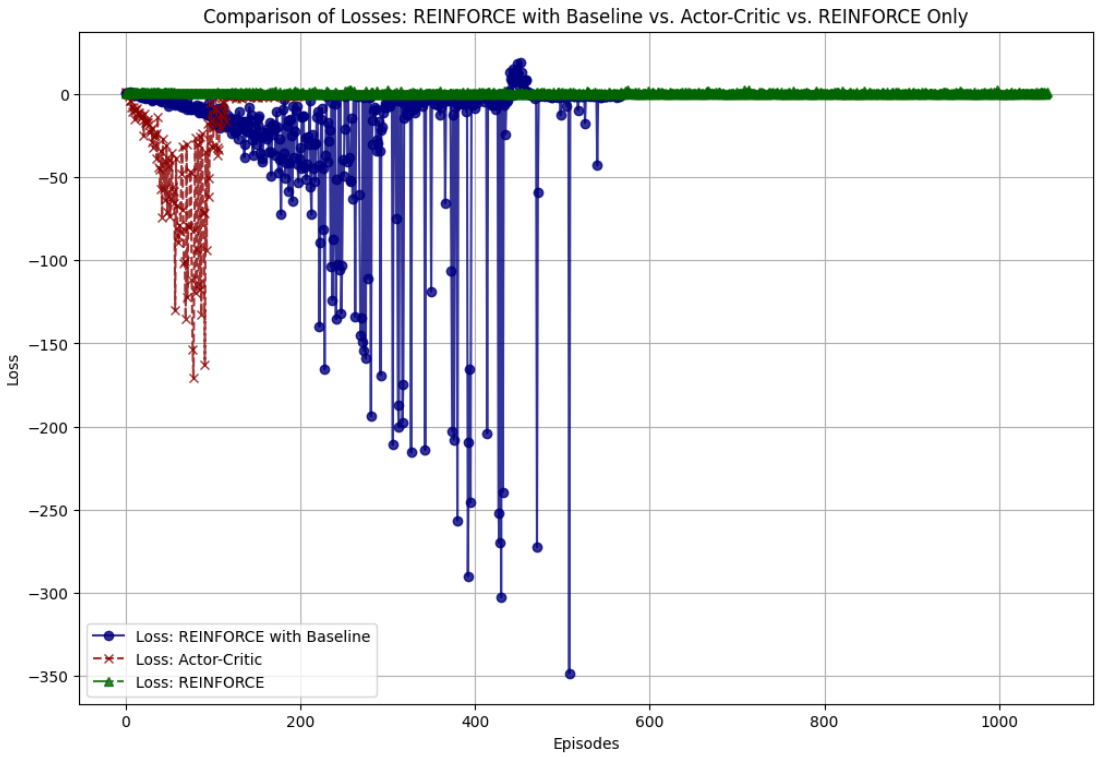
As you can see, the advantages is simplifies to the definition of the TD error. Therefore, updating the policy network parameters using either the TD-error or the advantage estimate would lead to similar adjustments in the policy.

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.



The convergence times observed were 1,056 episodes for REINFORCE, 568 episodes for REINFORCE with baseline, and just 202 episodes for the actor-critic method.

Following the approach outlined in the class, we applied adaptive and distinct learning rates for each network: 0.0007 for the value network and 0.01 for the policy network. Updates to the policy network are suspended when the reward levels demonstrate sufficient stability. We adopted a discount factor (gamma) of 0.99 and employed a different network architecture compared to the previous algorithm.



The loss observed for REINFORCE remained stable and low, whereas with actor-critic and REINFORCE with baseline, we encountered various issues, such as exploding gradients. However, ultimately, the loss converged to zero. It's worth noting that instances of negative loss were due to the subtraction applied to the loss term.