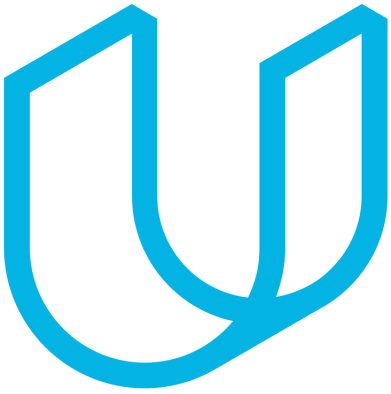
# Continuous control Project
## Follow a moving target

Jean-Baptiste Gheeraert



UDACITY

DEEP REINFORCEMENT LEARNING NANODEGREE
UDACITY

August 28, 2019

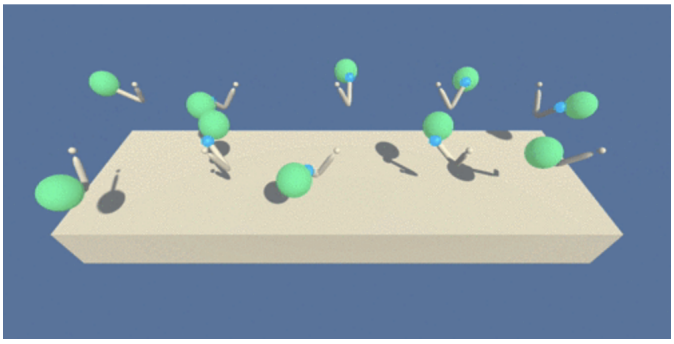
# Table des matières

# Chapitre 1

# Project description

## 1.1 Environment

In this project an agent (or several similar agent) aims to follow a target. A reward of +1 is provided for each step that the agent's hand is in the goal location. Thus, the goal of your agent is to maintain its position at the target location for as many time steps as possible.



The environment space is defined by 33 variables by agent (position, rotation, velocity, and angular velocities of the arm) and the action space contains 4 numbers corresponding to torque applicable to two joints.

## 1.2 Learning algorithm

The algorithm used here is a Deep Deterministic Policy Gradient (DDPG) [2]. A DDPG is composed of two networks : one actor and one critic.

During a step, the actor is used to estimate the best action, ie $argmax_a Q(s, a)$ ; the critic then use this value as in a DDQN to evaluate the optimal action value function.
Both of the actor and the critic are composed of two networks. On local network and one target network. This is for computation reason : during backpropagation if the same model was used to compute the target value and the prediction, it would lead to computational difficulty.
During the training, the actor is updated by applying the chain rule to the expected return from the start distribution. The critic is updated as in Q-learning, ie it compares the expected return of the current state to the sum of the reward of the choosen action + the expected return of the next state.

The first structure tried was the one from the ddpg-pendulum project of the nanodegree (with few modifications) and it gave very good results. Few things had to be adapted : the step function as we know have simulteanously 20 agents that return experiences and the noise as we have to apply a different noise to every agent (at first I did not change the noise, the training was running but the agent did not learn anything).

The actor is composed of 3 fc units :
- First layer : input size = 33 and output size = 128
- Second layer : input size = 128 and output size = 128
- Third layer : input size = 128 and output size = 4

The critic is composed of 3 fc units :

- First layer : input size = 33 and output size = 128
- Second layer : input size = 134 and output size = 128
- Third layer : input size = 128 and output size = 1

The second layer takes as input the output of the first layer concatenated with the choosen actions.

The training hyperparameters are as follow :

- Buffer size : 100,000
- Batch size : 128
- $\gamma$ : 0.99
- $\tau$ : 0.001
- learning rate actor : 0.001
- learning rate critic : 0.001
- weight decay : 0

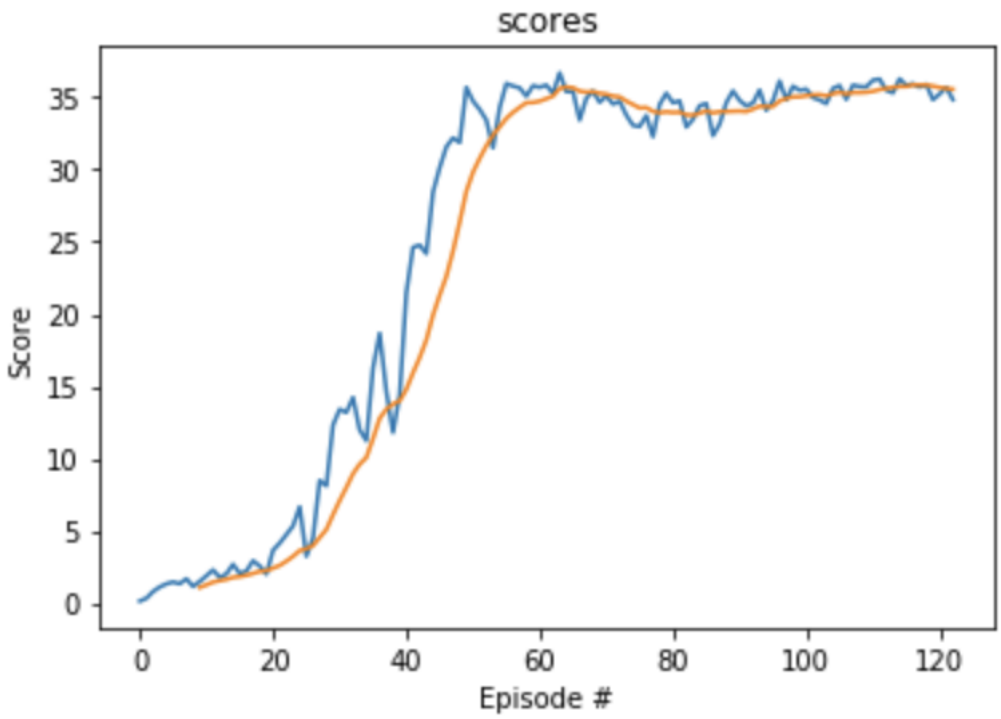
# Chapitre 2

# Plot of rewards

The environment has been solved in 23 episodes.

```
Episode 100     Average Score: 22.49
Episode 123     Average Score: 30.21
Environment solved in 23 episodes!      Average Score: 30.21
```

Here is the graph of the score evolution :



# Chapitre 3

# Ideas of future works

To improve the stability of the model, batch normalisation could be added. Batch normalisation permits obtaining a better gradient flow, insuring that the distribution of the inputs to a given layer during the training always follow a gaussian law.
We also could implement other model such as the ones seen in the course : PPO [4], A3C [3] or D4PG [1].

# Bibliographie


[1] Gabriel Barth-Maron et al. "Distributed Distributional Deterministic Policy Gradients". In : *CoRR* abs/1804.08617 (2018). arXiv : 1804.08617. URL : http://arxiv.org/abs/1804.08617.

[2] Timothy P Lillicrap et al. "Continuous control with deep reinforcement learning". In : *arXiv preprint arXiv :1509.02971* (2015).

[3] Volodymyr Mnih et al. "Asynchronous Methods for Deep Reinforcement Learning". In : *CoRR* abs/1602.01783 (2016). arXiv : 1602.01783. URL : http://arxiv.org/abs/1602.01783.

[4] John Schulman et al. "Proximal Policy Optimization Algorithms". In : *CoRR* abs/1707.06347 (2017). arXiv : 1707.06347. URL : http://arxiv.org/abs/1707.06347.