**6-2 Assignment: Cartpole Revisited**

Winnie Kwong

Southern New Hampshire University

CS-370 Current/Emerging Trends in CS

Professor Balogun

February 18th, 2024

**Solving Cartpole Problem with REINFORCE Algorithm**

The REINFORCE algorithm is an algorithm within reinforcement learning used to optimize an agent's policies. (Jayakody, 2023). In the case of the cartpole problem, the agent would control the cart by using forces in the left or right direction to keep the pole balanced. The agent would be rewarded for every action that keeps the pole balanced. For every action that is a miss-step, the agent would have negative consequences for the cart falling. The policy network that would be implemented would use the current state as the input and output of the various probabilities of how the cart would move.

**Solving Cartpole Problems with A2C Algorithm**

When looking at the Advantage, Actor-Critic (A2C) can help calculate the extra rewards when taking actions when reaching a particular state; in the cart pole problem, the state would be balancing the cart. According to Yoon, “On each learning step, we update both the Actor parameter (with policy gradients and advantage value) and the Critic parameter (with minimizing the mean squared error with the Bellman update equation).” (Yoon, 2019, para 27). Each interaction gathers experience in every algorithm process while the actor continuously adjusts its weights for future predictions. The best policy would be implemented during this process to determine how each consequence would impact the different states. The critic would choose the future rewards from the current state to ensure estimating values would be more accurate and consistent. In the cartpole scenario, the longer the pole is balanced, the reward is higher. Using the A2C algorithm leads to more stable and efficient training while exploring different techniques to improve performance.

**Policy Gradient Approaches Differ from Value-Based Approaches**

Although policy gradient and value-based approaches such as Q-learning use reinforcement learning, their methods differ when achieving goals. According to Yoon, “…the objective is to learn a policy that maximizes the cumulative future reward to be received starting from any given time t until the terminal time.” (Yoon, 2018, para 5). When digging into policy gradients, they focus on maximizing the expected return with the different parameters for better improvements. On the other hand, in Q-learning, each value function is evaluated to determine the best actions to take, making it more efficient when collecting data. The policy gradient method handles continuous action while removing the random features that come with predictions. Unfortunately, due to the methods used by policy gradient, it requires more data when compared to the value-based approaches. Furthermore, with value-based approaches, they cannot handle continuous action, which can lead to learning and training being slower compared to the policy gradient. Both methods are great to use; however, depending on the environment, one must consider suitable data efficiency and stability approaches.

**Actor-Critic Approaches Differ from Value- and Policy-Based Approaches**

With the actor-critic approach, the actor can learn the policy while the critic can determine the value of the state that the actor chooses. The value-based approach focuses on understanding the value, while the policy-based approach focuses on learning about the policy from the given states. However, one of the most significant key differences is the stability of the methods. With actor-critic approaches, since it is a well-rounded method, the stability is most stable compared to both value- and policy-based approaches because the critic can predict more long-term values while reducing the variance of gradient that enables the method to become stable policy updates. Another difference is the flexibility within the actor-critic approach because of the various functions of the given process. The actor-critic approach has a better chance to adapt to the problems, while both value- and policy-based approaches because of their limitations of either space complexity or high variance learning.

**References**

Jayakody, D. (2023, September 24). REINFORCE - A Quick Introduction (with Code) - Dilith Jayakody. *Dilith Jayakody*. <https://dilithjay.com/blog/reinforce-a-quick-introduction-with-code/>

Yoon, C. (2021, December 7). Deriving policy gradients and implementing REINFORCE. *Medium*. <https://medium.com/@thechrisyoon/deriving-policy-gradients-and-implementing-reinforce-f887949bd63>

Yoon, C. (2021b, December 7). Understanding actor critic methods and A2C - towards data science. *Medium*. <https://towardsdatascience.com/understanding-actor-critic-methods-931b97b6df3f>