Cartpole Revisited  
CS 370 Current/Emerging Trends in Computer Science

Sergio Mateos

Southern New Hampshire University

The REINFORCE algorithm is a type of policy gradient that maximizes the excepted reward by encouraging the actor/user to interact with its environment based on the policies. The cartpole problems are a classic dilemma of attraction of the cart to the pole via a hinge. The main result is to balance the pole as a vertical line by moving the cart left and right. The AI would reward the most accuracy for balancing the pole in the direction that will keep the pole in the center balanced with the mass point up.

A picture containing diagram, line, text, design

Description automatically generated

initialize policy parameters randomly.

repeat until convergence:

generate a trajectory by following the current policy.

compute the total reward for the trajectory.

compute the gradient of the policy objective with respect to the policy parameters.

update the policy parameters using the gradient and a learning rate.

Advance Actor-Critic or A2C is an algorithm with another policy with a single neural network to contemplate and estimate the policy and the value of the function. “How better taking that action at a state is compared to the average value of the state.” (AAC). One of the main differences between REINFORCE and A2C is the advantage function which is the difference between reward and action in the current state. The use of A2C to solve the cartpole problem is by selecting an actor and assigning an action that would be picked based on the critique. When the action is executed, the prediction would allow a balance in the pole, the longer the pole is balanced the higher reward for the assignment.

A diagram of a person with a stick figure

Description automatically generated with low confidence

Define the actor function.

Define the critic function.

Generate trajectories.

Calculate the Advantage function.

Calculate the policy gradients.

Calculate the critical loss.

Update the actor parameters.

Update the critical parameters.

Exit upon policy convergence.

The policy gradient approach is a type of reinforcement that estimates the value of a policy by computing the gradient focusing on the maximization of the reward. Q-learning is a value-based approach, the expected cumulative reward which is obtained by adjusting the action of the action value. The action is selected by the learned value pairs and calculated by the action-value function. The policy gradient approach and Q-learning differ since the policy gradient approach can handle continuous action, and Q-learning typically is limited to discrete action. Also, Q-learning updates the value of the reward based on an equation compared to the policy gradient approach which updates the reward based on the action taken.

Actor-critic approaches are methods that represent the policy independent of a function and memory structure is separated. The actor-critic approach combines elements of value and policy-based. The actor-critic use two neural networks, the actor, and the critic network. The actor-network is used to generate policy and the critic is used to estimate the value function.

Reference

Juliani, A. (2016, December 16). *Simple reinforcement learning with TensorFlow part 8: Asynchronous Actor-Critic Agents (A3C)*. Medium. https://medium.com/emergent-future/simple-reinforcement-learning-with-tensorflow-part-8-asynchronous-actor-critic-agents-a3c-c88f72a5e9f2

Advantage Actor Critic (A2C). (n.d.). Huggingface.co. <https://huggingface.co/blog/deep-rl-a2c>

Yoon, C. (2018, December 29). *Deriving policy gradients and implementing REINFORCE*. Medium. https://medium.com/@thechrisyoon/deriving-policy-gradients-and-implementing-reinforce-f887949bd63