Matthew Muller

**6-2 Assignment: Cartpole Revisited**

The REINFORCE algorithm is a policy-based approach that works by calculating a probability distribution that stochastically determines the action chosen at each given state based on the expected future rewards of that action. This means that as the agent learns which actions at a given state lead to higher rewards, the probability distribution is updated to increase the likelihood of that action being chosen in future episodes. Applying this algorithm to the CartPole problem begins with building a neural network to serve as the policy network that will accept a state vector as input and produce a discrete probability distribution over the possible actions (Joshi). The input (state) vector will have a length of 4 in order to store the cart position, cart velocity, pole position, and pole velocity at the tip. The output layer will be a SoftMax layer to produce the probability distributions over the actions. Possible pseudocode for the REINFORCE algorithm may look like this:

**function REINFORCE**

Initialize policy parameter arbitrarily

**for** each episode **do**

**for** time = 1 to Terminal time -1 **do**

Policy function parameterized by neural network

**end for**

**end for**

**return** policy parameter

**end function**

The Advantage Actor Critic (A2C) algorithm works to improve some of the issues of basic policy gradients with a new update equation that replaces the discounted cumulative award with the Advantage function. In Actor Critic methods such as this, the “Critic” estimates the value function, which could be the action-value (Q value) or state-value (V value). The “Actor” then updates the policy distribution in the direction suggested by the Critic. On each learning step, the Actor parameter is updated with policy gradients and advantage value and the Critic parameter works to minimize the mean squared error (Advantage value) with the Bellman update equation (Yoon). In the CartPole problem, the reward at each timestep stops when the cart strays too far from the center or the pole exceeds 15 degrees from vertical. The end of an episode is unexpected to the Critic and will commonly result in a larger negative Advantage value. This will cause the actor to decrease the likelihood of taking the actions that led to the episode ending again.

One of the main differences between policy gradient approaches and value-based approaches such as Q-learning is that policy gradient approaches eliminate the need to design an algorithm (e.g. epsilon-greedy) for selecting the best action. Instead, an action is chosen stochastically based on a probability distribution directly at each state rather than chosen based on an assigned value with a set chance to opt for exploration. Another difference between the two approaches is that in Q-learning, various methods such as experience replay and target networks are employed to stabilize the learning process while a policy network does not have these complexities (Joshi).

Actor-critic approaches seek to improve some of the issues of policy-based approaches such as noisy gradients, high variance, and lack of accounting for trajectories that have a cumulative reward of 0. The Actor-Critic approach consists of two networks, with the Actor learning based on a policy gradient approach and the Critic evaluating the Actor’s action by computing the value function. One of the most common ways that this approach reduces variance and increases stability is by subtracting cumulative rewards by a baseline. Thus, Actor-Critic methods combine both policy-based and value-based approaches so that the benefits of each can be utilized to create a more successful learning agent.

**Works Cited**

Joshi, NandaKishore. “Part 1&nbsp;: Policy Based Reinforcement Learning - A Detailed Study.”

Medium, Medium, 26 Feb. 2022, https://nandakishorej8.medium.com/part-1-policy-based-reinforcement-learning-a-detailed-study-1d4e9b8b5239.

Joshi, NandaKishore. “Part 2&nbsp;: Policy Based Reinforcement Learning - Openai's Cartpole

with Reinforce Algorithm.” Medium, Medium, 26 Feb. 2022, https://nandakishorej8.medium.com/part-2-policy-based-reinforcement-learning-openais-cartpole-with-reinforce-algorithm-18de8cb5efa4.

Karunakaran, Dhanoop. “The Actor-Critic Reinforcement Learning Algorithm.” Medium, Intro

to Artificial Intelligence, 30 Sept. 2020, https://medium.com/intro-to-artificial-intelligence/the-actor-critic-reinforcement-learning-algorithm-c8095a655c14.

Yoon, Chris. “Deriving Policy Gradients and Implementing Reinforce.” Medium, Medium, 23

May 2019, https://medium.com/@thechrisyoon/deriving-policy-gradients-and-implementing-reinforce-f887949bd63.

Yoon, Chris. “Understanding Actor Critic Methods.” Medium, Towards Data Science, 17 July

2019, https://towardsdatascience.com/understanding-actor-critic-methods-931b97b6df3f.