# Lab Report 2 Submission (2019)

Sylvain Lapeyrade, Reda Bourakkadi (sylla801, redbo196)

December 9, 2019

## Answers

### 1) Exercise 1: Trade-off Between Exploration and Exploitation

*Evaluate the effects of $\alpha$, $\gamma$ and $\epsilon$, and plot your accumulated reward for your best set of values. To get a good result you will need to update $\epsilon$ from a large to a small value during training. Study the values of the Q table for your best solution. What are the major difficulties for learning in this environment?*

By trying different values for $\alpha$, $\gamma$ and $\epsilon$, we found the best results with a value of 0.9 for each. As suggested, with have decreased $\epsilon$ from the initial 0.9 by 0.001 after each episode, until 0.

While we manage to have total reward of more than 0.5 constantly as instructed, our accumulated reward for our best set of value is 0.6661 and the corresponding best values are as in Table 1:

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 3 | 3 | 3 | 0 | 0 | 2 | 0 | 3 | 1 | 0 | 0 | 0 | 2 | 1 | 0 |

Table 1: Best values of the Q-tables of our best cumulated reward.

They are corresponding to the best action as evaluated by the algorithm. *Left* is designated by 0, *Down* by 1, *Right* by 2 and *Up* by 3. The Q-Table values are on in Figure 1. Indeed, since we took the epsilon greedy approach, where we randomly generate value between 0 and 1 and then we see if they are smaller than /*epsilon*. If so, a random action between the four mentioned earlier is chosen. Otherwise, we choose the action with the maximum value in the Q-table.

The major difficulties for learning in this environment is to find adequate values for $\alpha$, $\gamma$ and $\epsilon$ and also figure out how much to decrease $\epsilon$ on each episode. This is very important, because it determined how long we want to **explore** (i.e. randomly select actions to try new possibilities) which correspond to a big $\epsilon$ value, and how long to **exploit** (i.e. take the best action already found to pursue the best set of actions) which is a small $\epsilon$ value.

| Left | Down | Right | Up |
|---|---|---|---|
| 2.24034053e-02 | 2.05161925e-04 | 2.24713243e-04 | 2.15449416e-04 |
| 9.50456411e-06 | 4.06757021e-05 | 1.08813765e-05 | 4.65207716e-02 |
| 2.25865266e-05 | 1.78173882e-05 | 1.82306304e-05 | 1.63852399e-02 |
| 1.78228153e-05 | 6.52572124e-06 | 4.44221807e-06 | 1.76157626e-02 |
| 4.32316778e-02 | 2.69057491e-05 | 9.44212465e-05 | 3.66241274e-05 |
| 0.00000000e+00 | 0.00000000e+00 | 0.00000000e+00 | 0.00000000e+00 |
| 2.71115287e-08 | 8.88247889e-09 | 5.20025762e-02 | 2.37693514e-08 |
| 0.00000000e+00 | 0.00000000e+00 | 0.00000000e+00 | 0.00000000e+00 |
| 8.80245540e-05 | 6.97456017e-05 | 5.46894576e-05 | 4.40332313e-02 |
| 3.06572962e-05 | 7.03347894e-02 | 2.23675336e-05 | 1.03796446e-05 |
| 5.17980659e-02 | 1.69806541e-06 | 3.65062105e-06 | 4.20391920e-06 |
| 0.00000000e+00 | 0.00000000e+00 | 0.00000000e+00 | 0.00000000e+00 |
| 0.00000000e+00 | 0.00000000e+00 | 0.00000000e+00 | 0.00000000e+00 |
| 3.69007597e-03 | 8.46520530e-04 | 2.50595263e-01 | 5.88882913e-03 |
| 1.65424746e-02 | 9.35429600e-01 | 1.87956221e-02 | 1.74856936e-02 |
| 0.00000000e+00 | 0.00000000e+00 | 0.00000000e+00 | 0.00000000e+00 |

Table 2: The direction and their respective rewards.

## 2) Exercise 2: Cooperative Multi-Agent Deep Reinforcement Learning

deded

## 3) Exercise 3: Competitive Multi-Agent Deep Reinforcement Learning

dedededed