**(Note to the professor, the links on the page were not current. Two of the links to articles no longer exist)**

**Solving the Cartpole Problem Using Policy-Based and Actor-Critic Approaches**

The Cartpole problem, a classic reinforcement learning task, can be effectively solved using various learning techniques, including policy-based approaches like the REINFORCE algorithm. Unlike value-based methods, which estimate the expected return of actions, policy-based methods directly optimize the policy itself. The REINFORCE algorithm, a Monte Carlo policy gradient method, updates the policy by computing the gradient of expected rewards with respect to policy parameters. This is achieved by sampling trajectories, calculating rewards, and adjusting weights accordingly. The process involves three main steps: collecting episodes by interacting with the environment, computing the cumulative reward for each action taken, and updating the policy parameters using gradient ascent. The advantage of this method is its ability to handle continuous action spaces effectively, unlike Deep Q-Networks (DQN), which are limited to discrete action spaces. “In essence, policy gradient methods update the probability distribution of actions so that actions with higher expected reward have a higher probability value for an observed state.” (Yoon, 2018)

Another approach to solving the Cartpole problem is through the Advantage Actor-Critic (A2C) algorithm, which combines both value-based and policy-based methods. In A2C, two neural networks operate simultaneously: an actor network, which selects actions based on a learned policy, and a critic network, which estimates the value function to provide feedback on the action's quality. The critic evaluates how good an action is by calculating an advantage function, which helps reduce variance in training and improves stability. The actor network updates its policy based on the advantage estimate, ensuring the policy moves toward better decisions. By leveraging both policy and value estimations, A2C addresses some of the weaknesses of pure policy gradient methods, such as high variance and instability, while also improving sample efficiency.

Policy gradient approaches, such as REINFORCE, differ significantly from value-based approaches like Q-learning. Value-based methods estimate the value function to derive an optimal policy indirectly, requiring the storage of Q-values for state-action pairs. Conversely, policy gradient methods learn the policy directly by adjusting network weights based on observed rewards. This direct approach allows for more flexibility in handling continuous action spaces and stochastic policies, whereas value-based methods struggle in high-dimensional or continuous environments. The main drawback of policy gradients is their high variance, which often requires additional techniques such as baseline subtraction to stabilize learning.

Actor-critic methods, such as A2C, bridge the gap between value-based and policy-based approaches by combining the strengths of both. “A2C (Advantage Actor-Critic) is a specific variant of the Actor-Critic algorithm that introduces the concept of the **advantage function**. This function measures how much better an action is compared to the average action in a given state. By incorporating this advantage information, A2C focuses the learning process on actions that have a significantly higher value than the typical action taken in that state.“(Geeks, 2018) Unlike pure policy gradient methods, which rely solely on policy optimization, and value-based methods, which depend entirely on estimating action values, actor-critic models incorporate both components. The actor updates its policy using feedback from the critic, which evaluates the effectiveness of the chosen actions. This structure allows for more stable learning and reduced variance, addressing some of the fundamental limitations of standalone policy gradients. The use of an advantage function in actor-critic methods further refines the learning process by normalizing the updates and improving convergence rates.

In conclusion, solving the Cartpole problem using REINFORCE, A2C, and other reinforcement learning methods highlights the diverse approaches available in deep reinforcement learning. Policy-based methods offer advantages in continuous action environments, while value-based methods provide stability in learning discrete policies. Actor-critic methods, such as A2C, leverage the strengths of both approaches to achieve more efficient and stable learning. Understanding these methodologies provides deeper insights into reinforcement learning and its applications in more complex problems beyond Cartpole. Future advancements in hybrid reinforcement learning models could further improve learning efficiency, making these methods more applicable in real-world scenarios.

REFERENCES

Yoon, C. (2019, May 23). *Deriving Policy Gradients and Implementing REINFORCE*. Medium. <https://medium.com/@thechrisyoon/deriving-policy-gradients-and-implementing-reinforce-f887949bd63>

‌ *Actor-Critic Algorithm in Reinforcement Learning*. (2023, December 13). GeeksforGeeks. <https://www.geeksforgeeks.org/actor-critic-algorithm-in-reinforcement-learning/>

‌

‌