313326985 Shahar Shcheranski
206172686 Sarit Hollander

Report

Section 1 – Monte-Carlo Policy Gradient (REINFORCE)

1. The value of the advantage estimate reflects how better a specific action on state compared to the average of all other actions on the same state.

   It is better to follow the gradient computed with the advantage estimate instead of just the return itself because for many results there may be positive values that will affect the weights by not knowing which action is better, subtracting the average value of a state allows you to know how good an action from a state is compared to following the policy alone.

2. The reduction of the baseline does not introduce bias to the expectation of the gradient since:

$$\mathrm{E}_{\pi_\theta}[\nabla log\pi_\theta(a_t|s_t)b(s_t)] = 0$$

   Prove:

   First:

$$\nabla log\pi_\theta = \frac{\nabla\pi_\theta}{\pi_\theta} \rightarrow \nabla\pi_\theta = \pi_\theta \nabla log\pi_\theta$$

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

   Since $b(s_t)$ is not a function of a_t (constant):

$$\int \nabla\pi_\theta(a_t|s_t)b(s_t)\ d(a_t|s_t) = b(s_t)\nabla \int \pi_\theta(a_t|s_t)\ d(a_t|s_t)$$

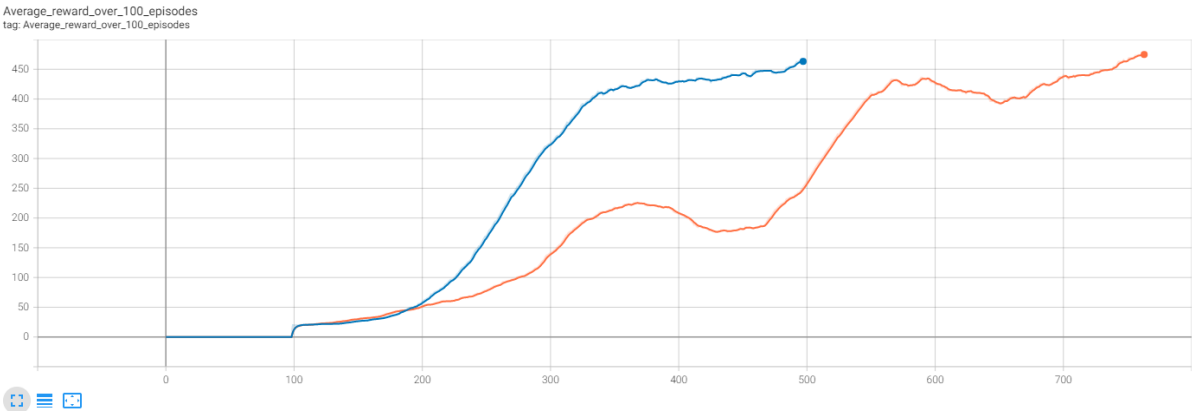
   We know that $\pi_\theta(a_t|s_t)$ is a probability distribution so its integral must be one, and the derivative of a constant is zero:

$$b(s_t)\nabla \int \pi_\theta(a_t|s_t)\ d(a_t|s_t) = b(s_t)\nabla 1 = 0$$

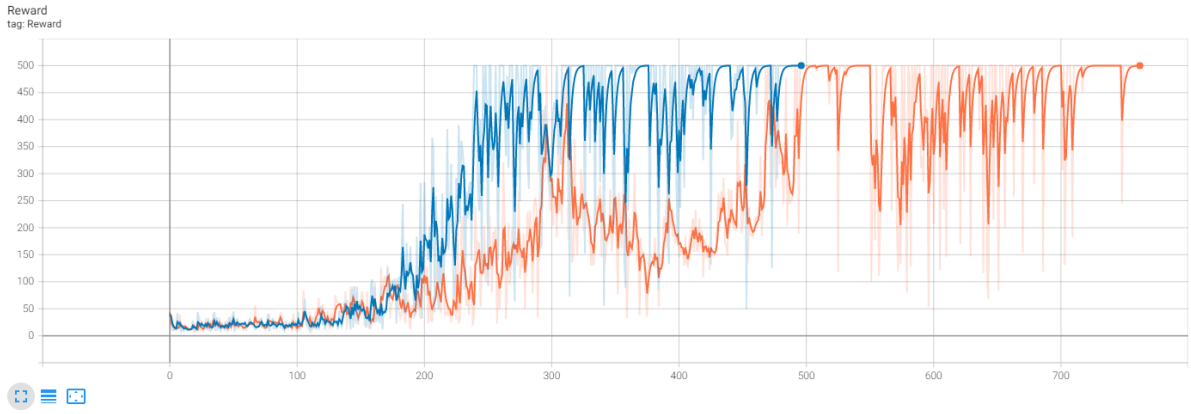
313326985 Shahar Shcheranski
206172686 Sarit Hollander

3. Comparing the results before and after the change:

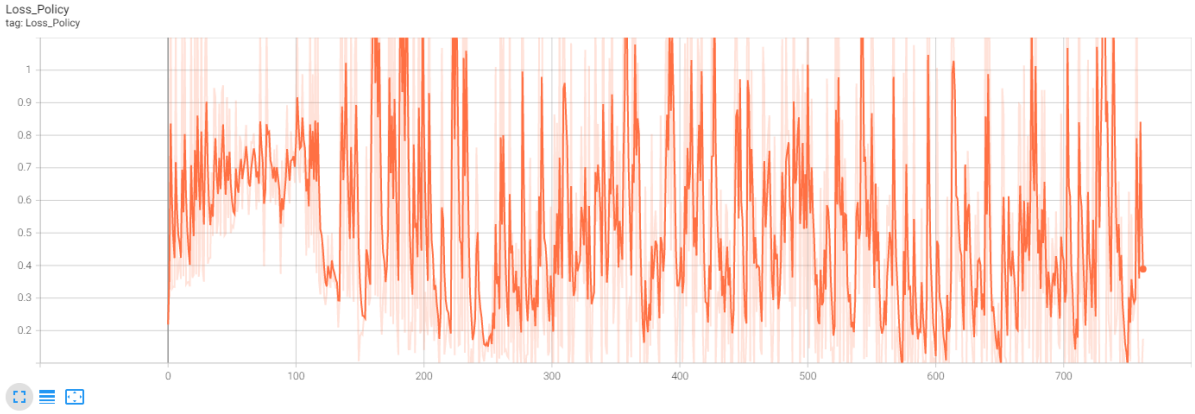
Episode average reward:



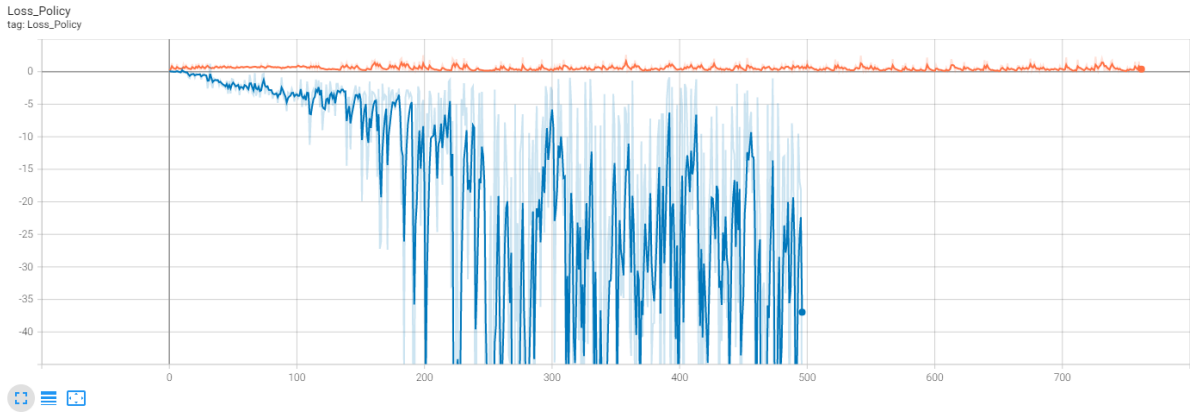
Episode total reward:



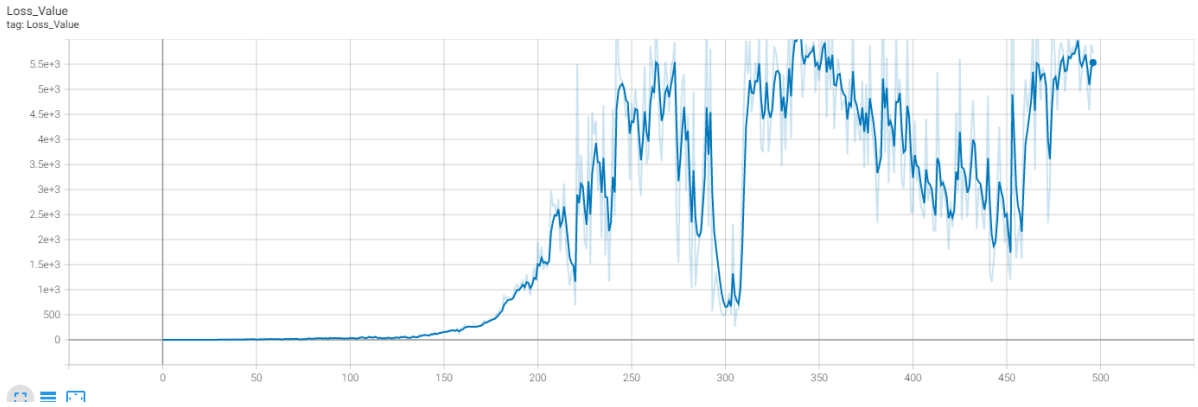
Episode policy loss:



313326985 Shahar Shcheranski
206172686 Sarit Hollander



Episode value loss:



The orange line represents the basic form of the REINFORCE algorithm and the blue line represents the REINFORCE algorithm using an advantage estimate with a value-function approximation baseline instead of the actual return of every episode.

The basic REINFORCE algorithm converged at episode 763 (after 481.428 seconds). The REINFORCE algorithm with the baseline converged at episode 511 (after 543.561 seconds).

313326985 Shahar Shcheranski
206172686 Sarit Hollander

Section 2 - Advantage Actor-Critic

1. For the value function V(s), the TD error:

   $\delta = r + \gamma V(s') - V(s)$ →

   $E[\delta \mid s, a] = E[r + \gamma V(s') \mid s, a] - V(s)$

   $since\ Q(s,a) = E[r + \gamma V(s')]$ →

   $E[\delta \mid s, a] = Q(s,a) - V(s) = A(s,a)$

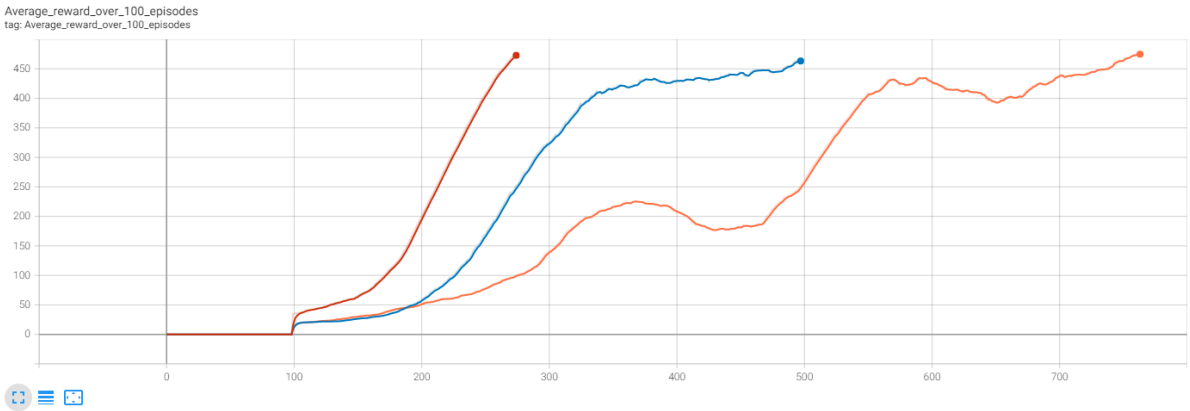
   We can see that the TD error is an unbiased estimate of the advantage function

   $A(s,a)$ so we can use TD error to update the policy gradient.

2. The actor is policy-based, which controls the agent behavior by learning the optimal policy. The actor takes as input the state and outputs the best action. The critic is a value-based which evaluates the action by computing the value function. It measures how good the action taken is.

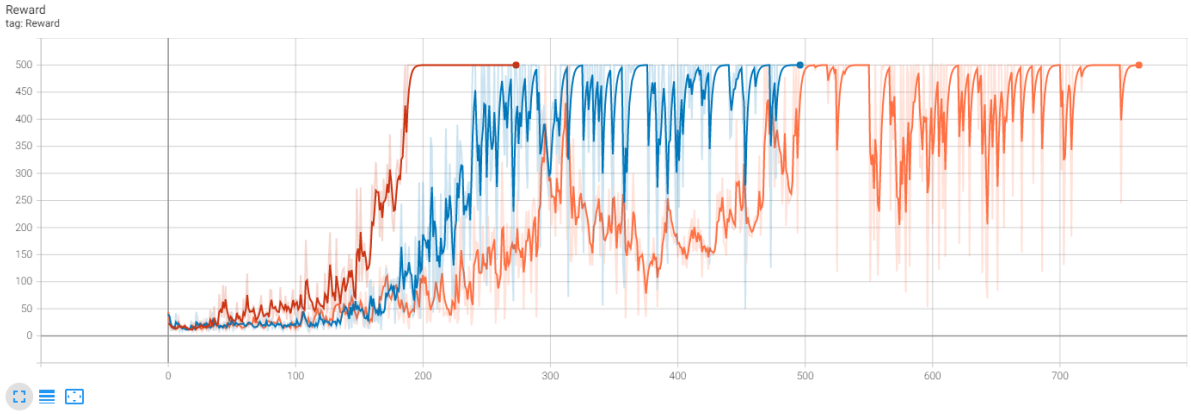
3. Comparing the results with the previous two models:

Episode average reward:

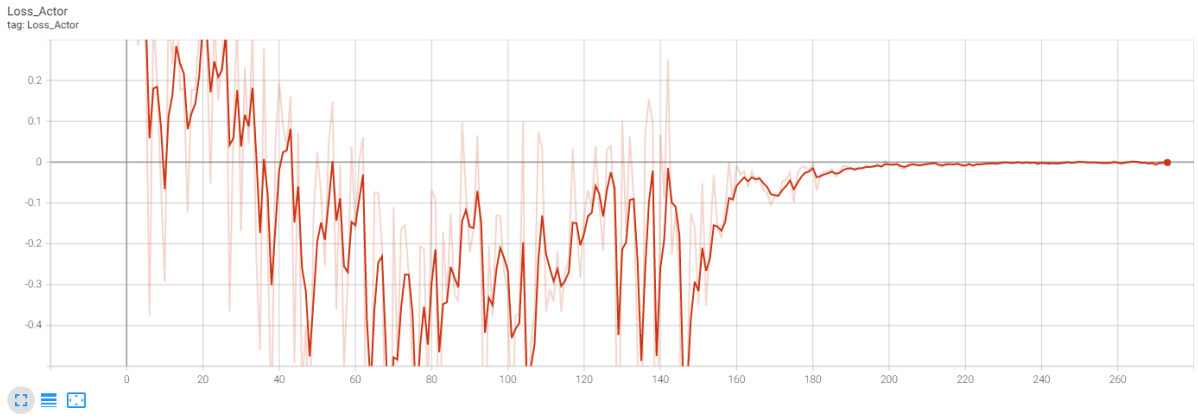


313326985 Shahar Shcheranski
206172686 Sarit Hollander

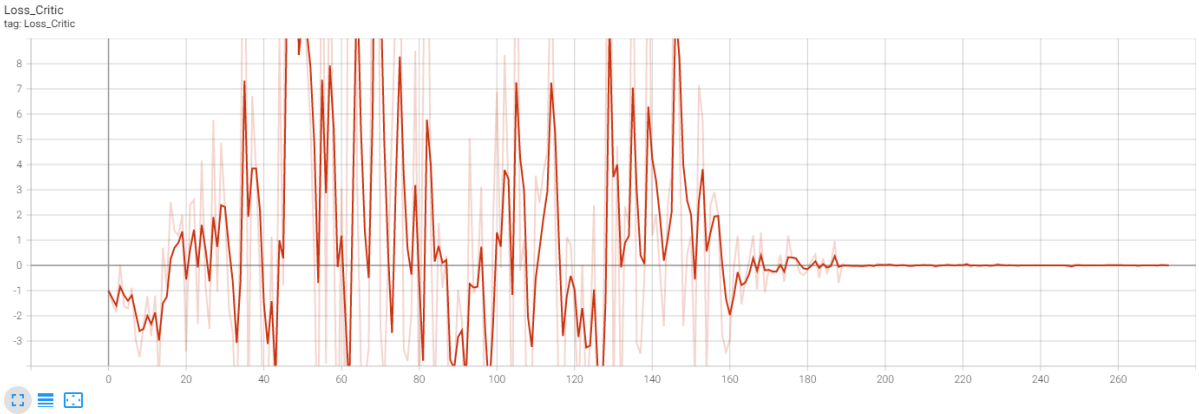
Episode total reward:



Episode actor loss:



Episode critic loss:



313326985 Shahar Shcheranski
206172686 Sarit Hollander

The orange line represents the basic form of the REINFORCE algorithm, the blue line represents the REINFORCE algorithm using an advantage estimate with a value-function approximation baseline instead of the actual return of every episode and the red line represents the Actor-Critic algorithm.

- The basic REINFORCE algorithm converged at episode 763 (after 481.428 seconds).
- The REINFORCE algorithm with the baseline converged at episode 511 (after 543.561 seconds).
- The Actor-Critic algorithm converged at episode 274 (after 283.092 seconds).

As we can see from the results, the Actor-Critic algorithm converged much faster than the previous two models.