Cartpole Revisited

CS 370 Current/Emerging Trends in CS

Nicholle Caudy

8/6/2025

REINFORCE is a method in reinforcement learning classified as a policy gradient algorithm, which is applied to optimize decision-making processes. The agent learns by trying actions and then adjusting to chances of those actions based on the total reward received afterward (GeekforGeeks, 2025). The policy serves as a framework for the agent, providing guidance on the appropriate actions to take in each state. It is refined iteratively until a solution to the environment is achieved (REINFORCE Algorithm, 2020). To solve the cartpole problem with this method we need to decide the rewards. Rewards are allocated for maintaining the pole in an upright position and for each time-step the pole remains upright on the cart. Set up the environment by defining the policy with rules, states, actions, and rewards at each step. The input is the state, and the output is a probability distribution over the actions. Compute the cumulative reward for the episode. Next, we initialize and update the weights of the policy network to increase the probability of actions that led to lower rewards. Run the training loop updating the policy network after each batch recording the states, actions, and rewards from each episode and compute the returns. The policy is subsequently revised according to the observed returns. Finally, we let the trained agent run in the environment without updating the policy and will see that the agent will choose actions based on gaining the highest rewards mastering the cartpole problem. (GeekforGeeks, 2025) (Yoon, 2019)

A diagram of a network

AI-generated content may be incorrect.

The Advantage Actor-Critic (A2C) algorithm is also a type of reinforcement learning algorithm. This algorithm integrates policy-based methods, where the actor is responsible for learning decision-making strategies, with value-based approaches, in which the critic evaluates the quality of those decisions. This combination enables agents to achieve more efficient learning by balancing evaluative feedback with the decision-making process. (Geek for Geeks, 2025) The advantage function evaluates the relative performance of a specific action by determining how much better or worse it is compared to preceding actions. This algorithm is intended to enhance training efficiency and minimize variance. To address the cartpole problem, the process begins with setting up the environment and specifying the actor and critic networks for the algorithm. The actor network maps the state to a probability distribution over actions, and the critic network estimates the state’s value. Then we define optimizers and loss functions for both networks. Next is the training loop where the agent interacts with the environment and calculates advantages and updates the actor and critic. The advantage function allocates greater rewards as the duration of maintaining the pole in an upright position increase. (GeekforGeeks, 2025)

A diagram of a business

AI-generated content may be incorrect.

The policy gradient approaches differ from value-based approaches, such as Q-learning in that the value-based approach focuses on the goal to learn accurate estimates of value function or Q value and policy gradient approach (REINFORCE) parameterizes the policy to maximize the expected return using a gradient ascent, making the policy the primary target for learning. Value-based methods tend to work best with discrete action spaces, selecting the best action that will produce the maximum Q-value. Policy gradient methods work well with continuous action spaces making it differ from the value-based approach. Value-based approaches tend to converge towards a deterministic optimal policy while policy gradient approaches can learn strictly through stochastic policies. (Comparing value-based and policy-based methods, n.d.)

Actor-critic approaches differ from value and policy-based approaches in that actor-critic approaches combine both value and policy-based approaches, where the actor selects an action based on the policy and the critic evaluates these actions by estimating the value function. This approach reduces variances and increases training speeds. Whereas the value-based approach learns based on the greatest reward given for an action using the Q value and policy-based approach learns making the policy the primary target of learning.

References:

Blog, G. (2020, November 24). *Reinforce algorithm: Taking baby steps in reinforcement learning*. Analytics Vidhya. <https://www.analyticsvidhya.com/blog/2020/11/reinforce-algorithm-taking-baby-steps-in-reinforcement-learning/>

GeeksforGeeks. (2025, June 5). *Reinforce algorithm*. <https://www.geeksforgeeks.org/machine-learning/reinforce-algorithm/>

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

GeeksforGeeks. (2025b, July 23). *Actor-critic algorithm in Reinforcement Learning*. <https://www.geeksforgeeks.org/machine-learning/actor-critic-algorithm-in-reinforcement-learning/>

*Actor-critic methods: Combining value-based and policy-based methods*. SoftwarePatternsLexicon.com. (2024, November 25). https://softwarepatternslexicon.com/neural-networks/viii.-reinforcement-learning-patterns/2.-advanced-rl-techniques/actor-critic-methods/

Comparing value-based and policy-based methods. (n.d.). https://apxml.com/courses/intro-to-reinforcement-learning/chapter-8-introduction-policy-gradient-methods/comparing-value-policy-based-methods