REINFORCE vs Actor-Critic

Both are policy gradient algorithms

**REINFORCE**

It optimizes the expected return by following the gradient of the log-policy scaled by the returns obtained from full episodes.

Algorithm summary:

-Collect full episodes using the correct policy (Monte Carlo)

-Compute return

-Update parameters with

Characteristics:

-Simple and unbiased gradient estimate

-No value function required

-High variance due to full-episode returns

-Inefficient sample usage (one update per episode)

-Convergence can be very slow

**Actor-Critic**

It combines REINFORCE and value-based methods. It uses two models:

Actor: learns the policy

Critic: learns a values function to estimate expected returns

It uses temporal-difference learning to estimate the advantage

Update rule

Characteristics:

-Lower variance due to baseline V(s)

-Online updates possible

-Faster and more stable learning

-Biased gradient estimate due to bootstrapping

-Additional model must be trained (extra complexity)

-Sensitive to the quality of the value estimate

**Environment**

Observation space: continuous state vector Box([-inf -inf -inf -inf -inf -inf -inf -inf -inf -inf -inf], [inf inf inf inf inf inf inf inf inf inf inf], (11,), float64)

Action space: continuous joint torques Box([-1. -1. -1.], [1. 1. 1.], (3,), float32)

**Results**

