**RESULTS OF CARTPOLE PROBLEM BUILT ON RL-STUDIO**

**Starting point:**

alpha: 0.7  
epsilon: 0.99  
gamma: 0.9  
epsilon\_discount: 0.9999999

**GOAL** → Open AI consider this environment solved when an average over the last 100 episodes of 195 is reached

* Reward = 1 if in the “center zone”, 0 otherwise
* Steps per run = 1000
* states

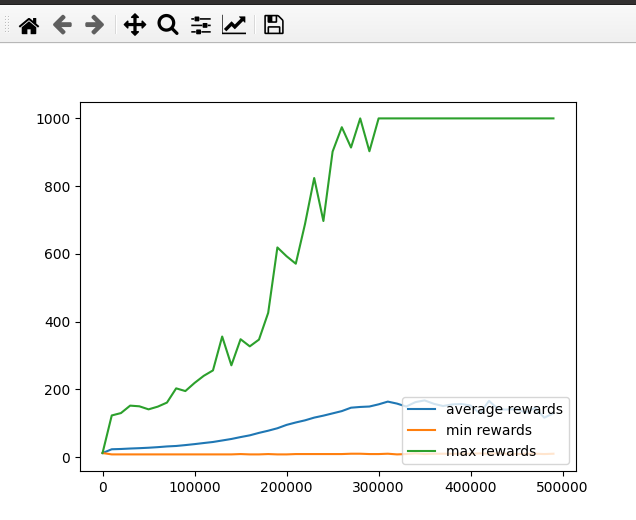
*| 0 | Cart Position | -4.8 | 4.8 |*  
*| 1 | Cart Velocity | -Inf | Inf |*  
*| 2 | Pole Angle | ~ -0.418 rad (-24°) | ~ 0.418 rad (24°) |*  
*| 3 | Pole Angular Velocity | -Inf | Inf |*

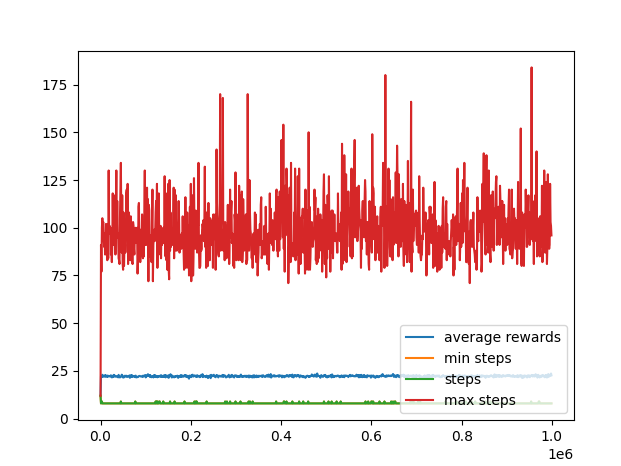
* Actions

*| 0 | Push cart to the left |*  
*| 1 | Push cart to the right |*

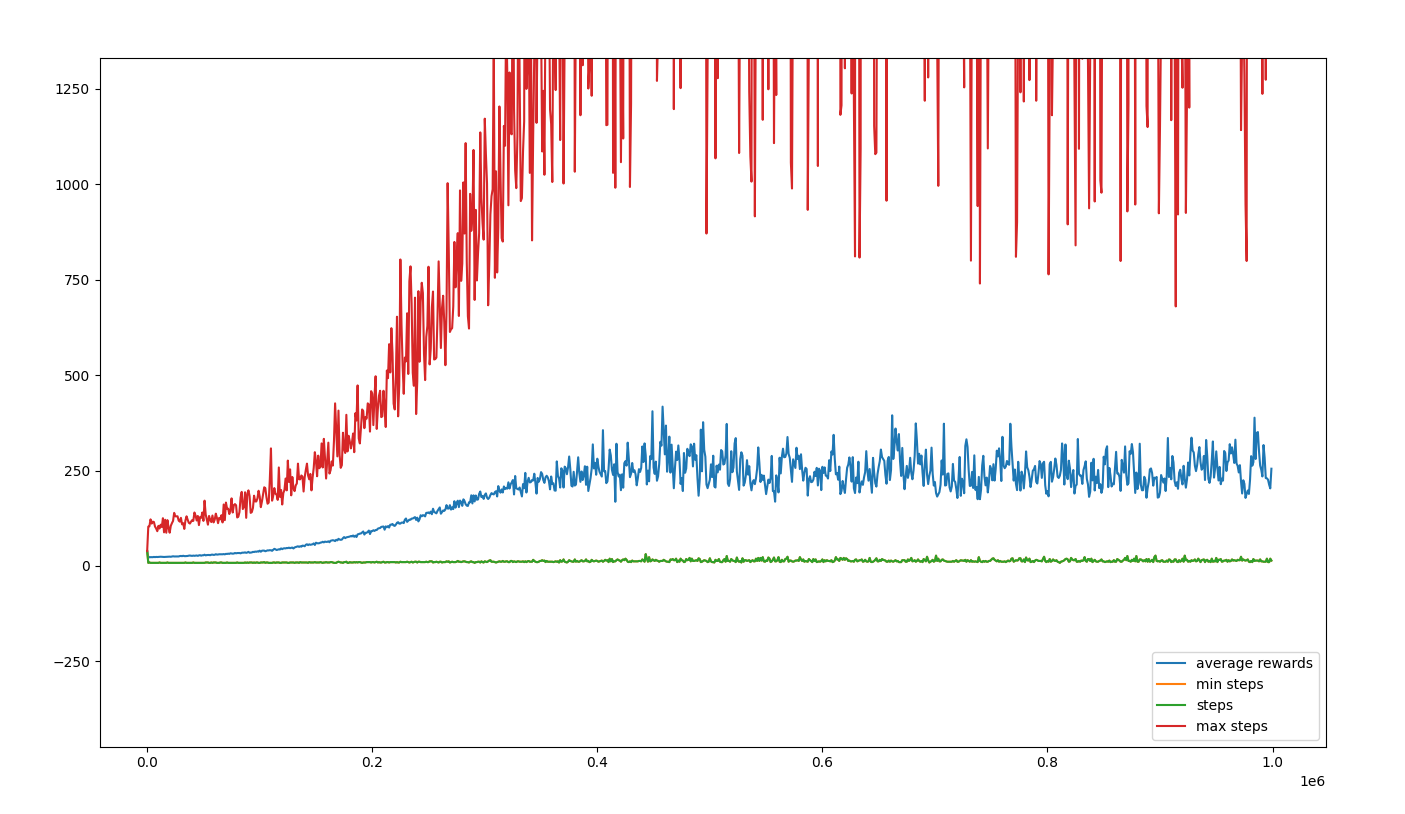
**Process**

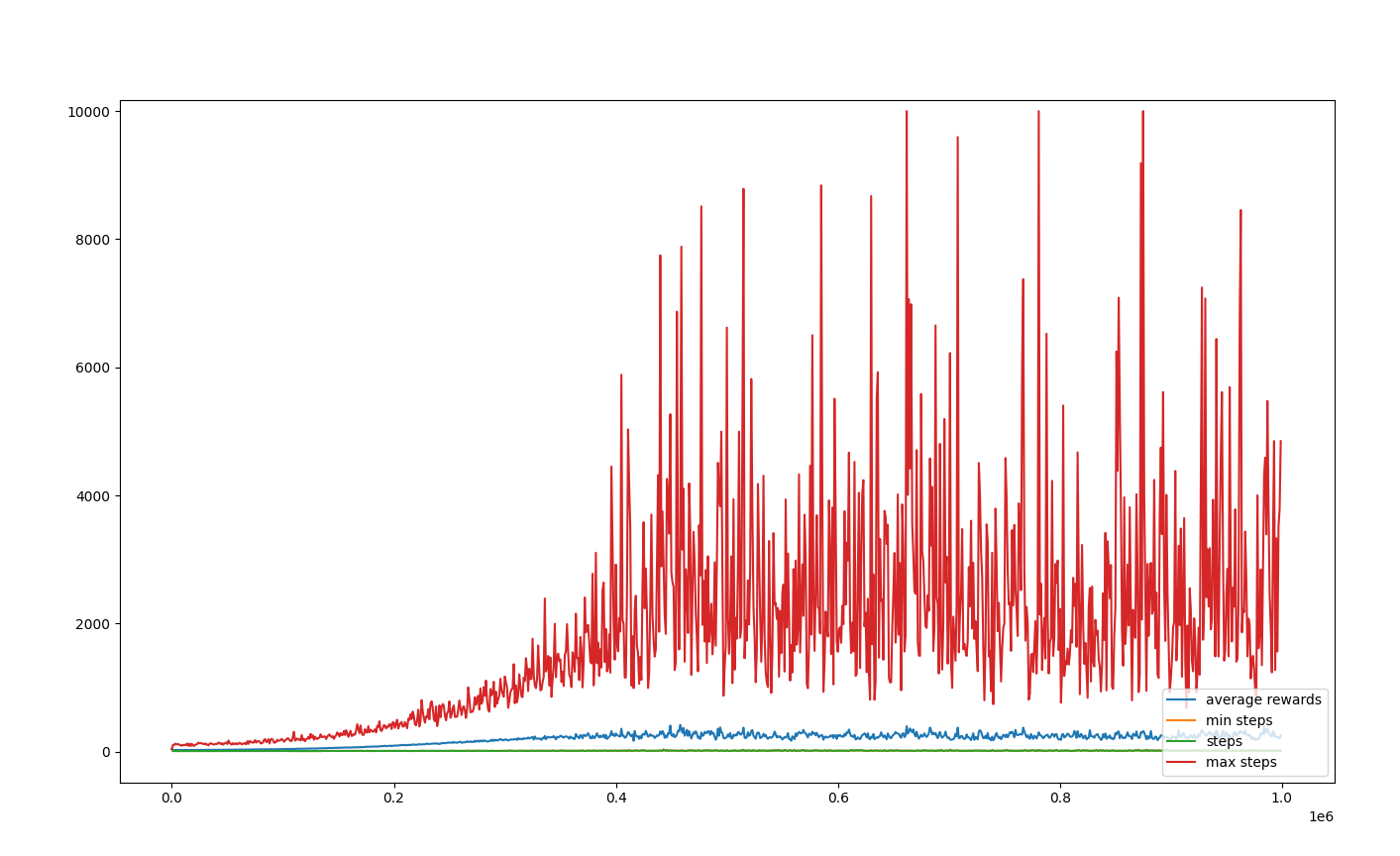
* First we observe that each ephoc (1000 episodes) we get some succesfull episodes in which the pole is kept vertical.The problem is that the average stop increasing and we also obtain always really bad episodes each epoch.

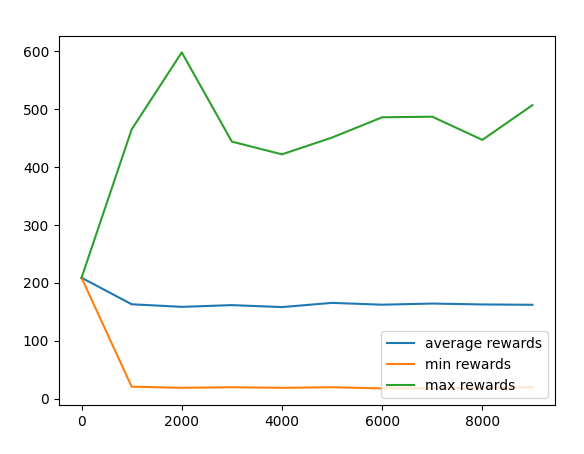


* We realized that the pole was not retrieved when it left the “center zone” and that it was not reacting enough soon to the pole falling, so we customized the reward in this way:
  + Reward = 1 - (|x\_pos| (0 is the center) + | angle |)
* It didnt worked, so we went back to the experiment in which we just reward the pole being up. This time, the approach was lighly d bgnt. We rewarded 0 when the step didnt fall the pole and -e when the gpisode finished. The attempt is to just punish failures which retrvvvbkogate to the previous wrong decissions. Additionally we increased to 10000 the maximum number of steps just to check what the agent is able to achieve.
* It worked even worse, so we decided to go back to positive rewards. This time we give a hint to the agent that the center is good rewarding him a little bit wen it is close to the center and keeping the negative reward when it falls.
  + Reward = 1 if it is up, additionally 0.1 if it is close to the center and -1 if it falls.
  + Hyperparameters:

alpha: 0.9  
epsilon: 0.99  
gamma: 0.9  
epsilon\_discount: 0.9999999





When applying this model in inference mode:

* Not bad, but still far from surpassing the 200 steps, so we started playing with the hyperparameters and making the trainment longer. It is what we obtained:
  + Hyperparameters:

alpha: 0.85  
epsilon: 0.99  
gamma: 0.95  
epsilon\_discount: 0.99999999

* + reward function back to just 1 when still running and 0 when pole felt (to not to confuse the agent)