Exploring Actor Critic Algorithms

1 Introduction

Actor-Critic methods are temporal difference (TD) learning methods that represent the policy function independent of the value function.In the Actor-Critic method, the policy is referred to as the actor that proposes a set of possible actions given a state, and the estimated value function is referred to as the critic, which evaluates actions taken by the actor based on the given policy. We implemented TD Actor Critic algorithm on 6*6 grid world, Cartpole and LunarLander.

$$Q_w(s,a) \approx Q_{\pi_\theta}(s,a)$$

parameters:- Actor-critic algorithms maintain two sets of parameters.

1. Critic:- Updates action-value function parameters w. The critic is solving a familiar problem of policy evaluation.
2. Actor :- Actor Updates policy parameters θ, in direction suggested by critic.

Actor-critic algorithms follow an approximate policy gradient

$$\nabla_\theta J(\theta) \approx E_{\pi_\theta}[\nabla_\theta \log \pi_\theta(s,a) Q_w(s,a)]$$
$$\Delta\theta = \alpha \nabla_\theta \log \pi_\theta(s,a) Q_w(s,a)$$

Unlike value based algorithms, there is no need to store the previous computed values. It is significantly faster than the value based algorithm.

2 Grid World

We have chosen 6 by 6 grid world. It is a deterministic environment with 4 possible actions (Left,Right,Up,Down). Reward distribution is as follows

the first row :- 0.1,0.2,0.3,0.4,0.5
the last col :- 0.5,0.6,0.7,0.8,1
rest all are given -1
the 5,5 position is the target position with reward of 10
The main Objective is to reach the end corner position (5,5)

3 Cartpole

Cartpole has two actions 0 to push left and 1 to push right. It gets a reward of 1 for every step taken, including the termination step. The threshold is 475. The objective is to make the pole stand for as long as possible.

Num	Observation	Min	Max
0	Cart Position	-2.4	2.4
1	Cart Velocity	-Inf	Inf
2	Pole Angle	~ -41.8°	~ 41.8°
3	Pole Velocity At Tip	-Inf	Inf

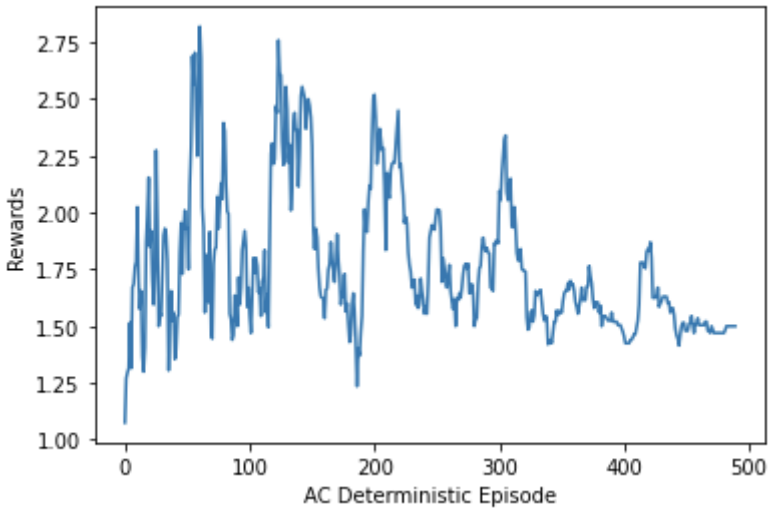
4 LunarLander

Landing pad is always at coordinates (0,0). Coordinates are the first two numbers in state vector. Reward for moving from the top of the screen to landing pad and zero speed is about 100..140 points. If lander moves away from landing pad it loses reward back. Episode finishes if the lander crashes or comes to rest, receiving additional -100 or +100 points. Each leg ground contact is +10. Firing main engine is -0.3 points each frame. It has 4 possible actions.
The main Objective is to land on the launch pad in between the flags.

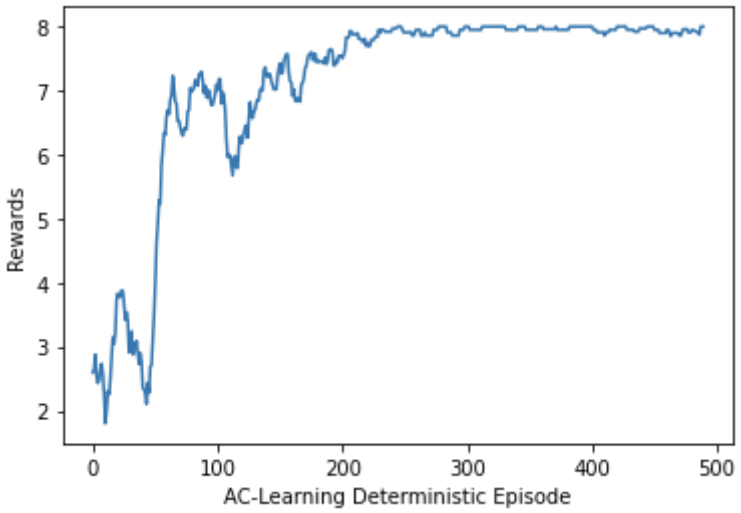
5 Results

5.1 Grid World

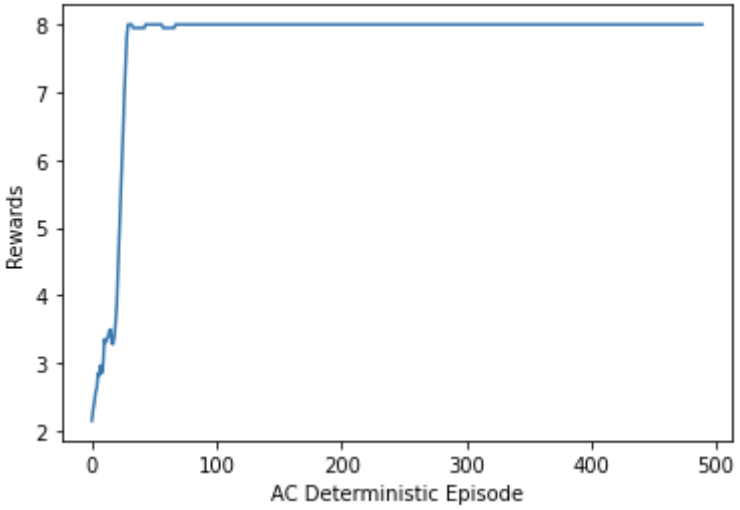
1. Parameter 1:- alpha=0.0001, beta=0.005, inputdims=2, gamma=0.9,fc1dims=16, fc2dims=32, nactions=4



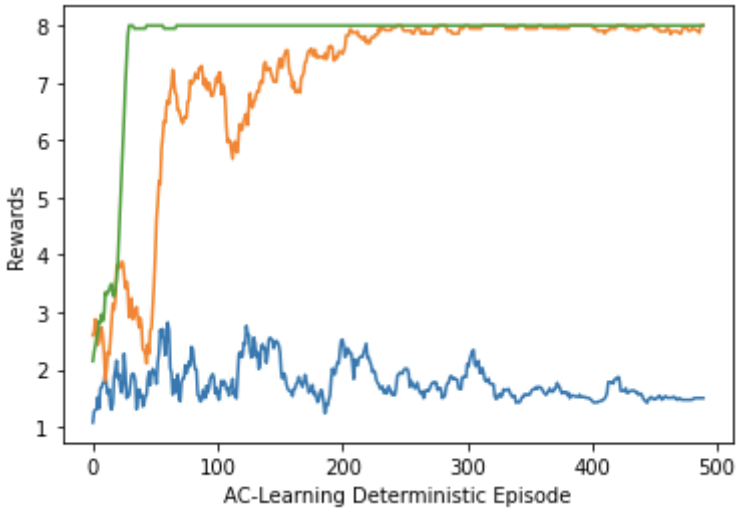
2. Parameter 2:- alpha=0.0001, beta=0.005, inputdims=2, gamma=0.9,layer1=16, layer2=32, actions=4



3. Parameter 3:- alpha=0.01, beta=0.5, inputdims=2, gamma=0.95,fc1dims=8, fc2dims=16, nactions=4

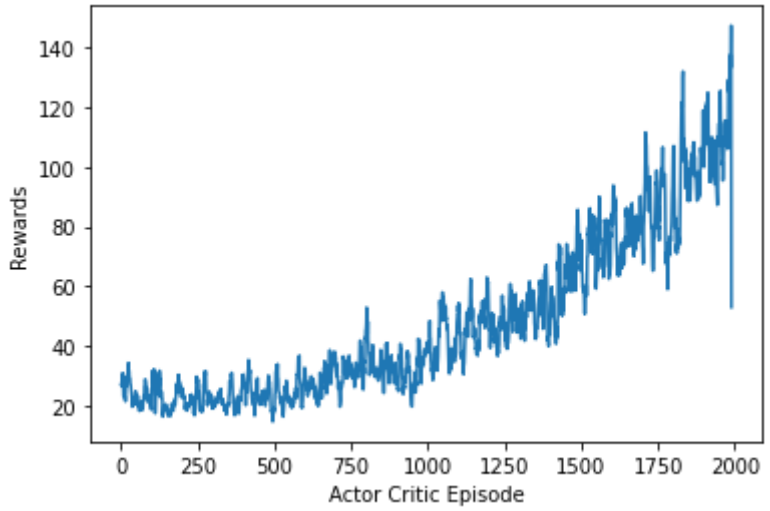


4. We got better results for parameter 3.

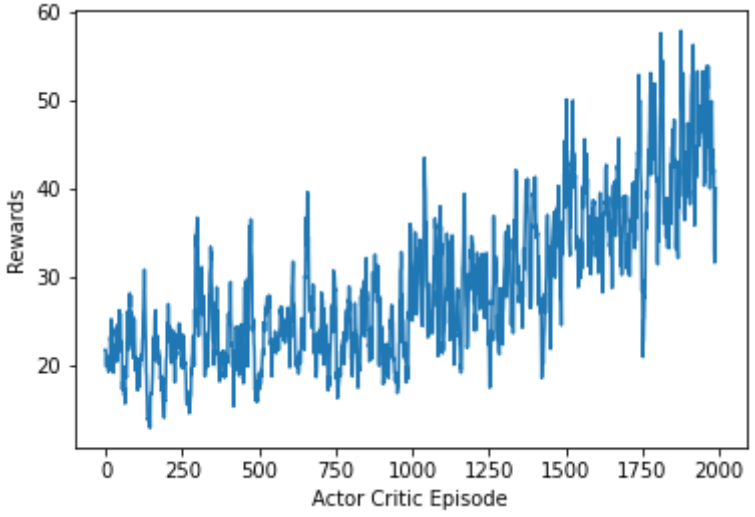


5.2 Cartpole Results

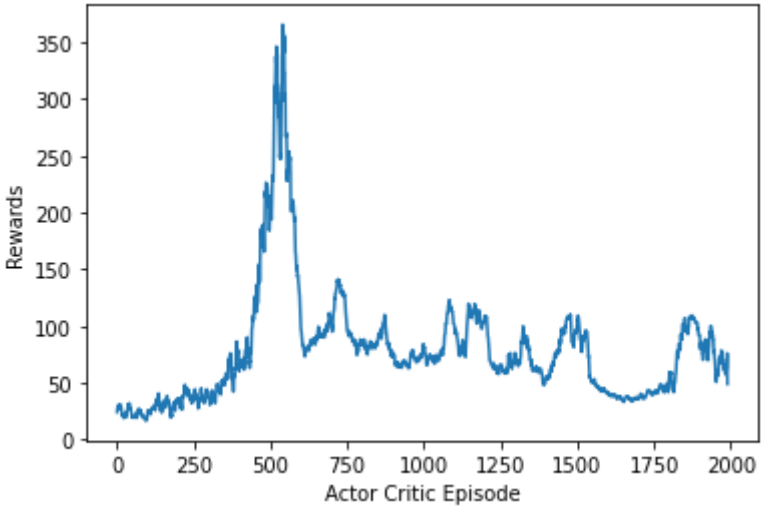
1. Parameter 1:- alpha=0.00001, beta=0.0005, input_dims=4, gamma=0.99, fc1_dims=32, fc2_dims=32, n_actions=2



2. Parameter 2:- alpha=0.00001, beta=0.0005, input_dims=4, gamma=0.99, fc1_dims=32, fc2_dims=32, n_actions=2



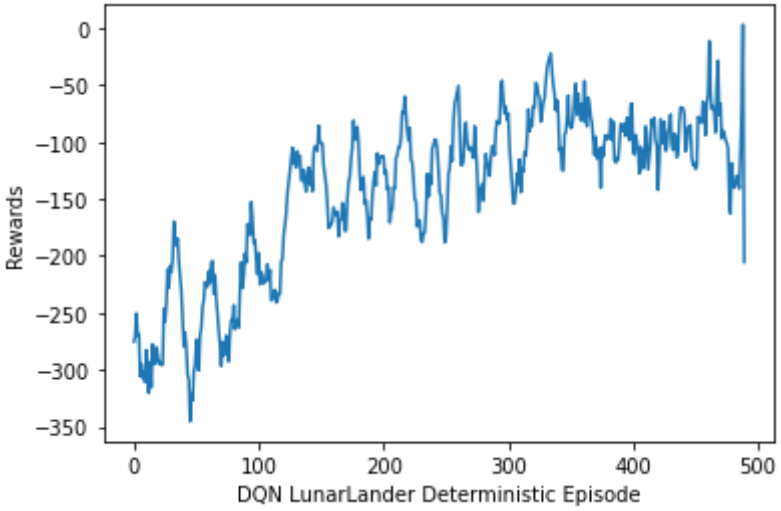
3. Parameter 3:- alpha=0.00001, beta=0.0005, input_dims=4, gamma=0.99, fc1_dims=32, fc2_dims=32, n_actions=2



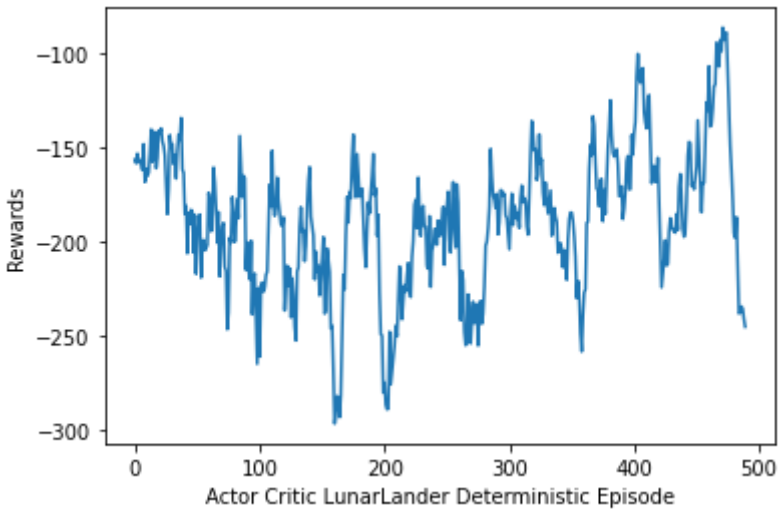
4. Got better results for Parameters 2

5.3 LunarLander

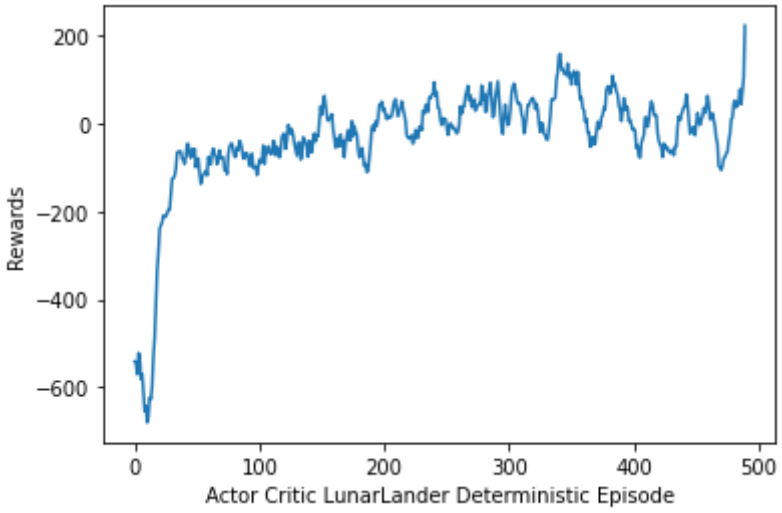
1. Parameter 1:- alpha=0.00001, beta=0.00001, input_dims=8, gamma=0.99,fc1_dims=512, fc2_dims=512, n_actions=4



2. Parameter 2:- alpha=0.00001, beta=0.00001, input_dims=8, gamma=0.99,fc1_dims=256, fc2_dims=256, n_actions=4



3. Parameter 3:- alpha=0.0001, beta=0.0001, input_dims=8, gamma=0.99,fc1_dims=256, fc2_dims=512, n_actions=4



4. Got better results for Parameter 3