Performance Comparison of PG, PPO, and TRPO on Benchmark Environments

Trijal Srivastava

This report presents a comparative performance analysis of three Policy Gradient-based reinforcement learning algorithms: **Policy Gradient (PG)**, **Proximal Policy Optimization (PPO)**, and **Trust Region Policy Optimization (TRPO)**. Each algorithm was trained across five benchmark continuous and discrete control environments: *Acrobot-v1*, *Ant-v5*, *HalfCheetah-v5*, *LunarLander-v3*, and *Reacher-v5*. The performance metric used is the average episodic reward over training timesteps.

Training Configuration

Each algorithm was trained for a total of **50 iterations**, where each iteration consisted of approximately **10,000 environment timesteps**. During each iteration, the agent interacted with the environment over multiple episodes, collecting experience tuples (s, a, r, m) until the timestep limit was reached.

The average episodic reward was computed after each iteration and stored for performance evaluation. Both the actor and critic networks were updated once per iteration using the collected batch of trajectories. In total, each training run comprised around **500,000 environment interactions**.

Results and Analysis

The following plots visualize the learning progress for each environment. Each curve represents the average performance over training iterations.

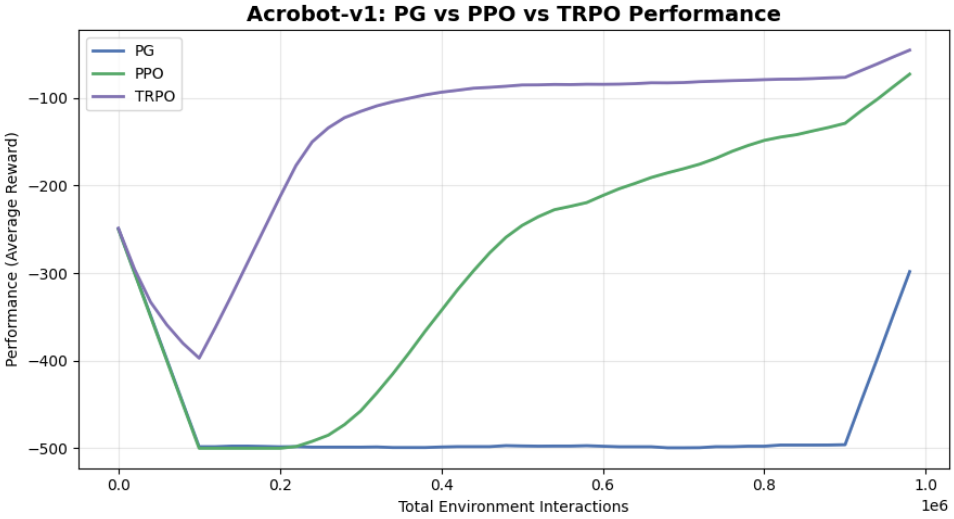

Figure 1: Performance comparison on Acrobot-v1.

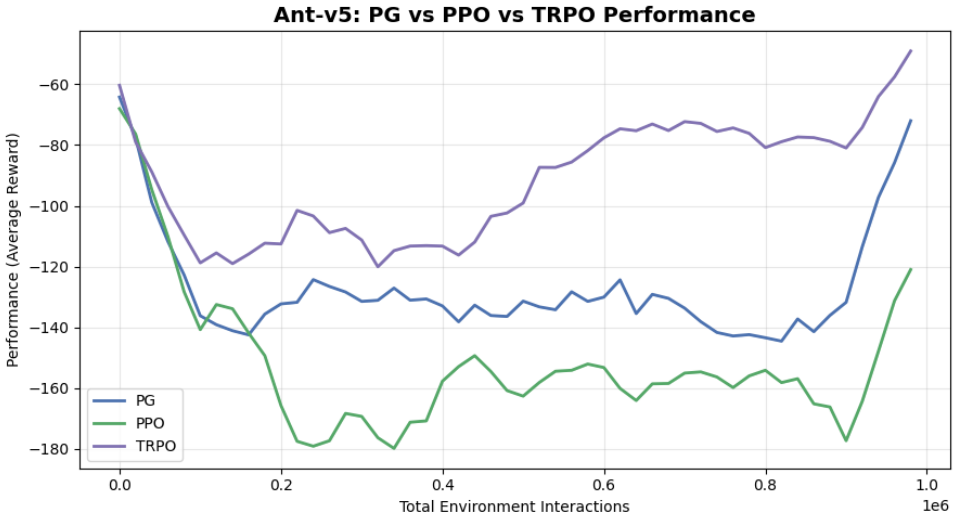

Figure 2: Performance comparison on Ant-v5.

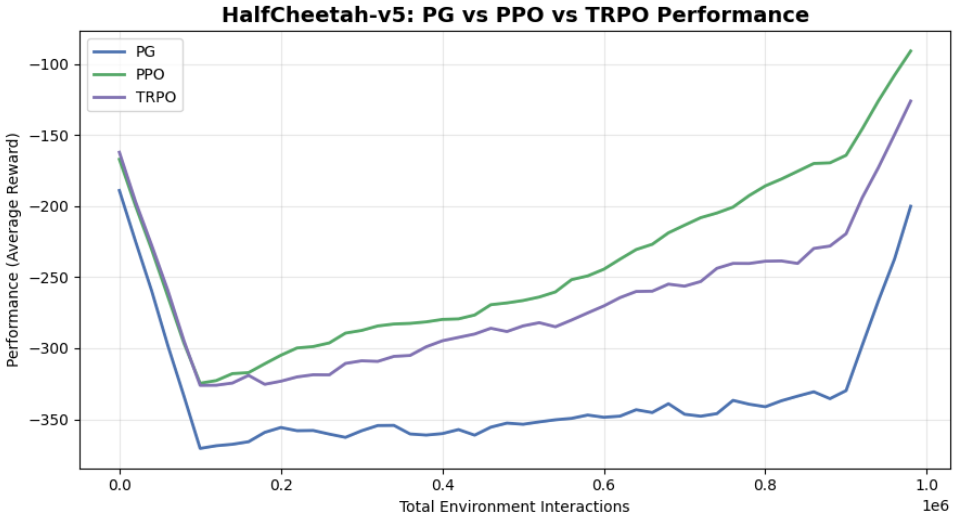

Figure 3: Performance comparison on HalfCheetah-v5.

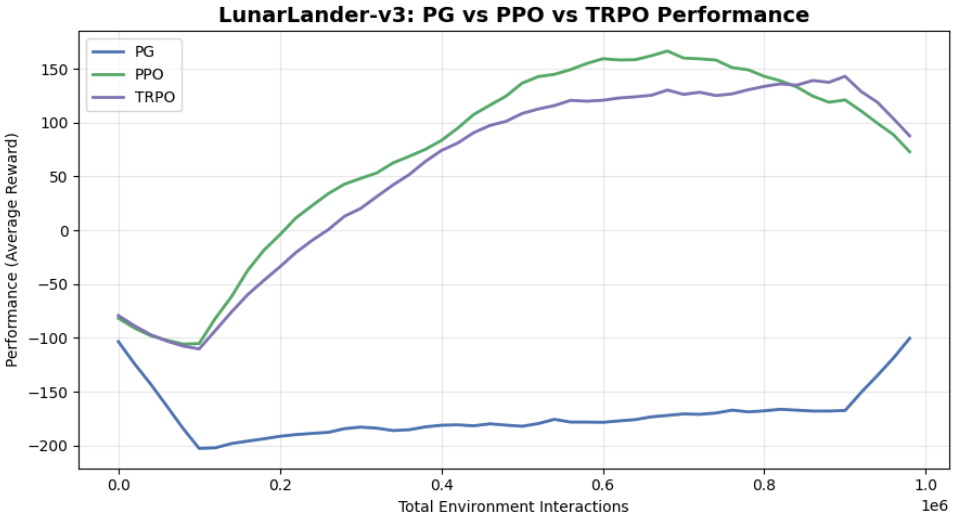

Figure 4: Performance comparison on LunarLander-v3.

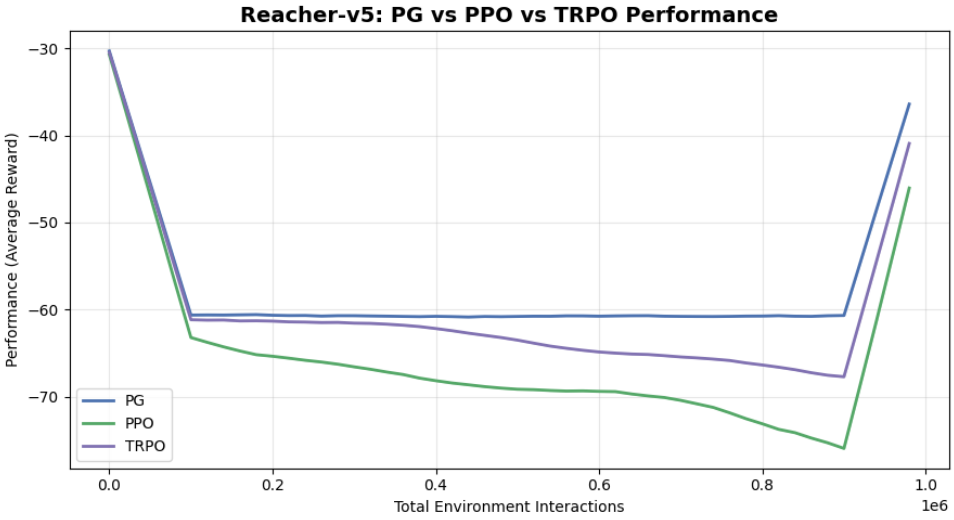

Figure 5: Performance comparison on Reacher-v5.