# CS6700 : Reinforcement Learning
Programming Assignment #3

Author: Solène Butruille
Roll number: CS20F001

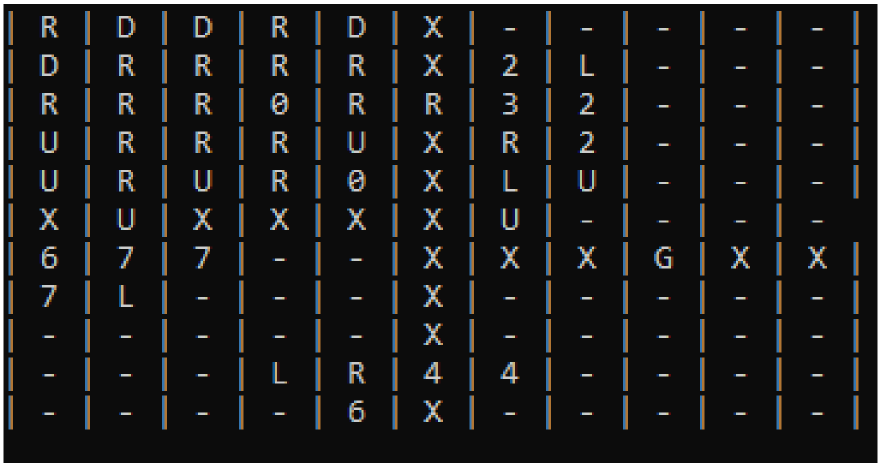
# 1 Hierarchical Reinforcement Learning

In order to visualize the Q values, I decided to print the grid word and to put in the squares the option which has the highest value in the Q Values table for this state. States represented with a ”-” means that they were never chosen as starting state. Below you can find the legend.
0 : option in room 1, goal = right hallway
1 : option in room 1, goal = hallway down
2 : option in room 2, goal = left hallway
3 : option in room 2, goal = hallway down
4 : option in room 3, goal = hallway up
5 : option in room 3, goal = left hallway
6 : option in room 4, goal = hallway up
7 : option in room 4, goal = hallway right
D : down
R : right
L : left
U : up

## 1.1 Solution without intra-option Q learning

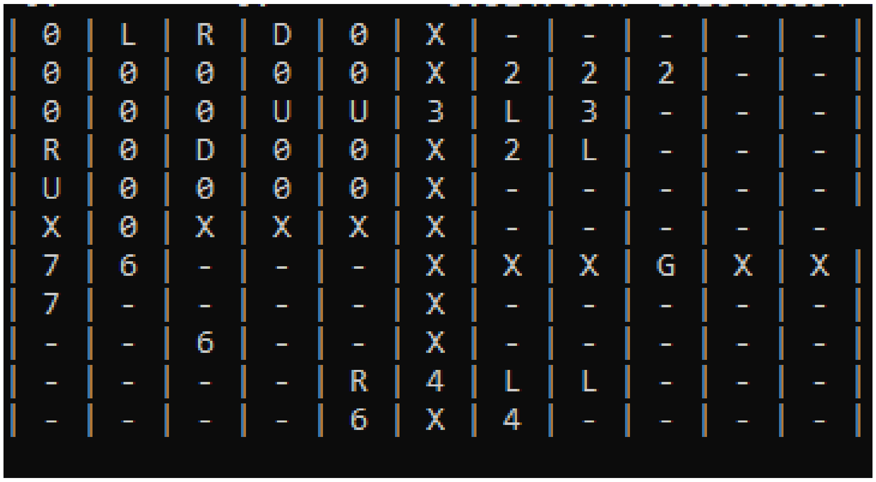
### 1.1.1 Q Values for a goal in a hallway

I started with a lot of tests on the parameters. I finally found that the $\alpha$ value doesn't affect much the final values as long as it is $> 0$ and we have an important number of iterations. In fact, the final result are not depending on the $\alpha$ value because I feel that they depend on the first chosen states which are actually random. As we can see bellow (they are 2 representations for $\alpha = 0.1$), with the exact same parameters, I can have results very different. It is not a real problem as they are multiple possibilities to reach the goal. The thing that doesn't change with the same parameters is the exploration (number of states with a value). I choose to take small values for $\alpha$ in order to have low computation. To have optimal results, I decided to run 10 000 episodes. I think it is a bit too much but it is not too slow and at least I am sure that it got close to convergence. I decided to plot my results for $\alpha = 0.3, 0.1(X2), 0.01$ as I explained above, the difference which we can see the most is the exploration.



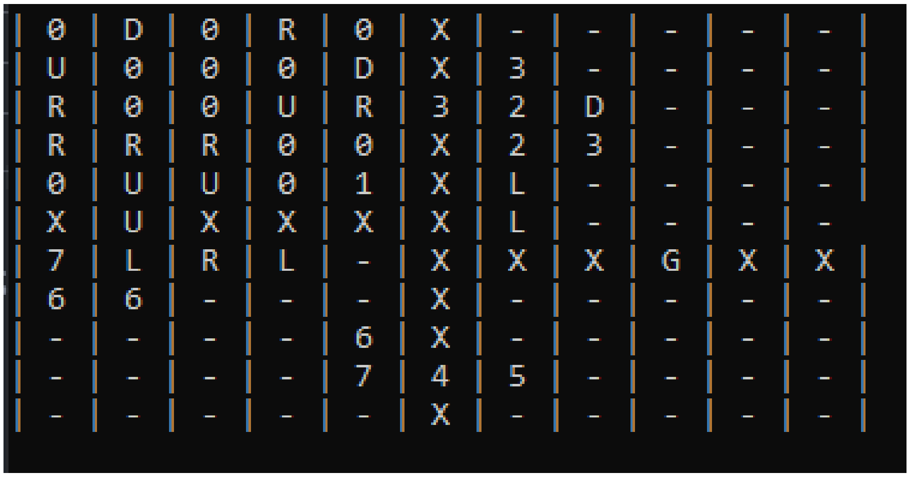
```
| R | D | D | R | D | X | - | - | - | - | - |
| D | R | R | R | R | X | 2 | L | - | - | - |
| R | R | R | 0 | R | R | 3 | 2 | - | - | - |
| U | R | R | R | U | X | R | 2 | - | - | - |
| U | R | U | R | 0 | X | L | U | - | - | - |
| X | U | X | X | X | X | U | - | - | - | -
| 6 | 7 | 7 | - | - | X | X | X | G | X | X |
| 7 | L | - | - | - | X | - | - | - | - | - |
| - | - | - | - | - | X | - | - | - | - | - |
| - | - | - | L | R | 4 | 4 | - | - | - | - |
| - | - | - | - | 6 | X | - | - | - | - | - |
```

Figure 1: $\alpha = 0.3$



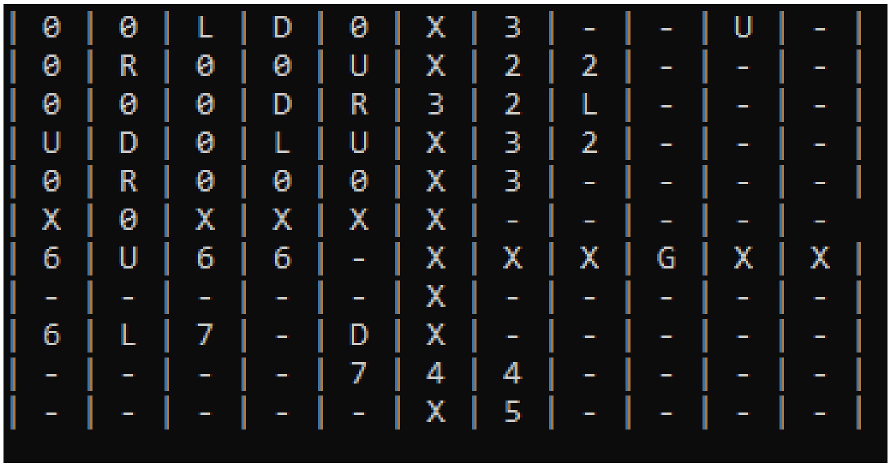
```
| 0 | L | R | D | 0 | X | - | - | - | - | - |
| 0 | 0 | 0 | 0 | 0 | X | 2 | 2 | 2 | - | - |
| 0 | 0 | 0 | U | U | 3 | L | 3 | - | - | - |
| R | 0 | D | 0 | 0 | X | 2 | L | - | - | - |
| U | 0 | 0 | 0 | 0 | X | - | - | - | - | - |
| X | 0 | X | X | X | X | - | - | - | - | -
| 7 | 6 | - | - | - | X | X | X | G | X | X |
| 7 | - | - | - | - | X | - | - | - | - | - |
| - | - | 6 | - | - | X | - | - | - | - | - |
| - | - | - | - | R | 4 | L | L | - | - | - |
| - | - | - | - | 6 | X | 4 | - | - | - | - |
```

Figure 2: $\alpha = 0.1$



```
| 0 | D | 0 | R | 0 | X | - | - | - | - | - |
| U | 0 | 0 | 0 | D | X | 3 | - | - | - | - |
| R | 0 | 0 | U | R | 3 | 2 | D | - | - | - |
| R | R | R | 0 | 0 | X | 2 | 3 | - | - | - |
| 0 | U | U | 0 | 1 | X | L | - | - | - | - |
| X | U | X | X | X | X | L | - | - | - | -
| 7 | L | R | L | - | X | X | X | G | X | X |
| 6 | 6 | - | - | - | X | - | - | - | - | - |
| - | - | - | - | 6 | X | - | - | - | - | - |
| - | - | - | - | 7 | 4 | 5 | - | - | - | - |
| - | - | - | - | - | X | - | - | - | - | - |
```

Figure 3: $\alpha = 0.1$

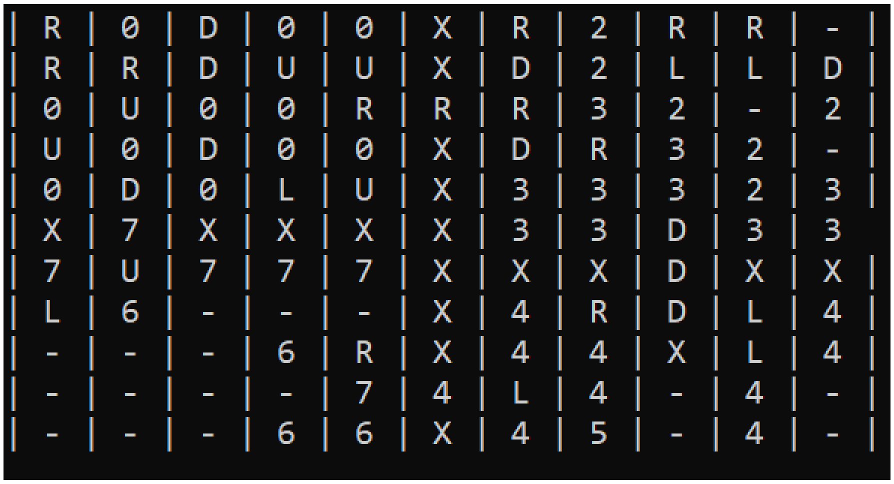


```
| 0 | 0 | L | D | 0 | X | 3 | - | - | U | - |
| 0 | R | 0 | 0 | U | X | 2 | 2 | - | - | - |
| 0 | 0 | 0 | D | R | 3 | 2 | L | - | - | - |
| U | D | 0 | L | U | X | 3 | 2 | - | - | - |
| 0 | R | 0 | 0 | 0 | X | 3 | - | - | - | - |
| X | 0 | X | X | X | X | - | - | - | - | -
| 6 | U | 6 | 6 | - | X | X | X | G | X | X |
| - | - | - | - | - | X | - | - | - | - | - |
| 6 | L | 7 | - | D | X | - | - | - | - | - |
| - | - | - | - | 7 | 4 | 4 | - | - | - | - |
| - | - | - | - | - | X | 5 | - | - | - | - |
```

Figure 4: $\alpha = 0.01$

### 1.1.2 Q Values for a goal in a room

If the goal is now inside a room instead of in a hallway, it changes a lot because the options can't find it. As the option's goal are to go to a hallway, it happens that the option goes through the goal but doesn't stop because it is not the sub-goal at this time. Therefor, we can see that with this goal, it will need to explore more to find the goal.

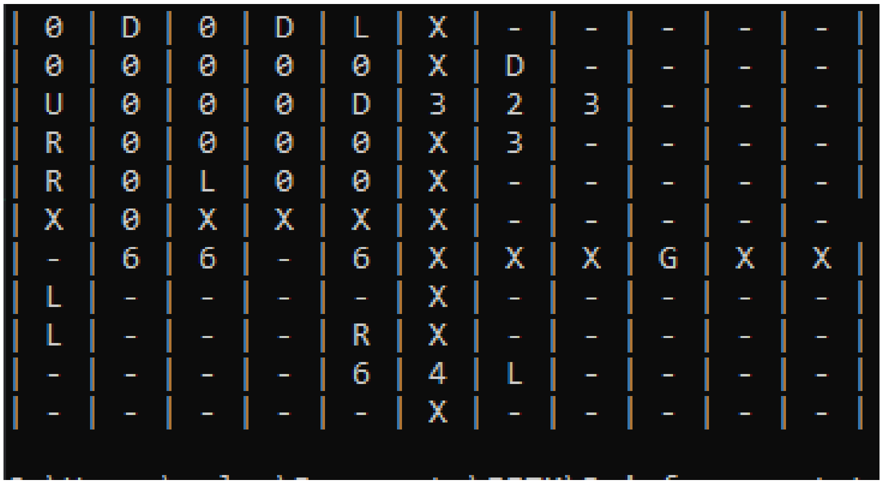


```
| R | 0 | D | 0 | 0 | X | R | 2 | R | R | - |
| R | R | D | U | U | X | D | 2 | L | L | D |
| 0 | U | 0 | 0 | R | R | R | 3 | 2 | - | 2 |
| U | 0 | D | 0 | 0 | X | D | R | 3 | 2 | - |
| 0 | D | 0 | L | U | X | 3 | 3 | 3 | 2 | 3 |
| X | 7 | X | X | X | X | 3 | 3 | D | 3 | 3
| 7 | U | 7 | 7 | 7 | X | X | X | D | X | X |
| L | 6 | - | - | - | X | 4 | R | D | L | 4 |
| - | - | - | 6 | R | X | 4 | 4 | X | L | 4 |
| - | - | - | - | 7 | 4 | L | 4 | - | 4 | - |
| - | - | - | 6 | 6 | X | 4 | 5 | - | 4 | - |
```

Figure 5: $\alpha = 0.1$, Goal inside the room

### 1.1.3 Comparison without noise

I also find it interesting to compare the result with the classic environment to results in an environment without noise. What I noticed is that sometimes a state is going to take quickly a big value (for example the hallway between the first and second room.) Then all states of room 1 will chose to go to that state even if there is another solution better closer. On the figure 5, we can even see that some states in the room under the room 1 are choosing to go back in room 1 and then to go to the hallway of room 2 despite the fact that it is longer. As explain above, my theory is that the hallway between room 1 and 2 as such an important value in the q table that it crushes all chances for the other possibilities.



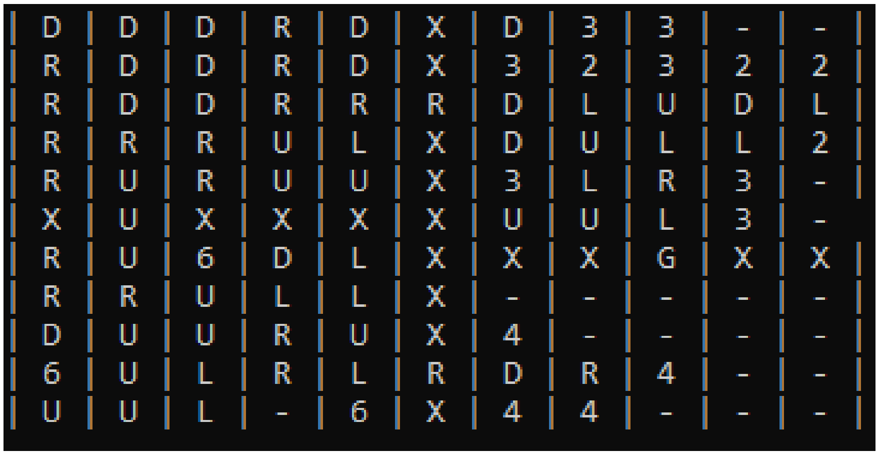
```
| 0 | D | 0 | D | L | X | - | - | - | - | - |
| 0 | 0 | 0 | 0 | 0 | X | D | - | - | - | - |
| U | 0 | 0 | 0 | D | 3 | 2 | 3 | - | - | - |
| R | 0 | 0 | 0 | 0 | X | 3 | - | - | - | - |
| R | 0 | L | 0 | 0 | X | - | - | - | - | - |
| X | 0 | X | X | X | X | - | - | - | - | -
| - | 6 | 6 | - | 6 | X | X | X | G | X | X |
| L | - | - | - | - | X | - | - | - | - | - |
| L | - | - | - | R | X | - | - | - | - | - |
| - | - | - | - | 6 | 4 | L | - | - | - | - |
| - | - | - | - | - | X | - | - | - | - | - |
```

Figure 6: $\alpha = 0.1$, Goal in the hallway, no noise

## 1.2 Solution with Intra option Q learning

If we now implement q learning inside the options, it is going to take much time to converge to the optimal policy. Because the options will have to be learn, there values will not be very good and the algorithm will chose more often to take simple values (U,D,R,L) instead of options.

### 1.2.1 Q Values for a goal in a hallway

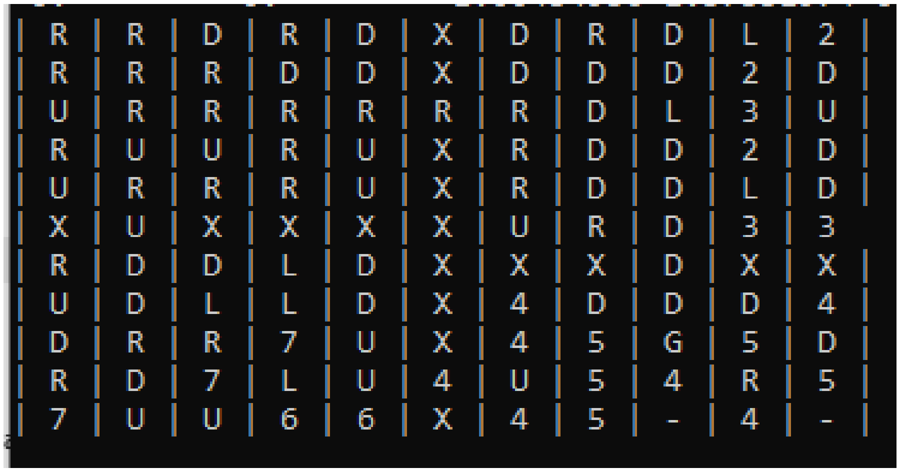


```
| D | D | D | R | D | X | D | 3 | 3 | - | - |
| R | D | D | R | D | X | 3 | 2 | 3 | 2 | 2 |
| R | D | D | R | R | R | D | L | U | D | L |
| R | R | R | U | L | X | D | U | L | L | 2 |
| R | U | R | U | U | X | 3 | L | R | 3 | - |
| X | U | X | X | X | X | U | U | L | 3 | -
| R | U | 6 | D | L | X | X | X | G | X | X |
| R | R | U | L | L | X | - | - | - | - | - |
| D | U | U | R | U | X | 4 | - | - | - | - |
| 6 | U | L | R | L | R | D | R | 4 | - | - |
| U | U | L | - | 6 | X | 4 | 4 | - | - | - |
```

Figure 7: $\alpha = 0.1$, Goal in the hallway

### 1.2.2 Q Values for a goal in a room

In this special case, we can see that the exploration is almost maximal. Almost every state will be at least once a starting state, whereas in the other situations, a lot of states are never starting state.



```
| R | R | D | R | D | X | D | R | D | L | 2 |
| R | R | R | D | D | X | D | D | D | 2 | D |
| U | R | R | R | R | R | R | D | L | 3 | U |
| R | U | U | R | U | X | R | D | D | 2 | D |
| U | R | R | R | U | X | R | D | D | L | D |
| X | U | X | X | X | X | U | R | D | 3 | 3
| R | D | D | L | D | X | X | X | D | X | X |
| U | D | L | L | D | X | 4 | D | D | D | 4 |
| D | R | R | 7 | U | X | 4 | 5 | G | 5 | D |
| R | D | 7 | L | U | 4 | U | 5 | 4 | R | 5 |
| 7 | U | U | 6 | 6 | X | 4 | 5 | - | 4 | - |
```

Figure 8: $\alpha = 0.1$, Goal in the room