5180: RL Project Proposal

Sreehari Premkumar

Problem Description

This project aims to develop and evaluate a Soft Actor-Critic (SAC) algorithm in OpenAI Gym's Box2D BipedalWalker-v3 environment. The environment simulates a four-joint bipedal walker that must navigate a terrain, where the goal is to maximize rewards by moving forward efficiently while using as little energy as possible through good motor control. This presents a continuous control problem that needs dynamic stability from the walker's motorized joints.

- **States**: The state space consists of a 24-dimensional vector that includes hull angle speed, angular velocity, horizontal and vertical speeds, joint angles, joint angular velocities, ground contact information, and lidar measurements for terrain awareness.
- **Actions**: The action space is represented as a continuous vector of size four, with values ranging from [-1, 1], which correspond to the motor speeds for the walker's hips and knees.
- **Rewards**: The agent receives rewards for moving forward while losing points for falling (-100 points) or using too much torque. The goal is to achieve a cumulative score of at least 300 points within 1600 time steps in the normal version.

My motivation for this project comes from my interest in robotics, especially in learning about how humanoids walk. I see the Bipedal Walker as a valuable step towards this goal.

Algorithms

The primary algorithm used in this project will be Soft Actor-Critic (SAC), a model-free, off-policy reinforcement learning method that is well-suited for continuous action spaces like that of the Bipedal Walker. SAC balances exploration and exploitation through maximizing entropy, making it strong in complex environments.

- **SAC Advantages**: Known for its efficiency in samples and stability in training, SAC is especially helpful for this project since I will be running experiments on CPU resources. The algorithm's use of an entropy term encourages exploration, which is important for avoiding getting stuck in suboptimal policies.
- **Algorithm Justification**: SAC has been successfully used in various continuous control tasks, often showing top performance. Its ability to explore is especially helpful given the challenges of the Bipedal Walker environment.

Results

The expected outcome is to evaluate the performance of SAC on the BipedalWalker-v3 environment, aiming to achieve scores close to or above the 300-point mark over multiple trials. The results will include:

- **Learning Curves**: Tracking cumulative rewards over episodes to show SAC's learning progress, with confidence bands to indicate variability across different training runs.
- **Reward Component Analysis**: Analyzing the breakdown of rewards, distinguishing between rewards for moving forward and penalties from using too much motor torque.
- **Parameter Sensitivity**: Investigating how performance changes with different settings for entropy scaling, learning rates, and replay buffer size.

Risks and Contingencies: Running the project without a GPU may limit the number of episodes I can complete in the given timeframe. If training does not converge due to computational limits, I will consider changes to the SAC setup that improve efficiency, like adjusting batch sizes or focusing on more recent samples in the replay buffer.

Future Work: If successful, I plan to extend this approach to the more challenging MuJoCo (3D) humanoid walking environment to further test SAC’s performance in complex control situations.