Assignment 2

Eyal Mazuz 208373977
Ariel Amsel 302269907

Section 1-

Question 1-

The value of the advantage function estimates the difference between the current policy and reward to our knowledge of the current state. It's better to follow the advantage estimate since the subtraction of the value function reduces variance in the gradients because the reward can fluctuate greatly.

Question 2-

The prerequisite condition for the baseline to not introduce bias to the expectation of the gradient is the estimator itself is unbiased (shouldn't depend on the action)

$$E_{\pi_\theta}[\nabla log\pi_\theta(a_t|s_t)b(s_t)] = E_{s_{0:t},a_{0:t-1}}[E_{s_{t+1:T},a_{t:T-1}}[\nabla log\pi_\theta(a_t|s_t)b(s_t)]] =$$

$$= E_{s_{0:t},a_{0:t-1}}[b(s_t) * E_{s_{t+1:T},a_{t:T-1}}[\nabla log\pi_\theta(a_t|s_t)]] = E_{s_{0:t},a_{0:t-1}}[b(s_t) * E_{a_t}[\nabla log\pi_\theta(a_t|s_t)]]$$

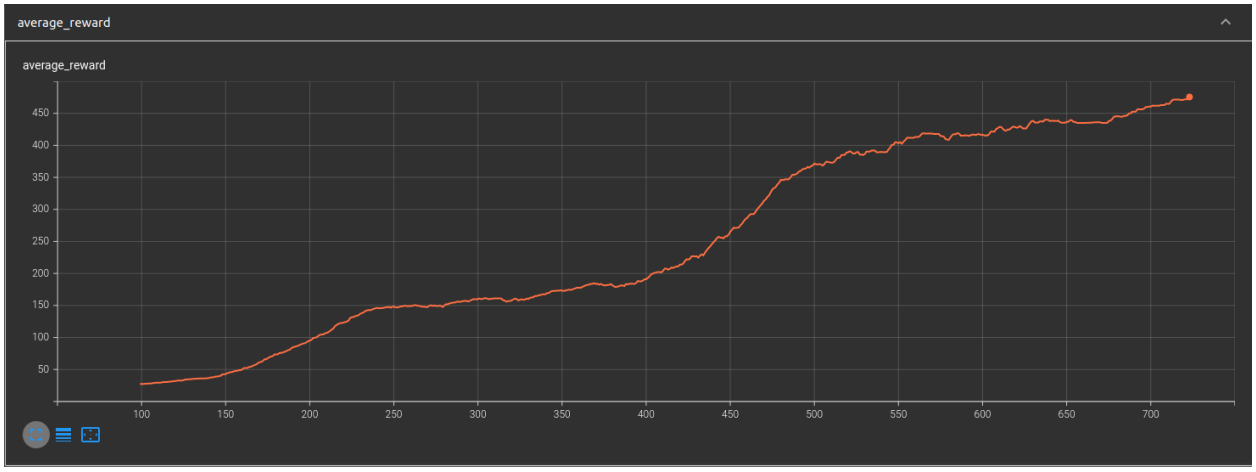
$$= E_{s_{0:t},a_{0:t-1}}[b(s_t) * 0] = 0$$

Code Section-

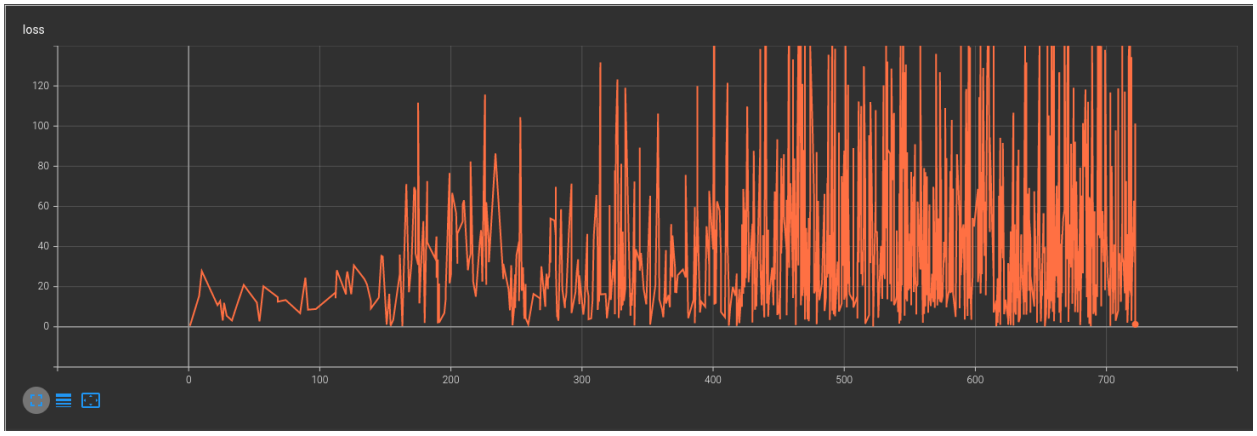
Reinforce:

The regular reinforce algorithm took 723 steps to converge

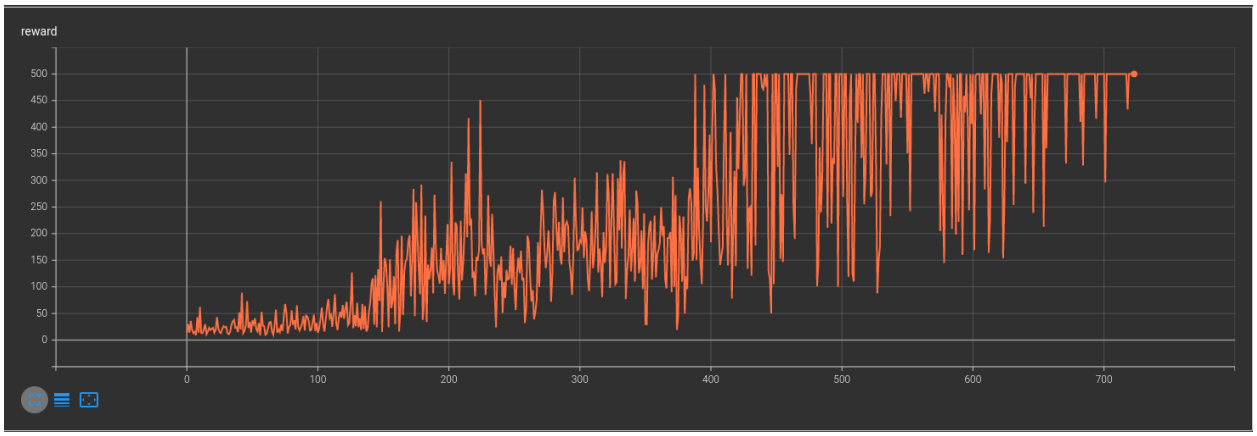
Average Reward



Policy Loss



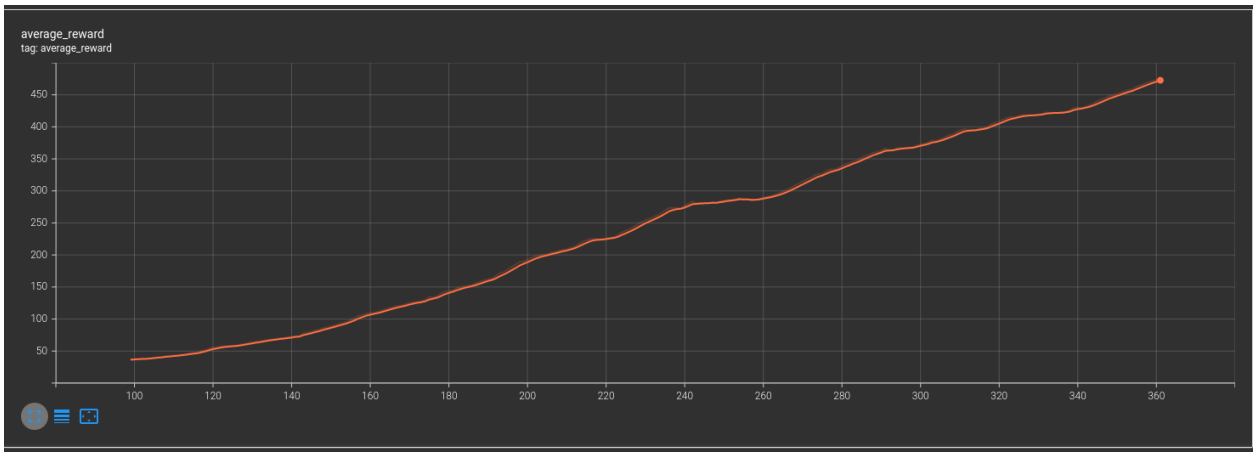
Reward per step



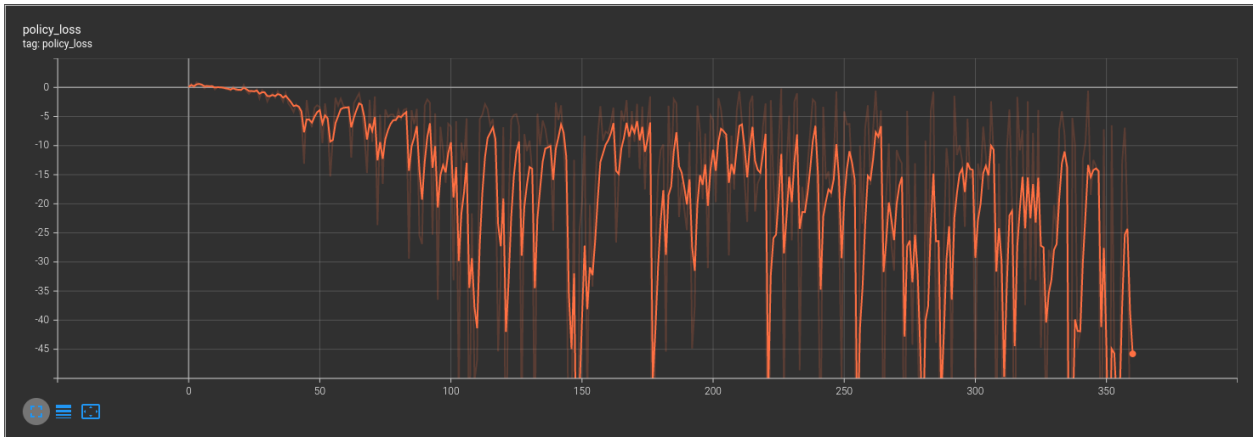
Reinforce with baseline:

The reinforce algorithm with baseline took 361 steps to converge

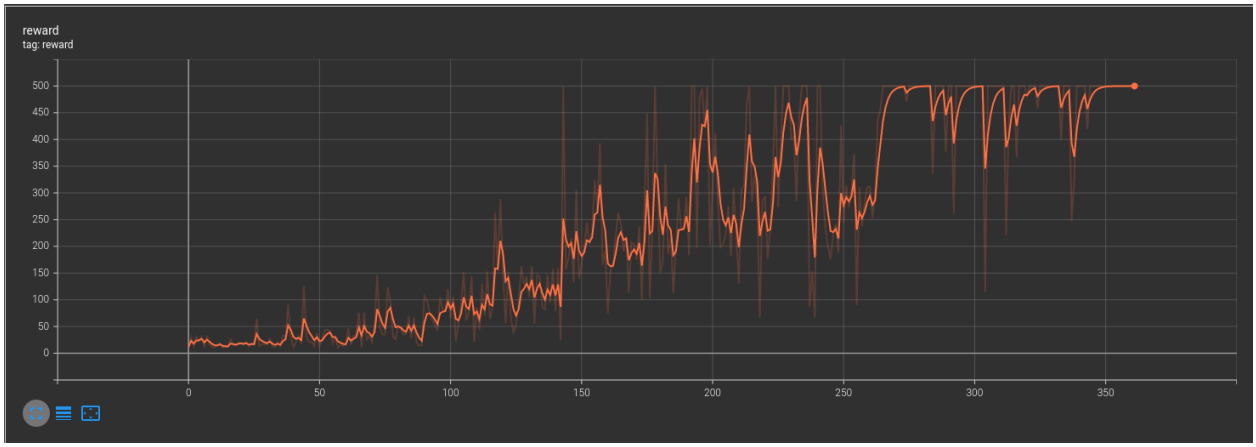
Average reward



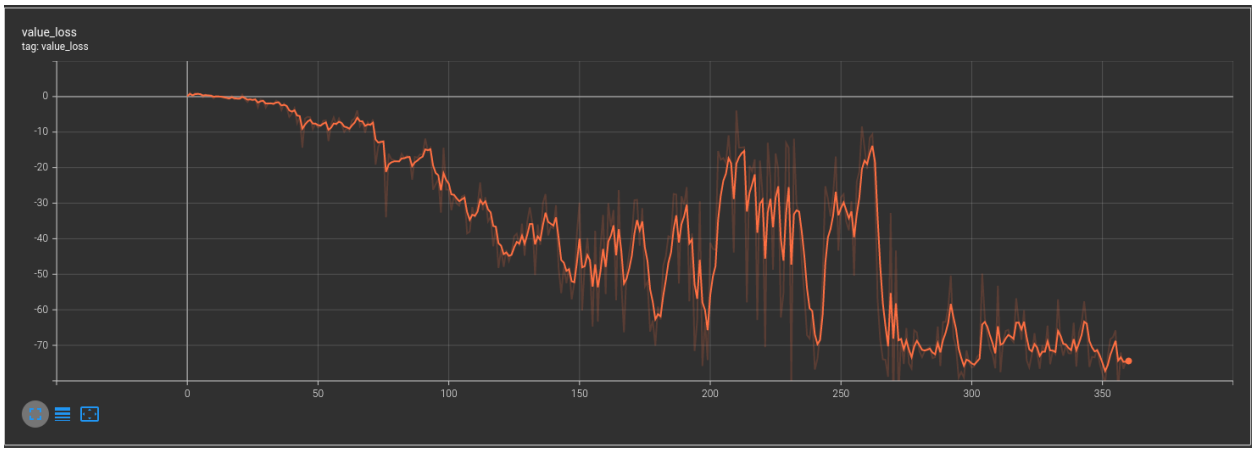
Policy Loss



Reward per step



Value Function Loss



Comparison between algorithms

We can see that the average reward is more smooth and linear rising than the reinforce algorithm without the baseline, as well as the reward per step which is more stable and has less noise. The convergence time of reinforce with baseline was around half the time it took without the baseline.

Section 2-

Question 1-

The TD error is an unbiased estimate of the advantage function

$$\delta^{\pi_\theta} = r + V^{\pi_\theta}(s') - V^{\pi_\theta}(s)$$

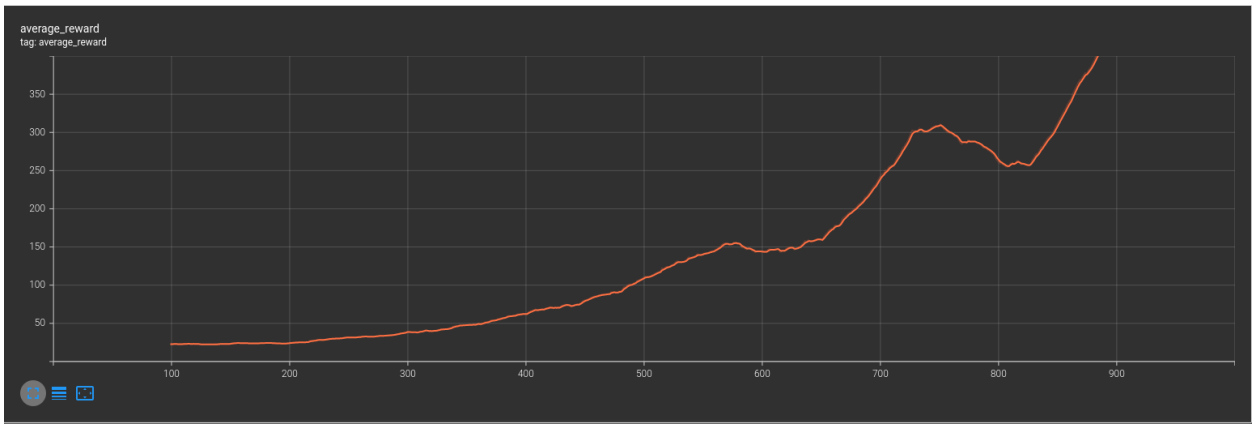
And $E_{\pi_\theta}[\delta^{\pi_\theta}|s,a] = E_{\pi_\theta}[r + V^{\pi_\theta}(s')] - V^{\pi_\theta}(s) = Q^{\pi_\theta}(s,a) - V^{\pi_\theta}(s) = A^{\pi_\theta}(s,a)$

Question 2-

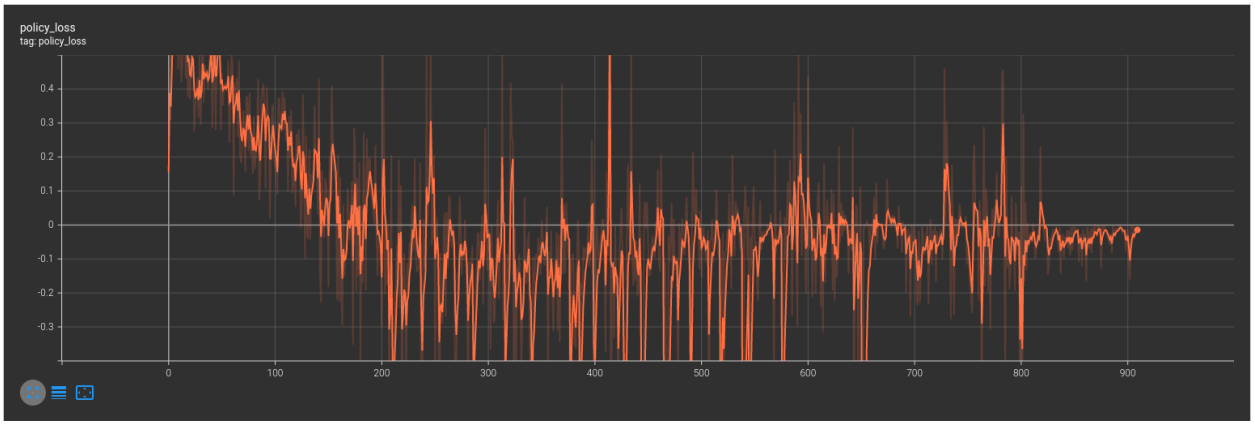
In the actor-critic methods, the actor is the policy estimator, and the critic is the value estimator. The role of the actor is to generate the policy to follow, and the role of the critic is to estimate how good the state the actor is currently at.

Solved at episode: 909

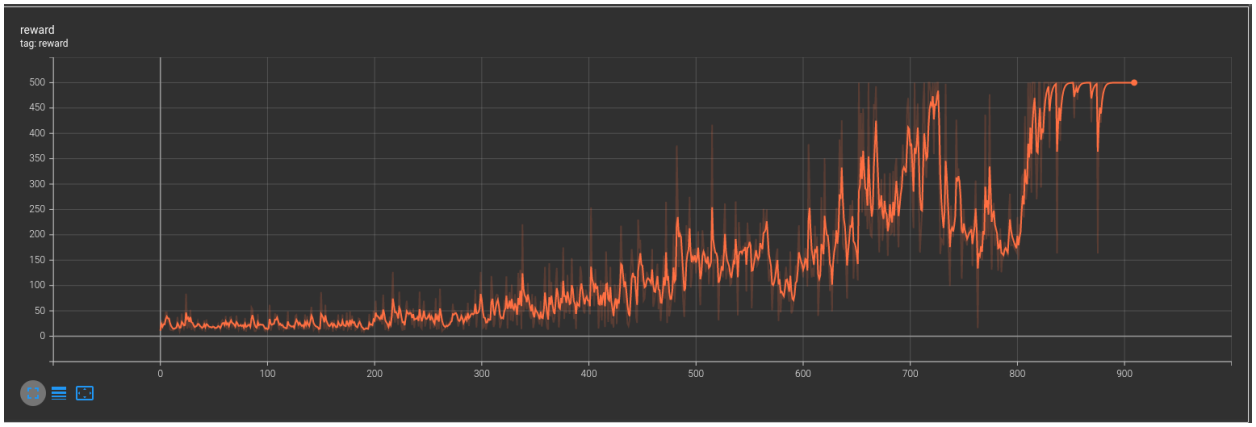
Average reward



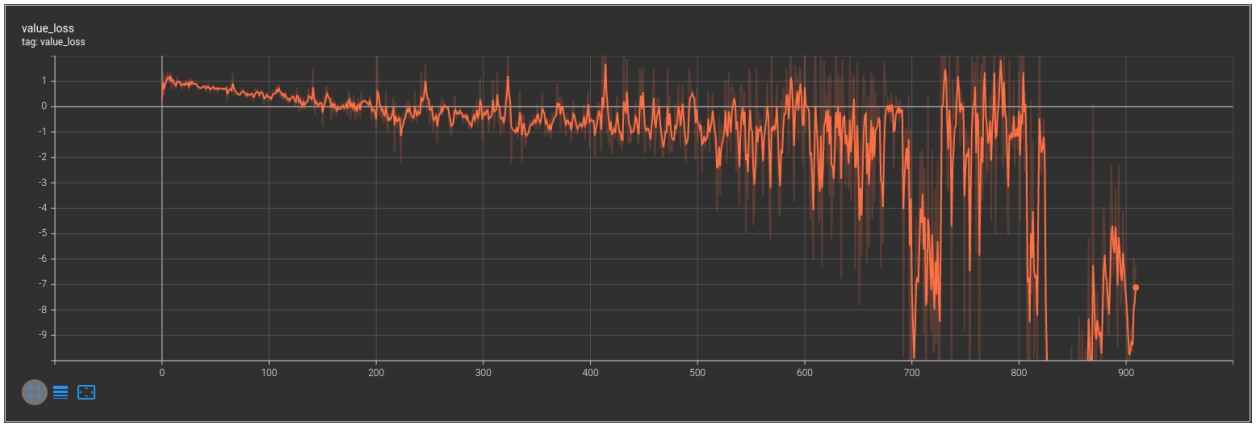
Actor loss



Reward per step



Critic loss



Comparison between algorithms

Actor critic method since they are online methods that update the parameters in each step, they introduce bias to the loss via the TD error with causes a slower convergence time than the regular reinforce algorithms but reduces the variance of the rewards and makes a smoother convergence