A brief Mathematical Proof of Policy Gradient, Actor Critic and PPO

1. Policy Gradient

The main idea of Policy Gradient is to maximize the expectation of reward gaining from a trajectory sampled from a policy. To maximize the total reward, we use gradient decent algorithm.

Thus,

Take a further step:

If we sample trajectory by let the “actor” with parameter to interact with the environment, let’s say, N trajectory is generated in form , the probability of each trajectory is

Plug into eqn(2)

Then we expand the item

We can get

With eqn(3):

Hence, the loss function of Policy Gradient is

1. Policy Gradient with Baseline

For some specific environment, the reward is always true, which is unbeneficial for the network to converge.

Rewriting eqn(4) by moving reward into gradient and adding a bias to reward

Here, b is a constant or function independent of , which indicates .

The equation remains valid.

Replace total reward with discounted total gain , and replace b with a function , which means discounted reward before ,

.

Other forms of replacement are, replace with advantages .

The key difference between baselined policy gradient with traditional policy gradient algorithm mentioned in section 1 is, it can be updated after on each step instead of over the whole trajectory.

*The core of those to algorithms is identical. The only difference is, algorithm in section 1 collects gradient of each step over the whole trajectory and execute gradient decent afterwards, while the algorithm in section 2 updated at each step.*

1. Actor Critic

Following the baseline policy gradient, we replace the weight with *TD(0)* error

, it is also a neural network which called *“Critic”* .

We update the value network at each step by minimizing TD-error.

And update Actor network with policy gradient.

Then we get Actor Critic algorithm