COMPETITIVE TASK - REPORT

The agents are trained to play tennis via DDPG, an algorithm derived from DQN optimized for continuous action spaces.

If an agent hits the ball over the net, it receives a reward of +0.1. If an agent lets a ball hit the ground or hits the ball out of bounds, it receives a reward of -0.01.

Training lasts as long as the score performed by the winner agent (the average over 100 episodes) is not higher than 0.50; in any case, training is stopped if the desired score is not achieved in 10000 episodes.

Each episode consists of 1000 steps at most.

For each step, every agent picks the *best action* predicted for its current state according to a collective model and actuates it (one action is a tuple of torque values). More details are below.

The action is followed by 'reaction' of the environment expressed by the next state (new positions/velocities) and one reward (positive in new position is in the target location)

The SARS sequences (current state, action, reward and next state) collected by all agents is hence 'memorized' in a collective memory buffer (the buffer can store at most 10.000 sequences – buffer size) and the 'experience' of 128 sequences (batch size) randomly picked from the buffer is used to train the collective model.

The model basically consists of two deep neural networks: one - the 'actor' - is trained to estimate the *best action* in a certain state, the other – the 'critic' - is trained to estimate the value (future expected rewards discounted with a Gamma factor of 0.99) of the action suggested by actor; in turn, this value is used to train the actor in its estimate.

Two additional clones of the networks (target networks) slowly track the previous ones (the parameters are slightly shifted in their direction, 1‰ at each iteration - Tau parameter) and are employed to the improve convergence properties of the training algorithm.

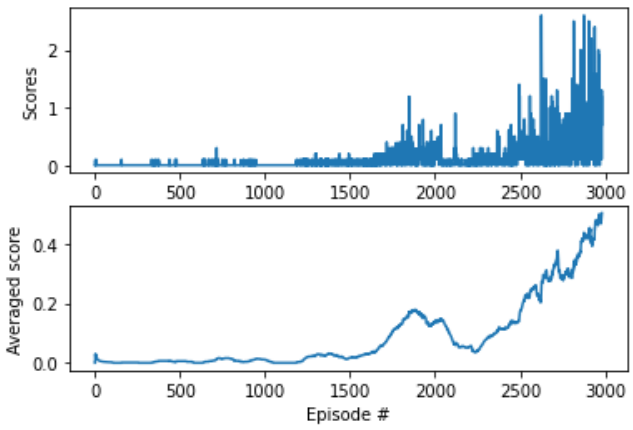
In details, the actor network takes in state vectors and takes out action vectors. It has two fully connected hidden layers with 256 and 128 units respectively and all of them are rectified non-linearities. The units in the output layer consist of hyperbolic tangent activation functions, suitable to bound actions between -1 and +1. Note that, the action takes out by the actor network is not directly applied by the agent but is 'flavored' with Ornstein-Uhlenbeck noise, useful to generate temporally correlated exploration for exploration efficiency in physical control problems (details in DDPG paper), and eventually 'clipped' to stay between -1 and 1 event once noise is introduced.

The critic network takes in state vectors and action vectors and takes out the Q-value (action value in the state). It has two fully connected hidden layers with 256 and 128 units respectively and all of them are rectified non-linearities. Actions are not included until the second hidden layer of the network.

The final layer weights and biases of both the actor and critic were initialized from a uniform distribution [−3e-3, +3e-3] to ensure the initial outputs for the policy and value estimates were near zero. The hidden layers were initialized from uniform distributions $[-\frac{1}{\sqrt{f}}, +\frac{1}{\sqrt{f}}]$ where f is the fan-in of the layer.
Experimental settings found in DDPG paper were replicated here.

The training consists of stochastic gradient descent (Adam optimizer) and back-propagation: learning rate is 1e-4 for the actor, 4e-4 for the critic; the weight decay is always 0.

With the described algorithm and hyperparameters, the desired score - the average reward of the best player on 100 episodes - has been achieved in less than 3000 episodes:



Future work may be dedicated to improving performances by testing other algorithms (like D4PG).

Further tests may be run to check also the DDPG performances after excluding velocity variables from the observation vectors: indeed, such information may be redundant since the three consecutive time stamps stacked in the observation vector implicitly enable the possibility to infer directions and velocity of racket and ball.