Assignment 1 – Reinforcement Learning on CartPole-v1

Course: Reinforcement Learning

Submitted By: Aryan Aarav

Date: 4th August 2025

1. Introduction

The goal of this assignment is to solve the CartPole-v1 environment using three reinforcement learning algorithms: Deep Q-Network (DQN), Policy Gradient (REINFORCE), and Actor-Critic (A2C). The models were implemented in PyTorch, trained for 1000–1500 episodes, and evaluated using reward plots and convergence metrics.

2. Methodology

• DQN – Value-based approach using replay buffer, target network, and ε-greedy exploration.

• REINFORCE – Monte Carlo policy gradient with softmax action probabilities and return-based updates.

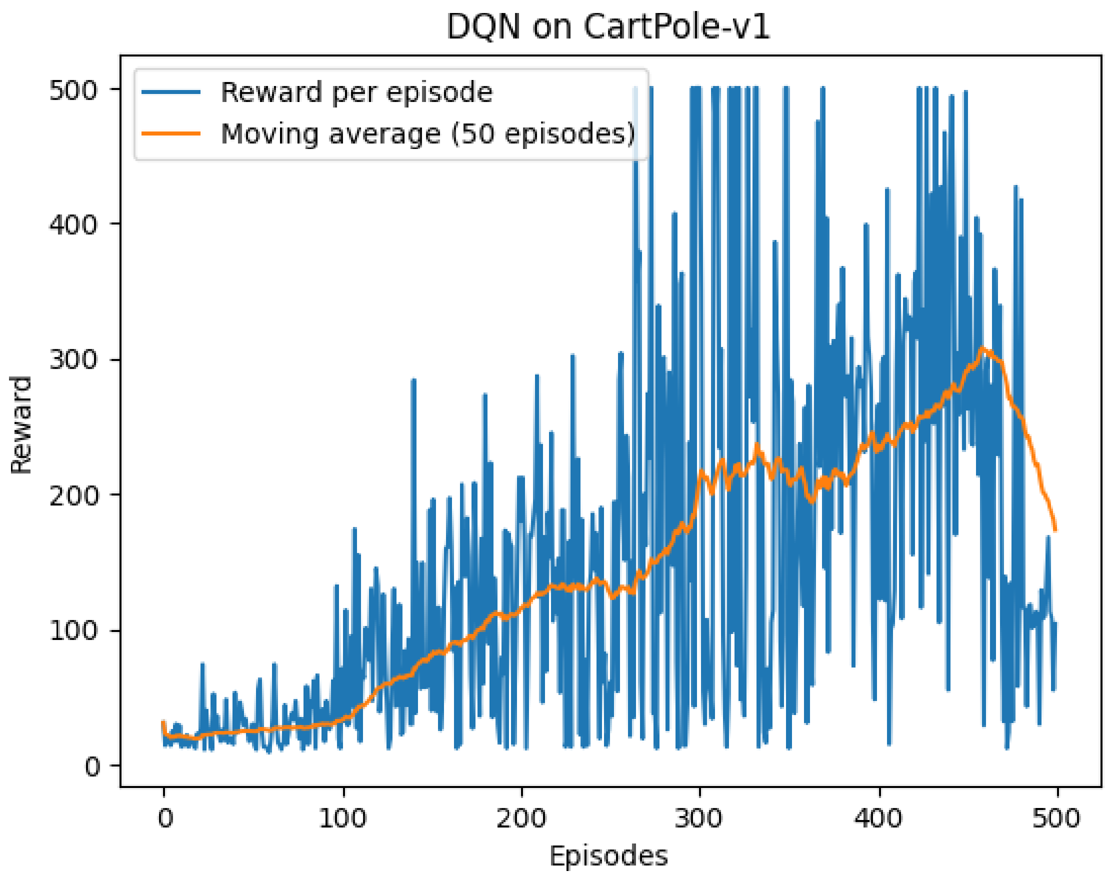
• A2C – Actor-Critic method using TD error, advantage estimation, and joint policy-value updates.

3. Hyperparameters

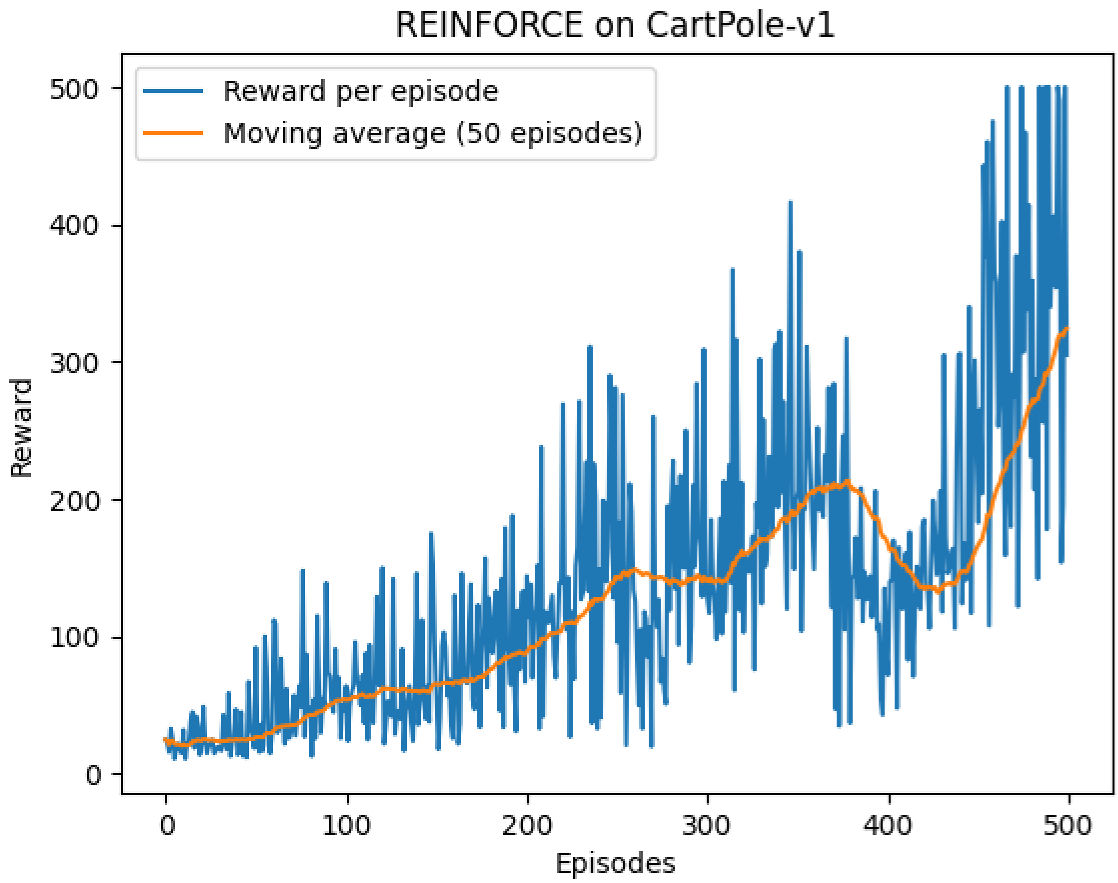
Algorithm	Learning Rate	Gamma	Episodes	Hidden Layers	Batch Size	Replay Buffer	Epsilon Decay
DQN	1e-3	0.99	1500	64-64	128	10000	0.995
REINFORCE	1e-3	0.99	1500	128	-	-	-
A2C	5e-4	0.99	1500	128	-	-	-

4. Results

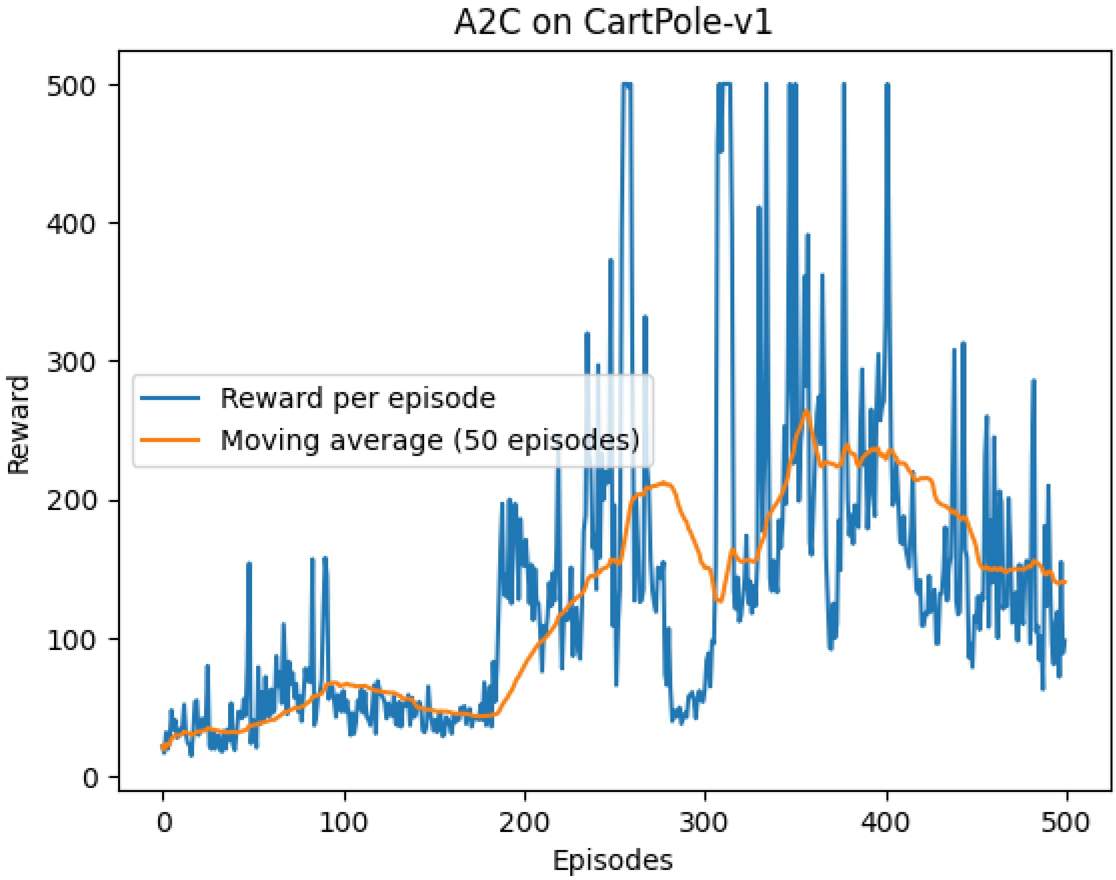
DQN Training Curve



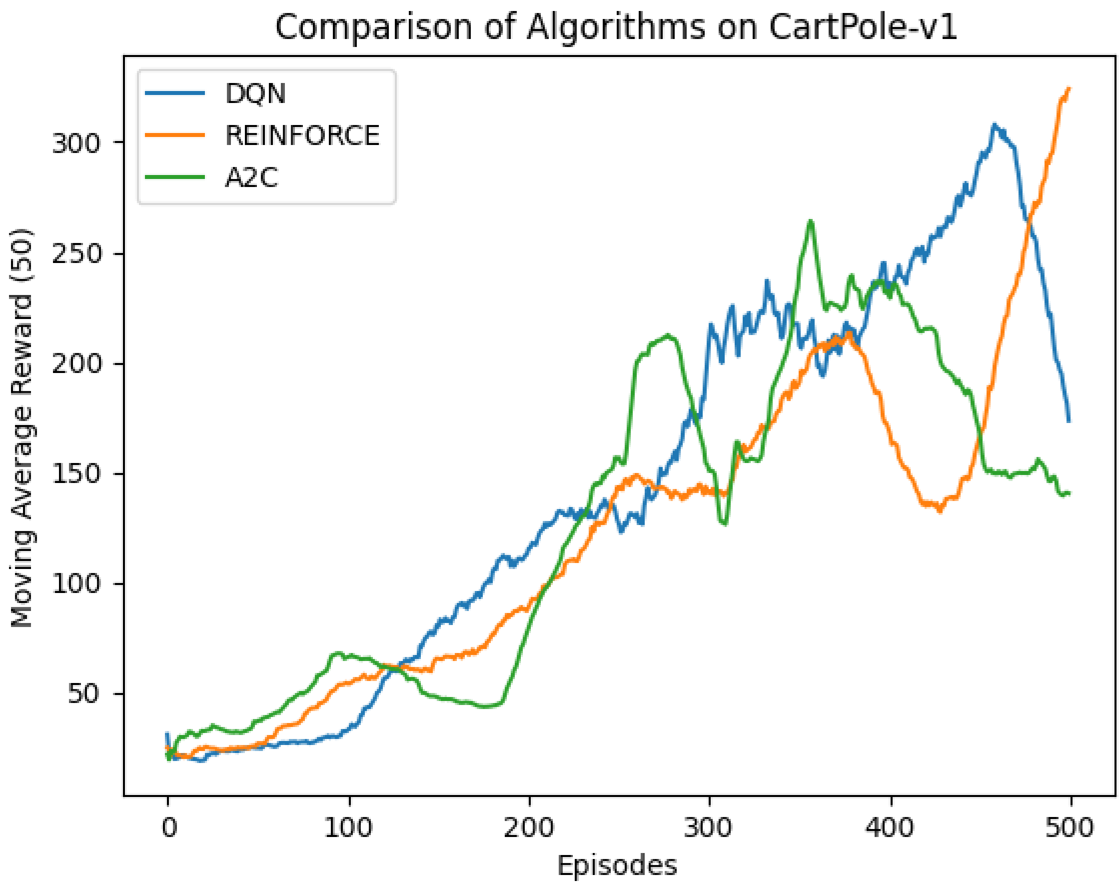
REINFORCE Training Curve



A2C Training Curve



Comparison of Algorithms



Final Results Table

Algorithm	Avg Reward (Last 100)	Convergence Episode
DQN	232.17	299
REINFORCE	244.27	351
A2C	155.39	259

5. Analysis & Discussion

The DQN agent successfully solved the environment, converging in ~299 episodes with an average reward of 232.17. REINFORCE achieved the best stability and performance, converging at ~351 episodes with a higher average reward of 244.27. A2C demonstrated partial learning, converging earlier (~259 episodes) but achieving a lower final average reward of 155.39, highlighting its instability in this simple environment.

6. Conclusion

This assignment compared three reinforcement learning algorithms on CartPole-v1. DQN and REINFORCE solved the environment reliably, while A2C was less stable. These results emphasize the differences between value-based, policy gradient, and actor-critic approaches in terms of convergence, stability, and performance.