**6-2 Assignment: Cartpole Revisited**

Aaron Ciminelli

SNHU

CS-370-T4235 Current/Emerging Trends in CS 23EW4

Venkatesh Baglodi, Ph.D.

April 9th, 2023

The REINFORCE algorithm is a policy gradient method that can be used to solve the Cartpole problem, a well-known challenge in reinforcement learning. This challenge requires balancing a pole on a cart by applying force in two directions to the cart. The REINFORCE algorithm learns to maximize the policy by developing a stochastic policy to determine the probability distribution of actions. This is accomplished by fine-tuning the policy settings based on the rewards received.

To run the REINFORCE algorithm, first set the policy parameters. Following that, for each episode, construct a trajectory using the existing policy, compute the total reward for the trajectory, compute the policy gradient using the trajectory, and update the policy parameters.

Create a policy that relates the current state to action in order to solve the Cartpole issue with the REINFORCE algorithm. A neural network can represent the policy, which takes the current state as input and produces a probability distribution across actions. The agent then chooses an action from the probability distribution and applies it to the cart.

The A2C algorithm, on the other hand, is another policy gradient approach for solving the Cartpole problem. To optimize the policy, it blends value-based and policy-based techniques. The policy-based approach establishes the probability distribution of actions, whereas the value-based approach estimates the state-value function.

To begin running the A2C algorithm, set the policy and value function parameters. Then, for each episode, construct a trajectory using the current policy, compute the total reward for the trajectory, compute the policy gradient and value function gradient using the trajectory, and update the policy and value function parameters.

Create a policy and a value function that convert the current state to an action and a state-value estimate, respectively, to solve the Cartpole issue using the A2C method. Separate neural networks that take the current state as input and output the matching values can be used to represent the policy and value functions.

Policy gradient techniques, such as REINFORCE and A2C, optimize the policy directly by altering policy parameters in response to the incentives obtained. Value-based techniques, on the other hand, such as Q-learning, estimate the ideal action-value function by iteratively updating the Q-values based on the rewards received.

Policy-based methods are frequently used in continuous action spaces, but value-based methods are more appropriate in discrete action spaces. Furthermore, policy-based methods are capable of dealing with stochastic policies, whereas value-based methods require deterministic policies.

Policy gradient and value-based approaches are both incorporated into actor-critic methodologies. A separate actor network determines the policy, while a critic network estimates the state-value function.

The actor network generates the policy, while the critic network estimates the state-value function. The policy gradient is used to update the actor network, while the temporal difference error between the expected and actual state values is used to update the critic network.

Because they use the critic network to reduce variance in policy gradient estimations, actor-critic methods are more stable than pure policy gradient or value-based systems. The critic network serves as a baseline for the policy gradient, decreasing noise in gradient estimates and increasing the stability of the learning process.

Furthermore, actor-critic approaches can learn deterministic or stochastic policies in both continuous and discrete action spaces. Deep reinforcement learning uses actor-critic approaches because they are efficient, stable, and can handle complicated environments.

Two common policy gradient approaches for solving the Cartpole problem are the REINFORCE and A2C algorithms. In how they optimize policy, policy gradient approaches differ from value-based approaches. Actor-critic techniques, which are popular in deep reinforcement learning, combine the benefits of policy gradient and value-based approaches. The nature of the problem and the application requirements influence algorithm selection.

**References:**

Fang, Jie, Yunqing Rao, and Mingliang Shi. 2023. “A Deep Reinforcement Learning Algorithm for the Rectangular Strip Packing Problem.” PLoS One 18 (3): e0282598.

Yoon, C. (2018, Dec). *Deriving policy gradients and implementing reinforce*. Medium. Retrieved April 8, 2023, from https://medium.com/@thechrisyoon/deriving-policy-gradients-and-implementing-reinforce-f887949bd63

Yoon, C. (2019, February). *Understanding actor critic methods*. Medium. Retrieved April 8, 2023, from https://towardsdatascience.com/understanding-actor-critic-methods-931b97b6df3f