Slide 1: Introduction

Policy Gradient (PG): A family of reinforcement learning algorithms that directly optimize the policy.

REINFORCE Algorithm: A Monte Carlo policy gradient method for optimizing policies.

Vanilla Policy Gradient: A basic form of policy gradient algorithm without additional enhancements.

Actor-Critic Algorithms: Methods combining policy-based (actor) and value-based (critic) approaches.

Proximal Policy Optimization (PPO): A policy gradient algorithm designed to balance stability and performance.

Slide 2: Contents

Policy Gradient Algorithms: Techniques to directly improve policies by maximizing rewards.

Actor-Critic Algorithms: Combines policy and value functions for learning.

PPO: An advanced policy optimization technique that imposes constraints to stabilize updates.

Slide 3: Policy Gradients

Optimal Policy: A policy that maximizes the expected cumulative reward.

Parameterized Function: A function where the policy is defined using parameters (e.g., neural network weights).

Slide 4: Deep Reinforcement Learning (DRL) Algorithm Classes

Model-Based RL: Algorithms that use a model of the environment to plan actions.

Model-Free RL: Algorithms that learn directly from environment interactions without modeling it.

Advanced Actor-Critic: Variations of actor-critic methods with additional features like eligibility traces.

Slide 5-6: Policy-Based Approaches

Continuous Action Spaces: Spaces with infinite possible actions, making value-based approaches computationally expensive.

Stochastic Policies: Policies that output a probability distribution over actions rather than deterministic actions.

Slide 7-8: Policy Parameterization

Expected Return: The sum of rewards over a trajectory, weighted by discount factors.

Trajectory: A sequence of states, actions, and rewards in an episode.

Slide 9-10: Gradient Ascent

Gradient Ascent: An optimization technique to maximize a function by updating parameters in the direction of the gradient.

State Distribution: The probability distribution of states encountered by the policy.

Slide 11-13: REINFORCE Algorithm

Monte Carlo Sampling: A method to estimate expected values by sampling trajectories.

Log Probability (Log π): Used in gradients to simplify computation and avoid exponentiation.

On-Policy Learning: Learning where the policy used for collecting data is the same as the one being optimized.

Slide 14-16: Vanilla Policy Gradient

Advantage Function: A measure of how much better an action is compared to a baseline, used to reduce variance in gradient estimation.

Bias or Baseline: A value (e.g., state-value) subtracted from rewards to stabilize learning.

Slide 17-18: Trajectory Dataset

Softmax Layer: A function used in the policy to output probabilities for actions.

Slide 19-20: Actor-Critic Methods

Actor: The component responsible for policy updates.

Critic: The component responsible for value function updates, which aids the actor in learning.

Slide 21: One-Step Actor-Critic

TD Error: Temporal difference error used to update both actor and critic.

Slide 22-25: Proximal Policy Optimization (PPO)

Policy Ratio: The ratio of new to old policy probabilities, used to constrain updates.

Clipped Surrogate Objective: A method to limit large updates and stabilize training.

Slide 26-29: Continuous Action Spaces

Normal Distribution: A probability distribution used to model continuous action spaces.

Parameterized Approximators: Functions representing mean and standard deviation to define the policy.

Slide 30-32: Summary

Continuous Action Space: Actions defined in a continuous range rather than discrete choices.

Policy Gradient Techniques: Methods to improve policies by directly maximizing rewards.

Actor-Critic: Combines policy and value-based methods for efficient learning.

PPO: A robust policy gradient algorithm designed for stability.