Learning algorithm

The algorithm used to solve this environment is Deep Deterministic Policy Gradient (DDPG), an extension of DQN to continuous action spaces on an actor-critic manner. The main difference is that instead of just one network that outputs the value for every action, we have two: the actor, a maximizer that outputs the action with the biggest value, and the critic, that takes this action and a state and outputs the value.

To update the value network, the critic, we use an target actor and critic (fixed parameters) to create an Q used as target to train the learning critic.

Then, we use our updated critic to calculate a value with which we will do gradient ascent on the actor.

Also, an Ornstein-Unlenbeck process is implemented as noise generator to favor exploration.

Hyperparameters

```
BUFFER_SIZE = int(5e5)  # replay buffer size
BATCH_SIZE = 256        # minibatch size
GAMMA = 0.99            # discount factor
TAU = 1e-3              # for soft update of target parameters
LR_ACTOR = 2e-4         # learning rate of the actor
LR_CRITIC = 2e-4        # learning rate of the critic
WEIGHT_DECAY = 0.0      # L2 weight decay
OU_MU = 0.1
OU_SIGMA = 0.1
OU_THETA = 0.15
```

The hyperparameters selected are variations from the original DDPG paper, adapted by trial and error. The original values of the changed parameters are:

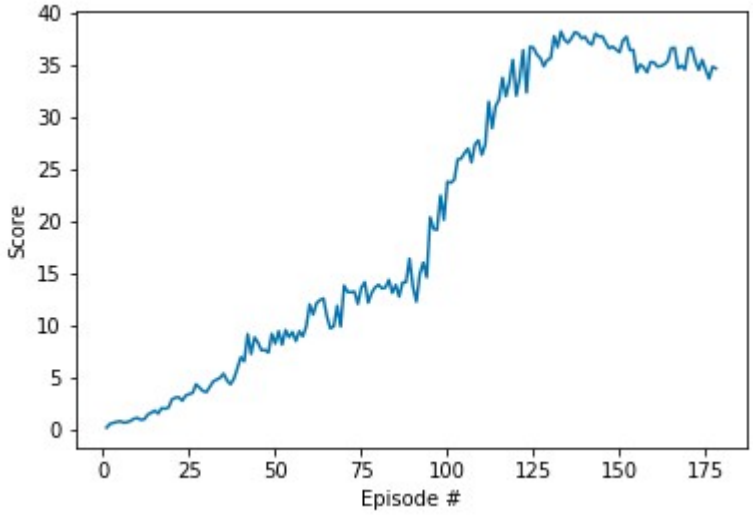
- Buffer size :1e5
- Batch size: 64
- Actor learning rate: 1e-4
- Critic learning rate: 1e-3
- Weigth decay: 1e-2.
- OU mu: 0
- OU theta: 0.1

Model architectures

Both are neural networks with 2 hidden with 400 and 300 with batch normalization on the first hidden layer and ReLU as activation function. For the critic, the actions come in in the second hidden layer.

Plot of rewards

Objective: The agent is able to receive an average reward (over 100 episodes, and over all 20 agents) of at least +30.




```
Episode 10      Average Score: 0.74     Time:2.15 min..
Episode 20      Average Score: 1.25     Time:4.51 min..
Episode 30      Average Score: 2.00     Time:7.18 min..
Episode 40      Average Score: 2.77     Time:9.90 min..
Episode 50      Average Score: 3.82     Time:12.62 min.
Episode 60      Average Score: 4.75     Time:15.35 min.
Episode 70      Average Score: 5.71     Time:18.07 min.
Episode 80      Average Score: 6.65     Time:20.79 min.
Episode 90      Average Score: 7.47     Time:23.52 min.
Episode 100     Average Score: 8.55     Time:26.28 min.
Episode 110     Average Score: 11.08    Time:29.03 min.
Episode 120     Average Score: 14.08    Time:31.79 min.
Episode 130     Average Score: 17.26    Time:34.56 min.
Episode 140     Average Score: 20.52    Time:37.31 min.
Episode 150     Average Score: 23.42    Time:40.07 min.
Episode 160     Average Score: 26.05    Time:42.84 min.
Episode 170     Average Score: 28.45    Time:45.61 min.
Episode 178     Average Score: 30.20    Time:16.17 sec.Environment solved in 178 episodes!      Average Score: 30.2
0       Time:2867.58
```

Ideas for future work

- Prioritized replay experiences
- Using Generalized Advantage Estimation.
- Using on-policy training by not using the replay buffer.
- Try other algorithms: A3C, A2C, TRPO, TNPG, D4PG.