Project Report

Learning Algorithm

The following algorithms was used as part of the solution:

1 – PPO – Proximal Policy Optmization: https://arxiv.org/abs/1707.06347

2 – GAE - Generalized Advantage Estimation: https://arxiv.org/abs/1506.02438

3 – Based on Advantage-Actor-Critic methods

The architecture used was:

A policy with an actor and critic network as follows:

```
----------------
Actor:
FCNetwork(
  (linear1): Linear(in_features=33, out_features=500, bias=True)
  (linear2): Linear(in_features=500, out_features=250, bias=True)
  (linear3): Linear(in_features=250, out_features=4, bias=True)
)
----------------
Critic:
FCNetwork(
  (linear1): Linear(in_features=33, out_features=500, bias=True)
  (linear2): Linear(in_features=500, out_features=250, bias=True)
  (linear3): Linear(in_features=250, out_features=1, bias=True)
)
----------------
```

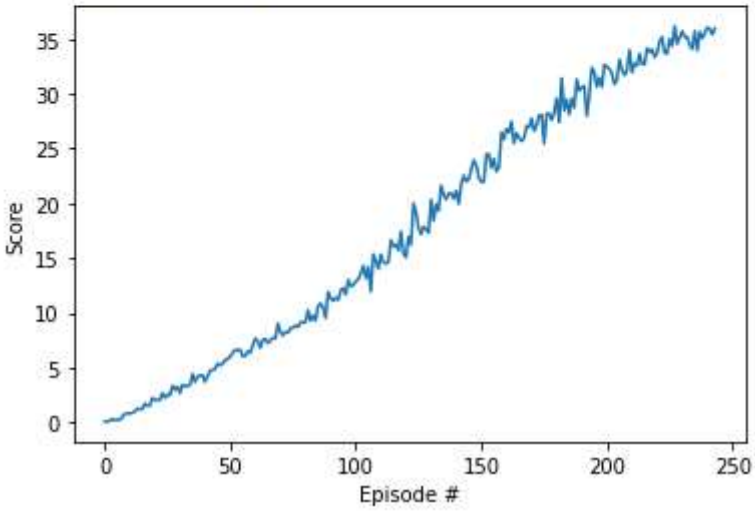
Hyperparameters

LR = 1.0e-4
SGD_EPOCH = 4
DISCOUNT = 1
GAE_LAMBDA = 0.95
BATCH_SIZE = 64
EPSILON_PPO_CLIPPING = 0.1

Plot of Rewards

```
Episode 0       Average Score: 0.10
Episode 20      Average Score: 0.91
Episode 40      Average Score: 2.04
Episode 60      Average Score: 3.31
Episode 80      Average Score: 4.51
Episode 100     Average Score: 5.89
Episode 120     Average Score: 8.66
Episode 140     Average Score: 11.86
Episode 160     Average Score: 15.35
Episode 180     Average Score: 19.13
Episode 200     Average Score: 22.95
Episode 220     Average Score: 26.54
Episode 240     Average Score: 29.69

Environment solved in 143 episodes!
```



Ideas for future work

1– Implement “DDPG: Deep Deterministic Policy Gradient”