Cartpole problem being solved using REINFORCE-

The cartpole problem poses a classic challenge in reinforcement learning, requiring the maintenance of balance for a pole attached to a cart. The REINFORCE algorithm is applied to this problem by first randomly initializing policy parameters. The algorithm then engages with the environment, generating trajectories—sequences of state-action pairs—based on the current policy. From these trajectories, returns (accumulated rewards) are computed, considering future rewards through discounting. Subsequently, the policy gradient is determined for each state-action pair, guiding the adjustment of policy parameters to favor actions leading to higher returns. Through gradient ascent, these parameters are updated to maximize the expected return, and this iterative process continues until convergence to an optimal solution.

Cartpole problem being solved using A2C-

The A2C (Advantage Actor-Critic) algorithm presents another approach to tackling the cartpole problem. By training an actor (policy) and a critic (value function) concurrently, A2C optimizes the policy and value estimates. After initializing policy and value function parameters randomly, the algorithm interacts with the environment, generating trajectories. Using a gamma factor for discounting, the total reward for each trajectory is computed, and the advantage for each state-action pair is determined by evaluating the difference between the value function estimate and the actual return. Subsequently, both the value function and policy parameters are updated using the computed returns and advantages, respectively. These updates are repeated until convergence, combining policy evaluation through the critic and policy improvement through the actor.

How policy gradient differs from Q-learning-

Policy gradient approaches and value-based approaches, like Q-learning, offer distinct reinforcement learning methods. Policy gradient methods learn the policy function directly, using gradient ascent to optimize the policy parameters for maximum expected cumulative reward, while value-based approaches estimate the value function for state-action pairs, focusing on iteratively updating the Q-values.

Actor-critic differs from value- and policy standards-

Actor-critic approaches bridge the gap between these two methods by concurrently maintaining an actor and critic. The actor learns the policy, and the critic evaluates the actions, facilitating effective policy improvement. This fusion allows actor-critic methods to balance exploration and exploitation efficiently during learning. The choice between these approaches depends on the specific problem and the desired trade-offs in exploration and exploitation.

References

Phy, V. (2019, November 4). *Reinforcement learning concept on CART-pole with DQN*. Medium. https://towardsdatascience.com/reinforcement-learning-concept-on-cart-pole-with-dqn-799105ca670

Balawejder, M. (2022, February 20). *Solving Open AI’s CartPole Using Reinforcement Learning Part-2*. Medium. https://medium.com/analytics-vidhya/solving-open-ais-cartpole-using-reinforcement-learning-part-2-73848cbda4f1

*The advantages and disadvantages of policy-gradient methods - Hugging Face Deep RL Course*. (n.d.). Huggingface.co. Retrieved July 31, 2023, from https://huggingface.co/learn/deep-rl-course/unit4/advantages-disadvantages?fw=pt

‌ Yoon, C. (2019, July 17). *Understanding Actor Critic Methods*. Medium. https://towardsdatascience.com/understanding-actor-critic-methods-931b97b6df3f

‌