Project Proposal: PPO on Atari Pong

Idea -

The PPO algorithm is a standard on Atari, but highly sensitive to small implementation adjustments or details. I will reproduce the standard PPO algorithm on ALE/Pong-v5 using Stable-Baseline3 as a baseline. Following that I will implement my own rendition of the PPO algorithm, modifying anywhere from 3-5 features from the original algorithm, and assess sample efficiency along with stability. I will also visualize the training diagnostics with heat maps to visualize policy/critic behavior over the game states for both implementations of the algorithm.

Environment and Baseline -

- Environment: gymnasium.make("ALE/Pong-v5")
- Baseline: SB3 PPO with default Atari design and RL Zoo references to use as the performance anchor.

My PPO Rendition -

I'll start by following the "37 Implementation Details of PPO" to ensure correctness, then change a small set of details to test learning speed and stability. As for now these are some options for the features of PPO that I am considering modifying, but they are subject to change: clip range schedule, target-kl early stop, GAE, entropy bonus schedule, and value function loss settings.

Evaluation -

- Metrics: mean episodic return vs environment steps. I'll be tracking sample-efficiency with the number of steps to reach a return >= 18, as well as stability/variance across seeds.
- Protocol: 5-10 random seeds, report +- 95% confidence interval, significance tests following: https://arxiv.org/abs/1709.06560?utm_source=chatgpt.com
- Heat Map: this will visualize where and when the agent is confident/uncertain and which actions dominate.

Heat Map Visualization -

- Action usage: over (ball_y - paddle_y) bins across episodes which will show paddle preference relative to ball geometry.
- Value(V(s)): take the course grid and accumulate average predicted V(s) per cell. This will reveal what parts of the course grid correlate with high/low value.
- Policy-entropy: will use the course grid to highlight uncertainty regions, this will help me understand areas where the algorithm is favoring exploitation or exploration.

Summary -

Upon following the above project proposal I believe I will meet all of the learning objectives outlined in the project instructions.