Reinforcement Learning Final Project

(Winter 2019-2020)

Continuous Control Task

Policy Gradient Algorithms

REINFORCE [D, C]	PPO [D,C]	DDPG [C]

D: Discrete Actions
C: Continuous Actions

D2C:

Discrete to Continuous Actions
For **REINFORCE** and **PPO**

Progress

REINFORCE

- Monte Carlo Rewards
- Advantage for Variance
- Handles Continuous Tasks
- on-policy, off-line

PPO

- MC Rewards [> single episodes]
- Advantage for Variance *V(s)*
- Trust Region
- Entropy based Exploration
- off-policy, on-line

DDPG

- TD Critic + Experience Replay
- Target networks for Variance
- Deterministic Actor + Noise
- off-policy, on-line

Sparse Reward

Sparse Reward

- Requires high exploration
- High variance
- Learning may never start!

Carrots & Sticks

- Reward and punishment are functions of angle
- Learning will plateau quickly if more reward than punishment. Will lead to quick rotations.

Slow Rotation

- Punish high angular & cart velocity at desired regions
- Coefficient will dictate the behavior (Smooth vs Erratic)
- Combination of continuous and step-wise rewards

HPO pipeline

Pipeline

- ConfigSpace library
- Google Drive + Colab
- Initially large parameters search space
- Manually gaging the performance based on reward plots
- Gradually decreasing the number of relevant parameters and their search space

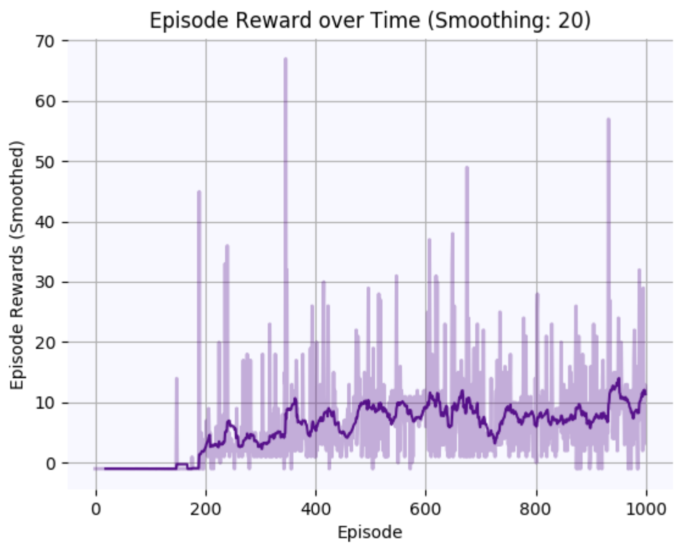
Insights

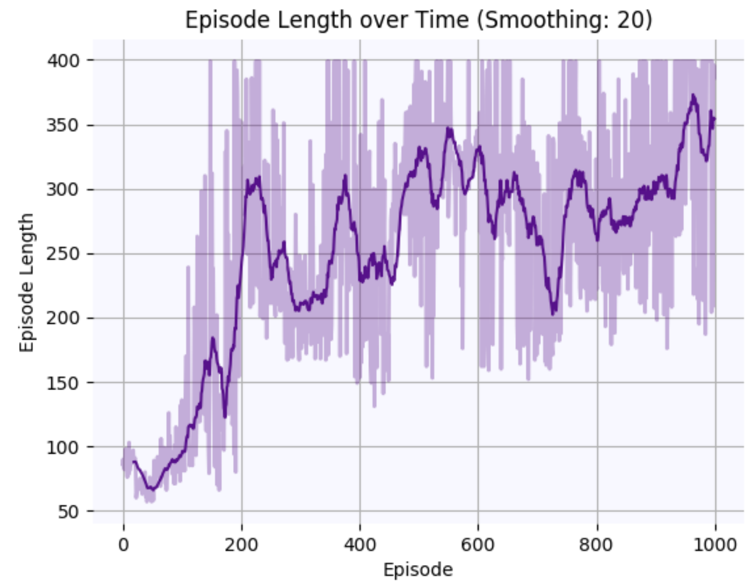
- Different algorithms (very) different hyper parameters
- Hidden dimension of the network was very important
- Pipeline was used for selecting between reward functions as well as fine tuning their coefficients
- Using ADAM optimizer reduced the impact of LR

Group: Mehran Ahkami, Seyed Mahdi Basiri Azad, Victor Vadakechirayath George

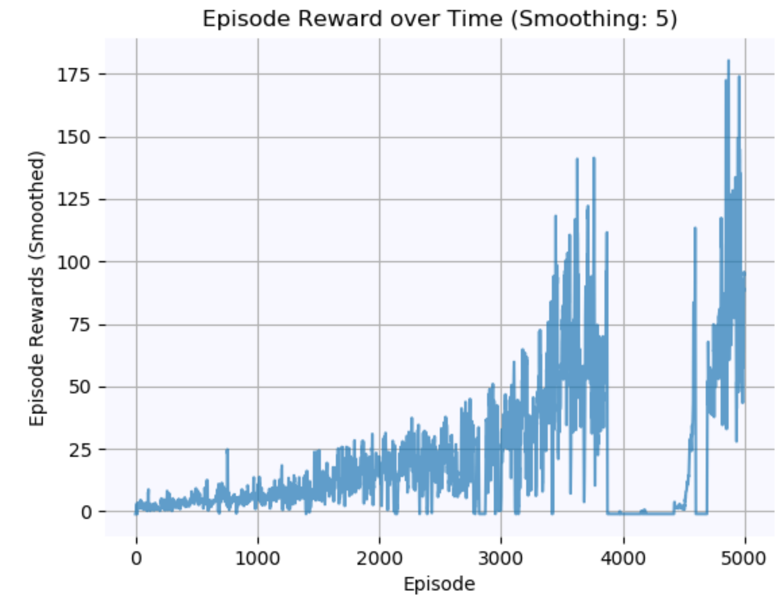
Sample Plots

PPO: High Consistency for Actions



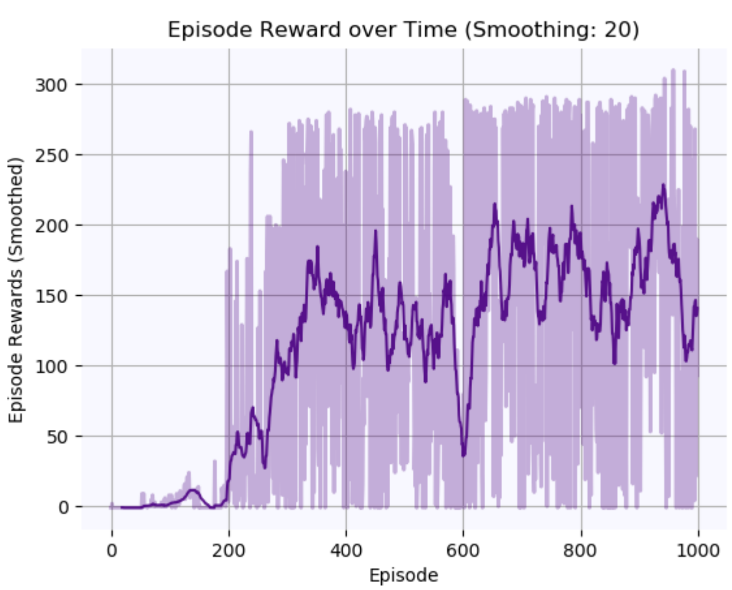


REINFORCE: Sparse Reward (Only Once!)

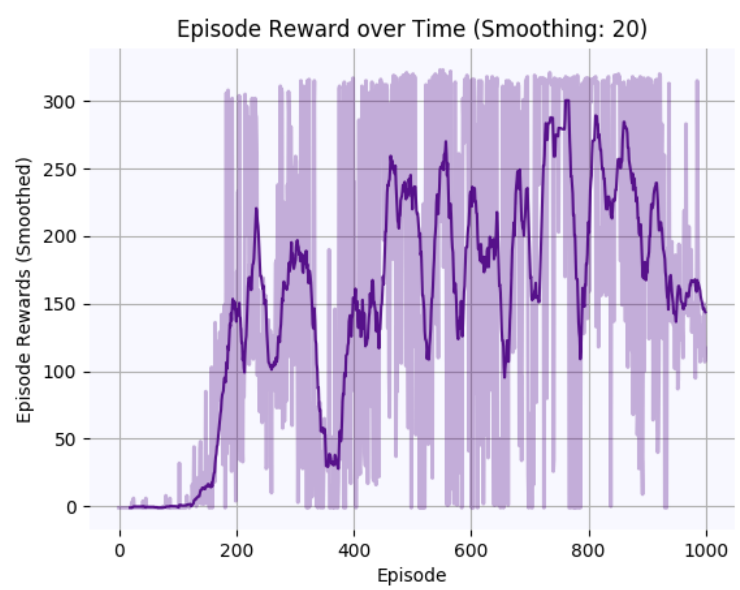


DDPG: Effect of Normal Noise on Exploration

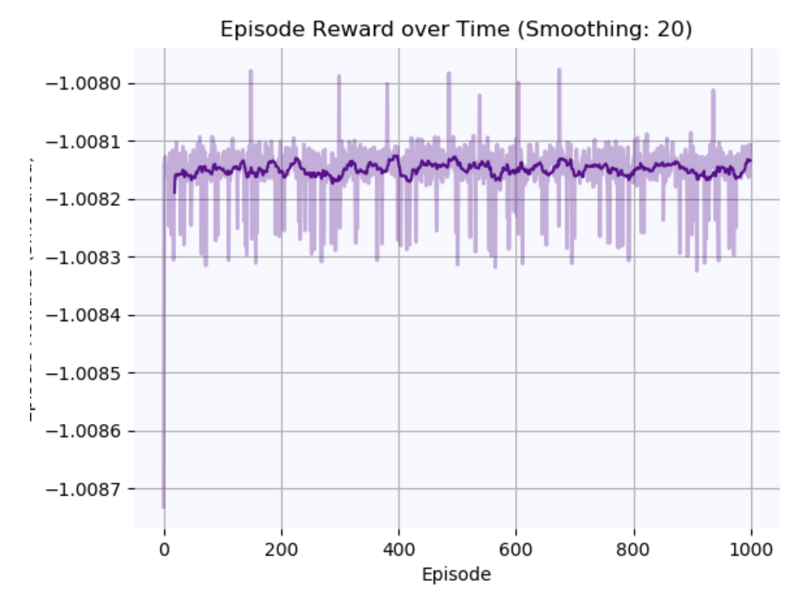
Noise Std: 1.0



Noise Std: 0.5



Noise Std: 0.01



Group: Mehran Ahkami, Seyed Mahdi Basiri Azad, Victor Vadakechirayath George