Lab 3

Axel Holmberg (axeho681)

10/4/2020

Task 2.2

Below are the environments after 10, 100 and 1000 and 10000 runs done with the Greedy Policy. Questions and answers under images.

- *What has the agent learned after the first 10 episodes?*

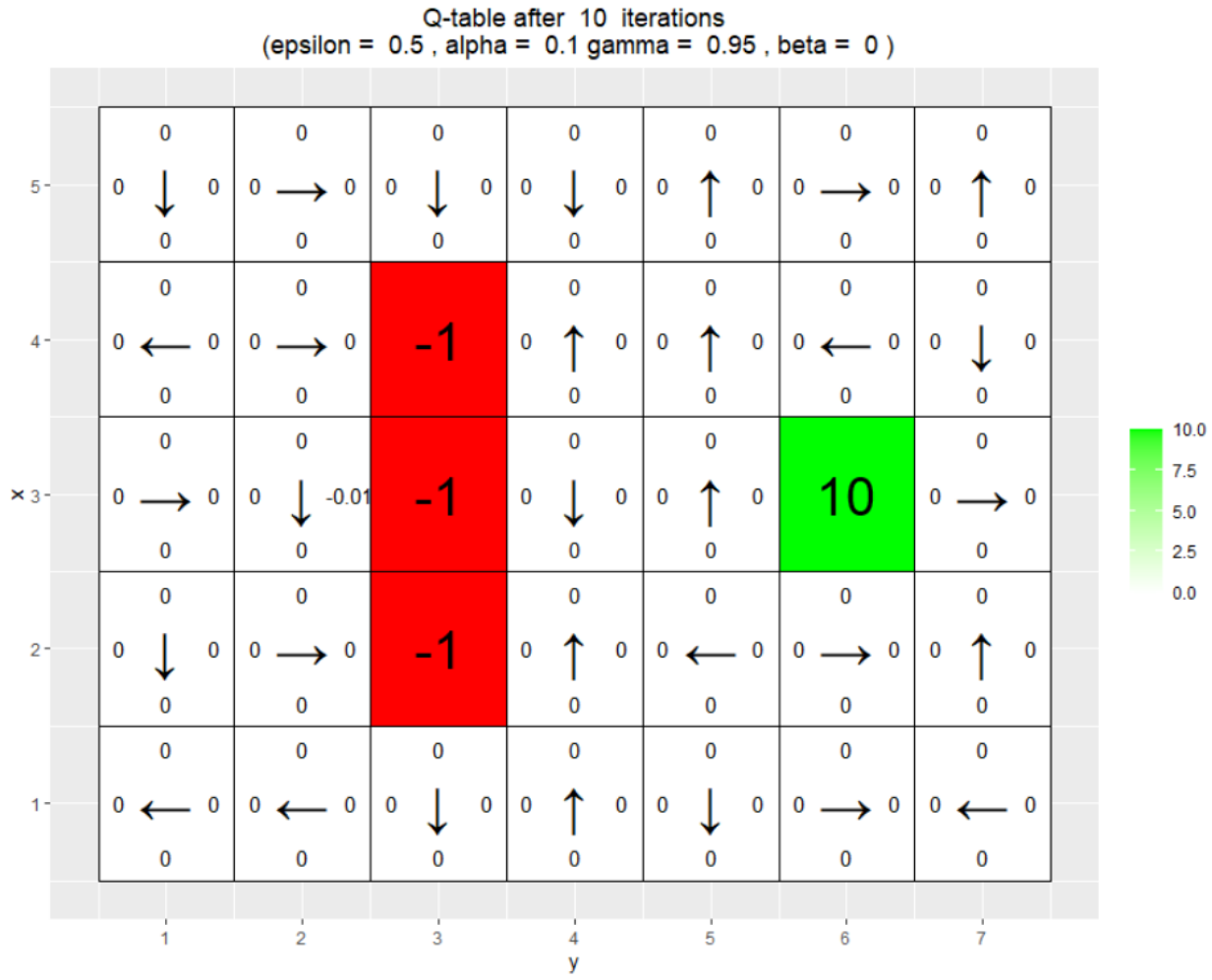


Figure 1: 10 iterations

As one can see in the image above the agent has not learned much in the first 10 iterations. The only q-value thas has been updated is after having gotten a negative reward in $x = 3, y = 2$.

- *Is the final greedy policy (after 10000 episodes) optimal? Why / Why not ?*

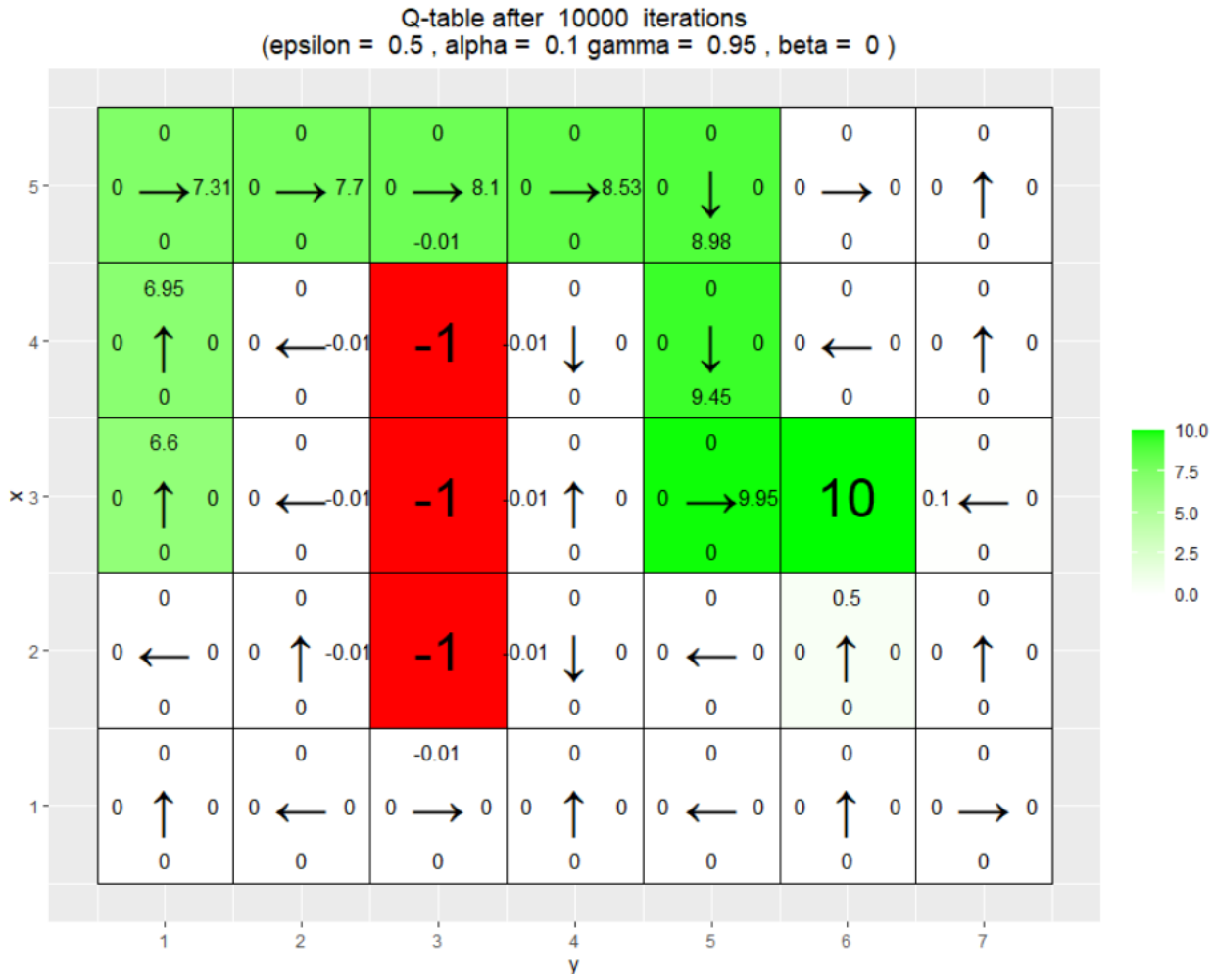


Figure 2: 10 000 iterations

The image above shows the path after 10000 iterations. The agent takes one of the most optimal paths as it takes the least amount of steps possible to get to the reward. Alhough there are many possible solutions for this where the agent takes exactly 9 steps.

- *Does the agent learn that there are multiple paths to get to the positive reward ? If not, what could be done to make the agent learn this ?*

No, the agent only learns one path. The reason for this is that there is no incentive for it to look for other paths other than the one that it has gotten some reward for as it doesn't have any randomness in the policy for its' movement.

Task 2.3

As γ denounces the discount factor that means that it adjusts how much of the values of the Q-table should affect the new updated Q-value, called correction. One can see this quite clearly if one compares $\gamma = 0.75$ with $\gamma = 0.95$.

The big difference one can see in the images above is how all the q-values are a lot higher when $\gamma = 0.95$.

One more thing that is clearly visible in these images is the effect of ϵ. ϵ is the threshold for exploration. As

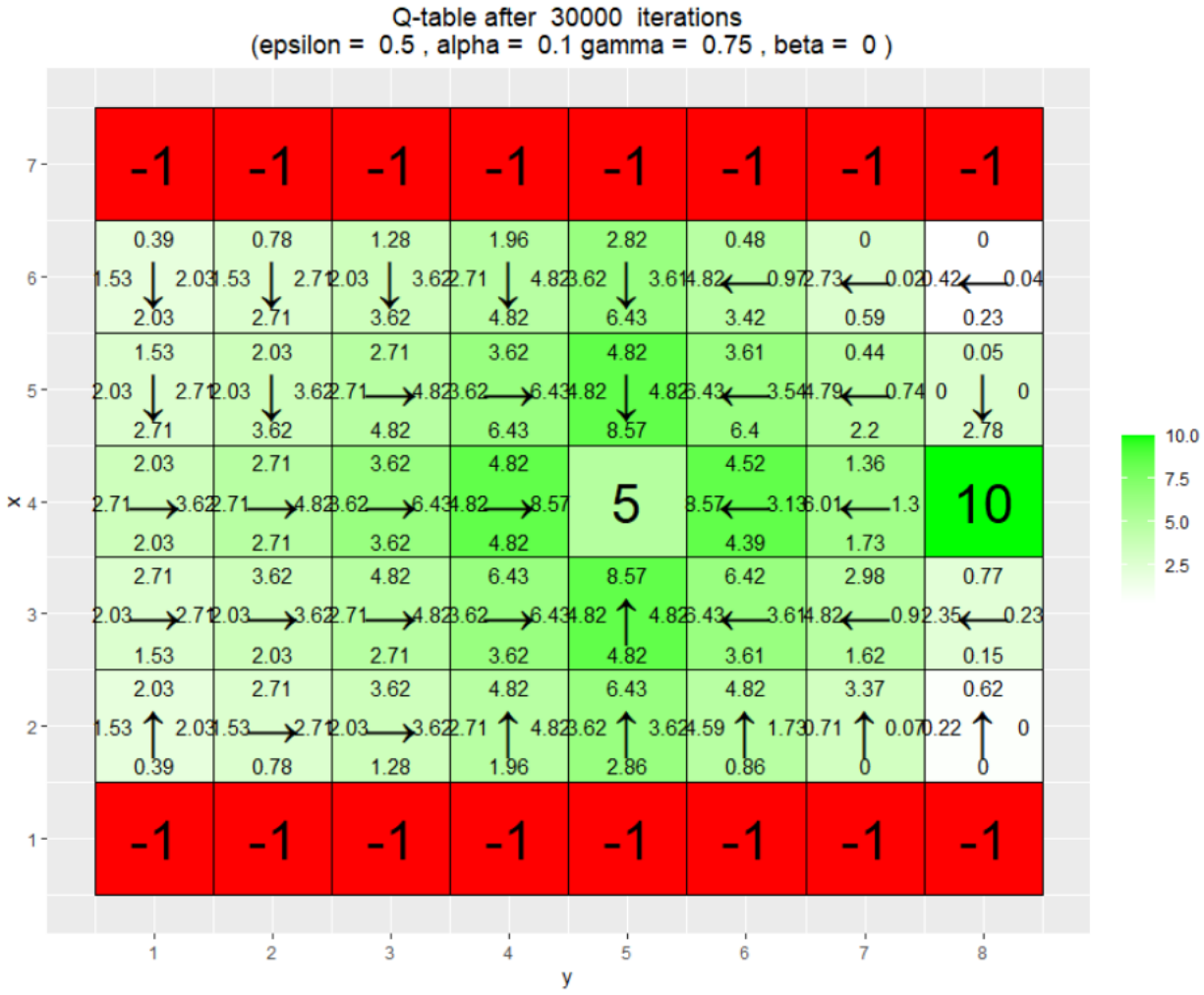


Figure 3: $\gamma = 0.75, \epsilon = 0.5$

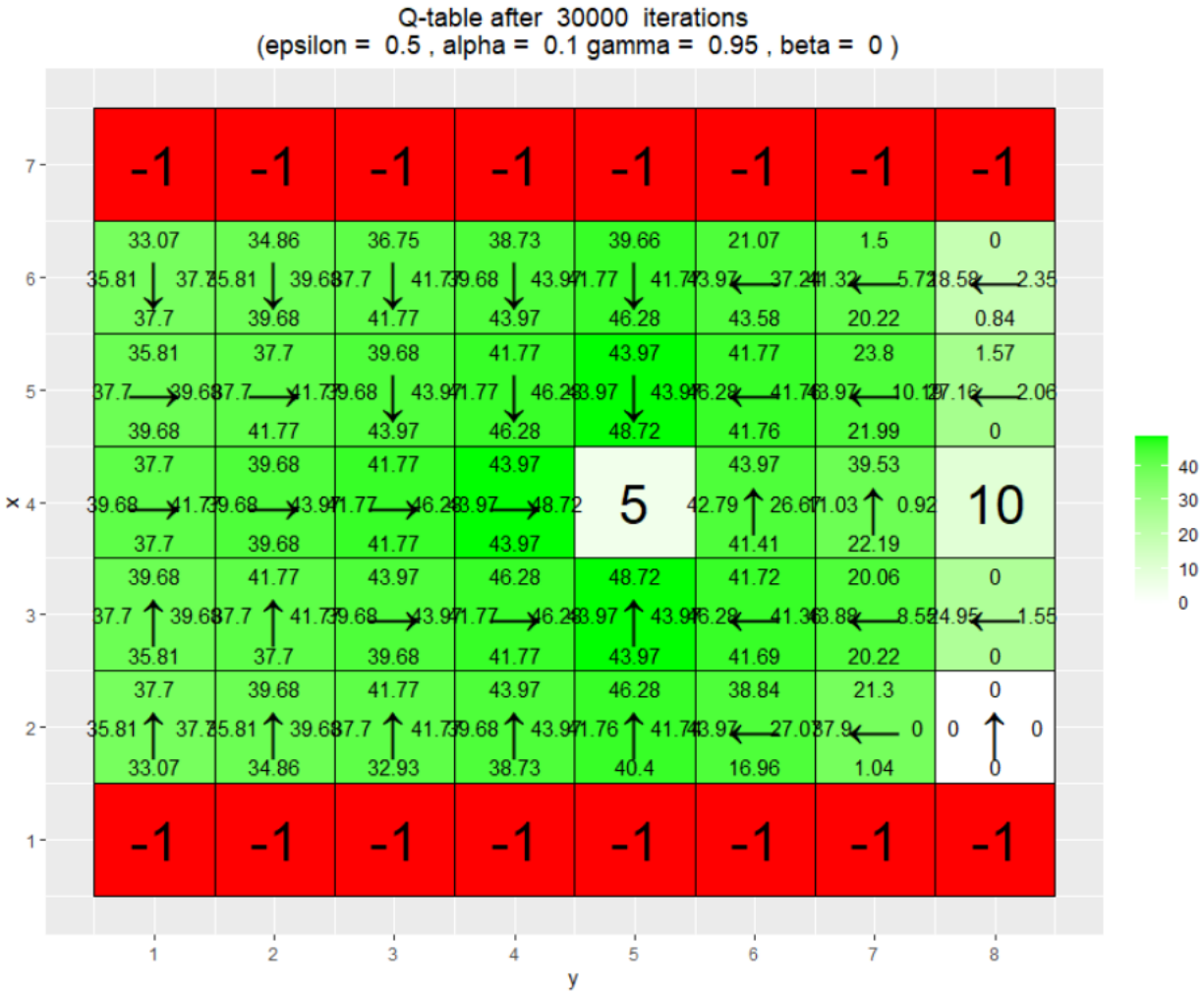


Figure 4: $\gamma = 0.95, \epsilon = 0.5$

$\epsilon = 0.5$ in both of the above that means than in each evaluation of the ϵ-greedy policy there is a 50% chance that it instead takes a random step. That means that it has a high chance of exploring more of the tiles. If one compares this with where $\epsilon = 0.1$ like the images below the results is quite clear.

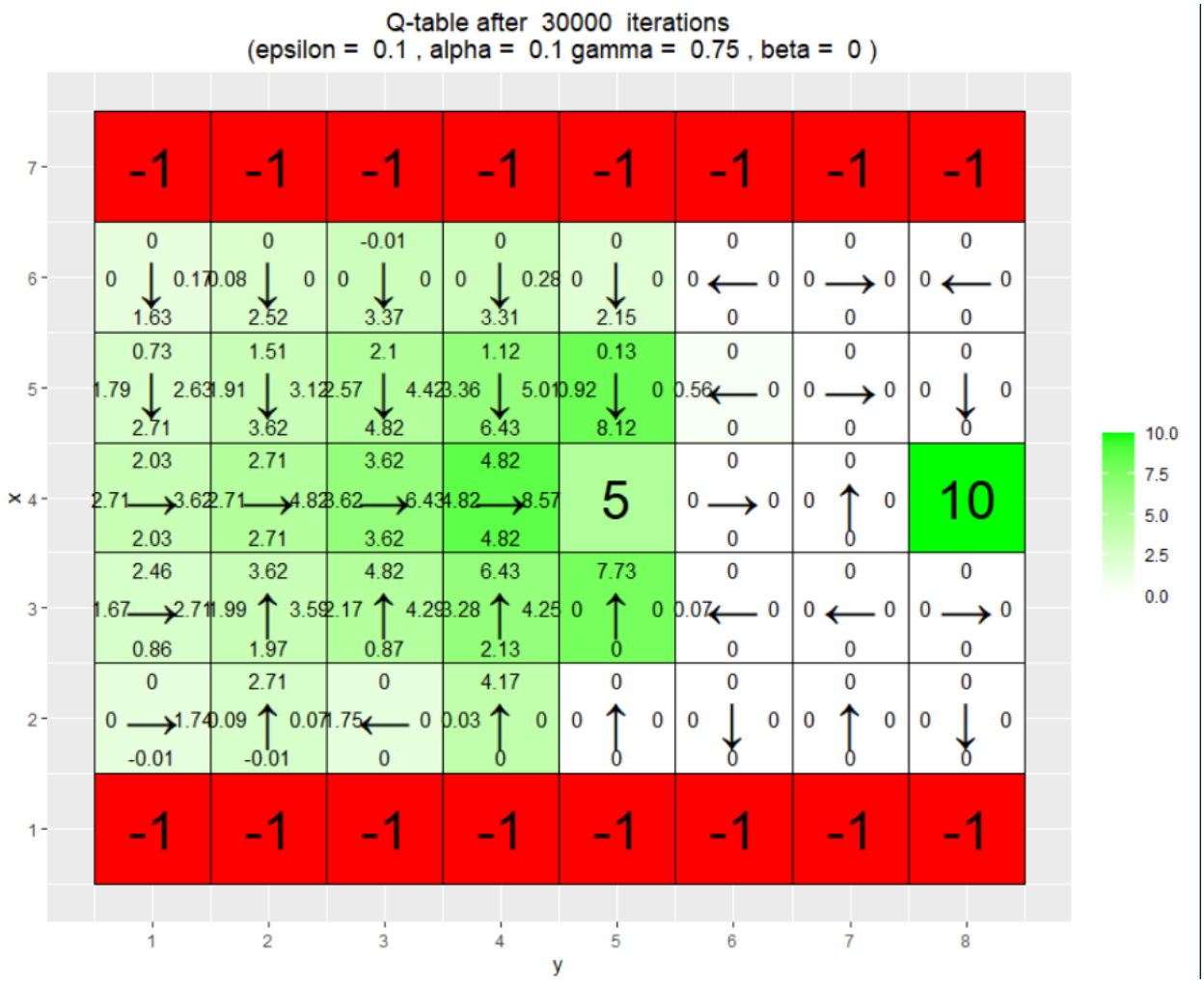


Figure 5: $\gamma = 0.75, \epsilon = 0.1$

As one can see it almost never takes a step beyond the tile with a reward of 5 as the low exploration rate makes it go straight to the tile with a reward of 5. This can also be seen in the graph below showing the rolling mean of the rewards where it aslmost always 5.

One can also look at the correction graph. The correction graph (see below) shows the correction of each step. This shows how much of a correction that occurs with each step.

The difference between these two are quite significant. The higher value of ϵ leads to more variation and mainly a higher correction overall. What this means is that with each episode the correction of the q values are higher. The reason for is that the higher the probability of acting greedily is the more different paths and more will be discorvered and the more correction is done each step and episode.

There are more to be analysed from the data and the graphs, but above is the key takeaways that I did of the graphs.

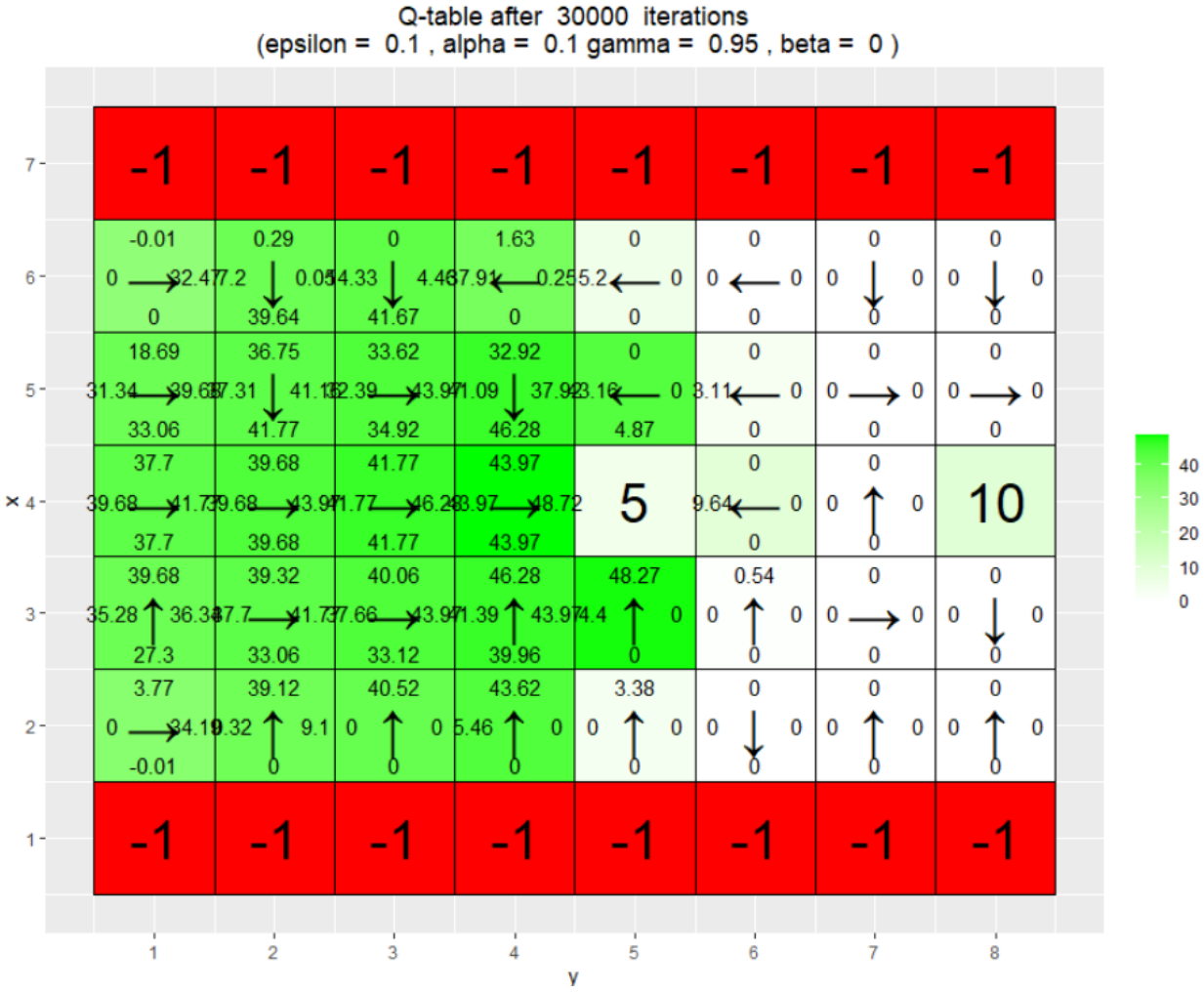


Figure 6: $\gamma = 0.95, \epsilon = 0.1$

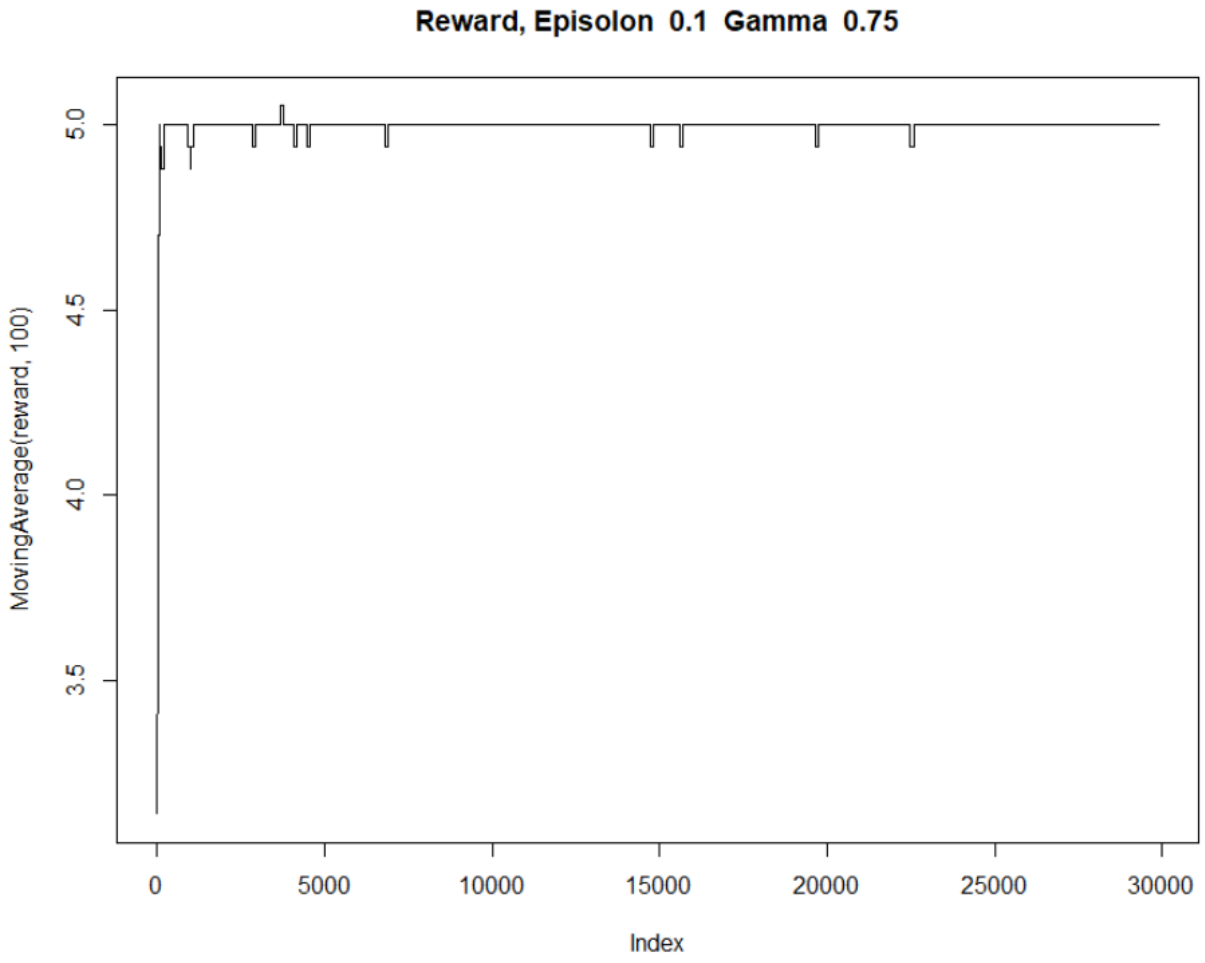


Figure 7: $\gamma = 0.75, \epsilon = 0.1$

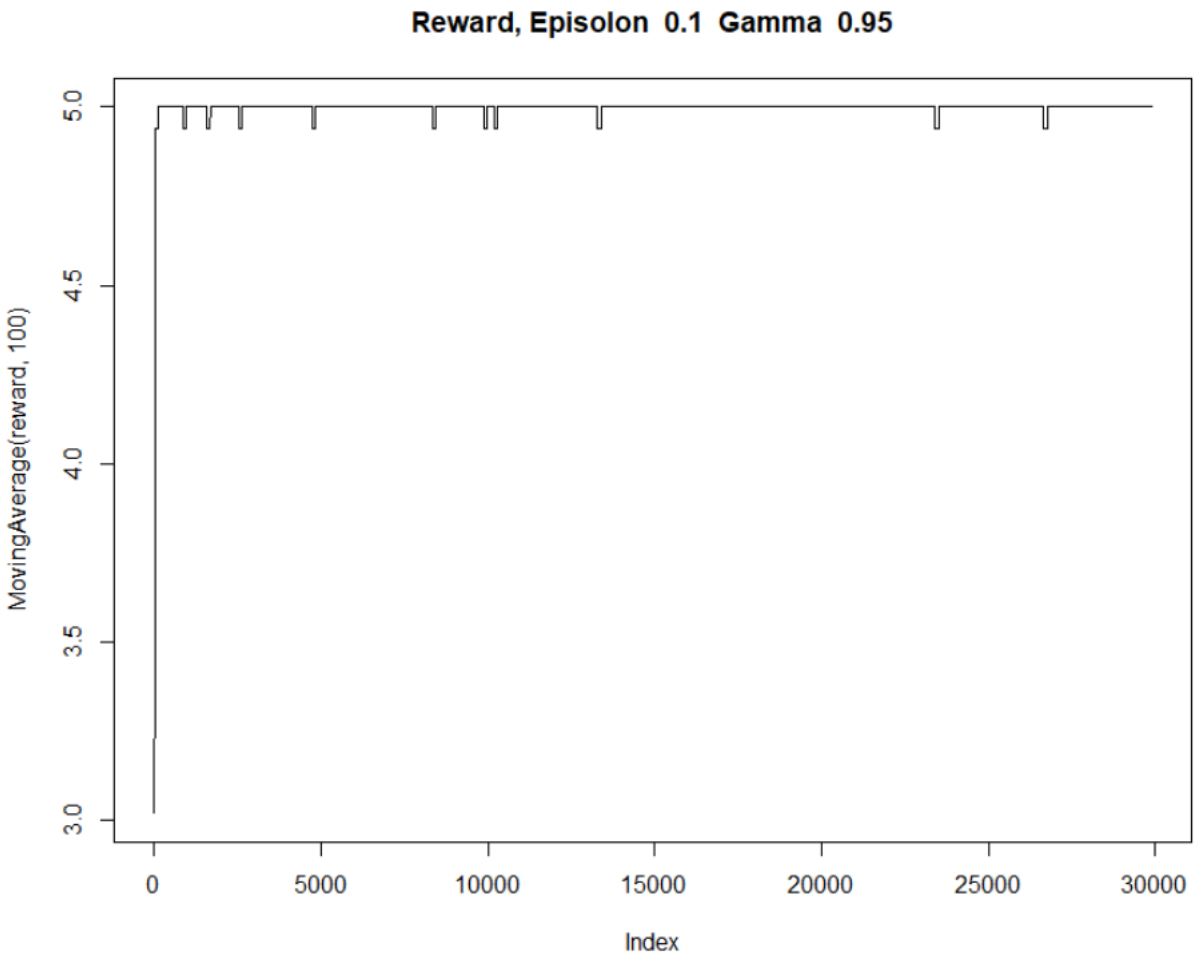


Figure 8: $\gamma = 0.95, \epsilon = 0.1$

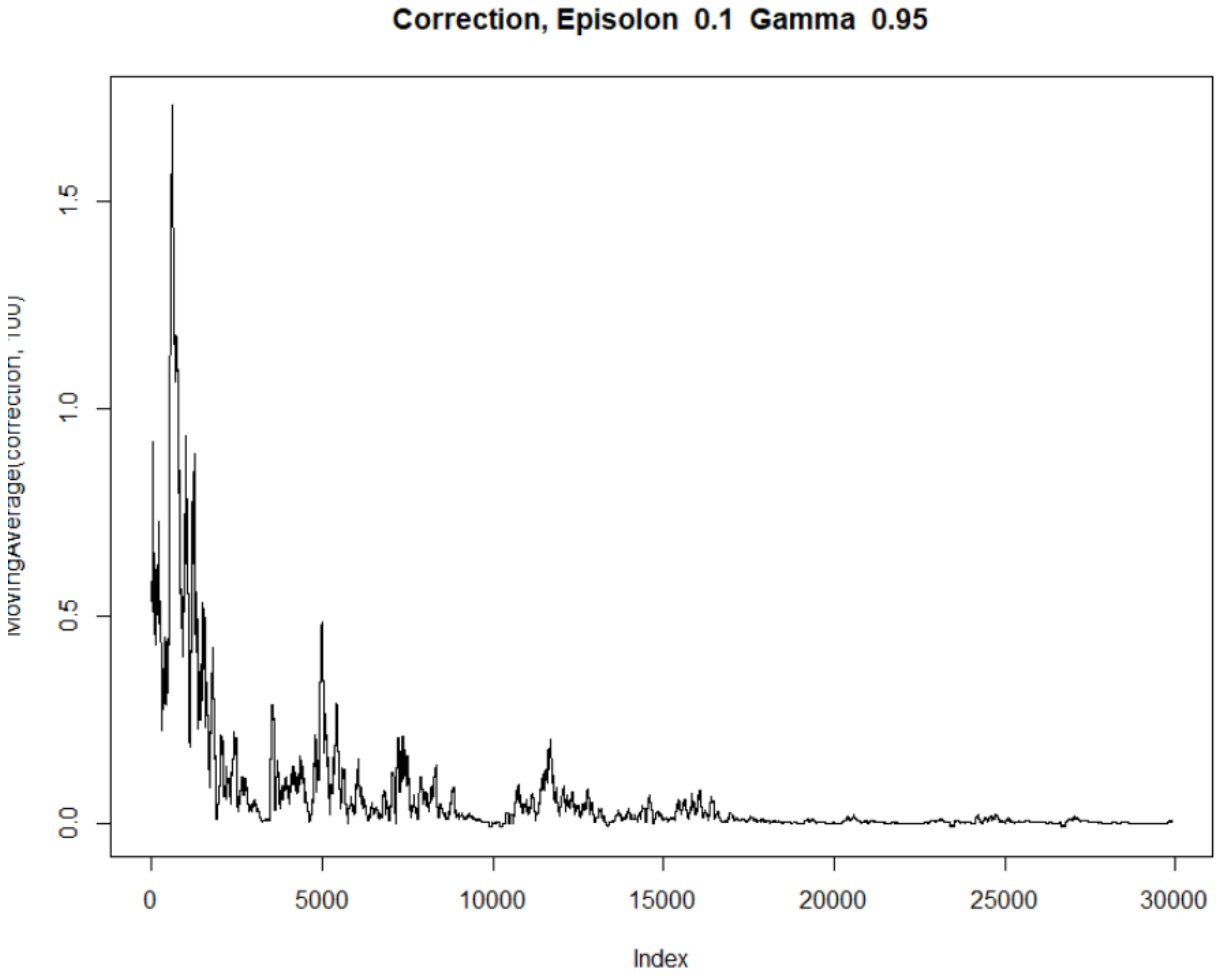


Figure 9: $\gamma = 0.95, \epsilon = 0.1$

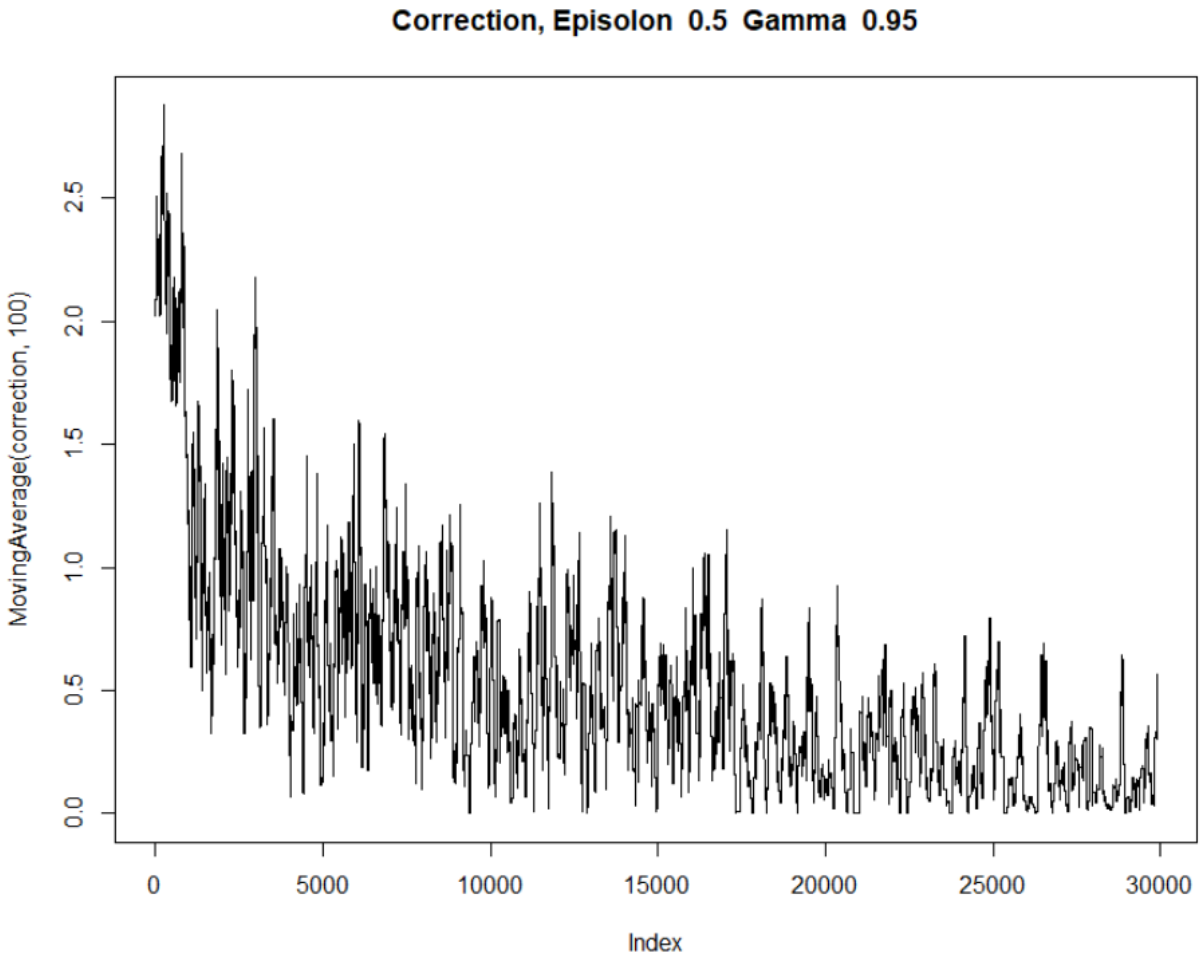
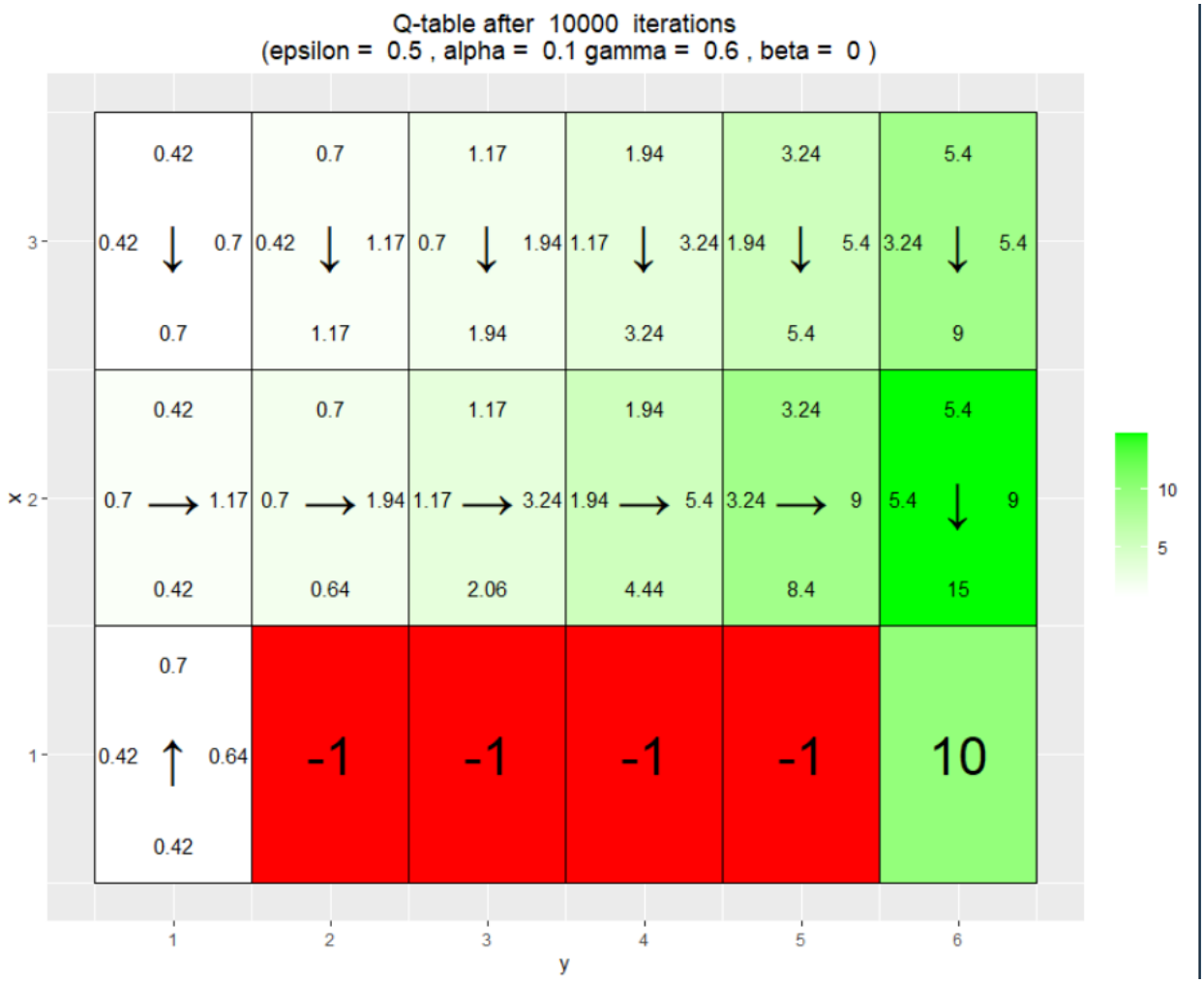
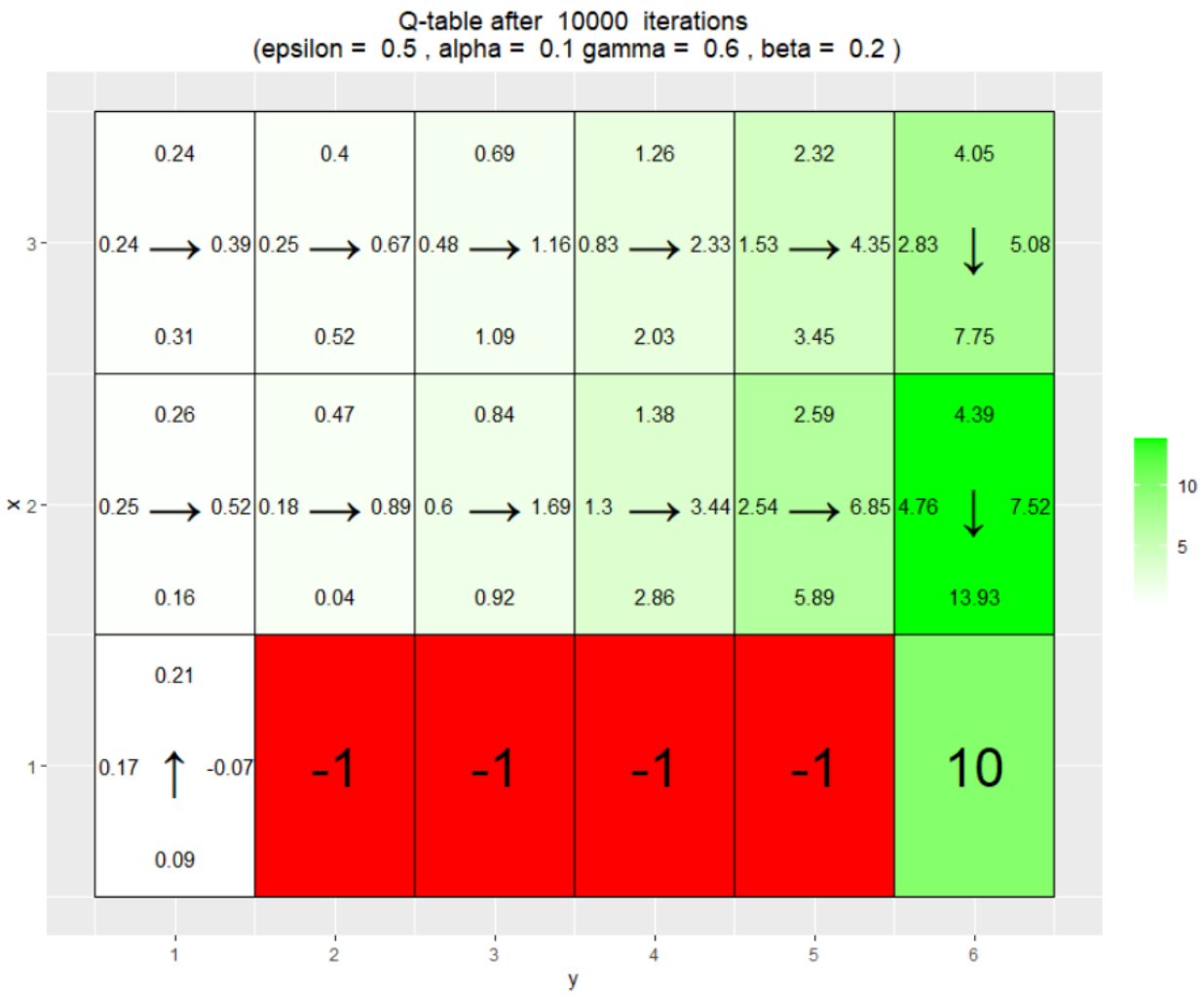


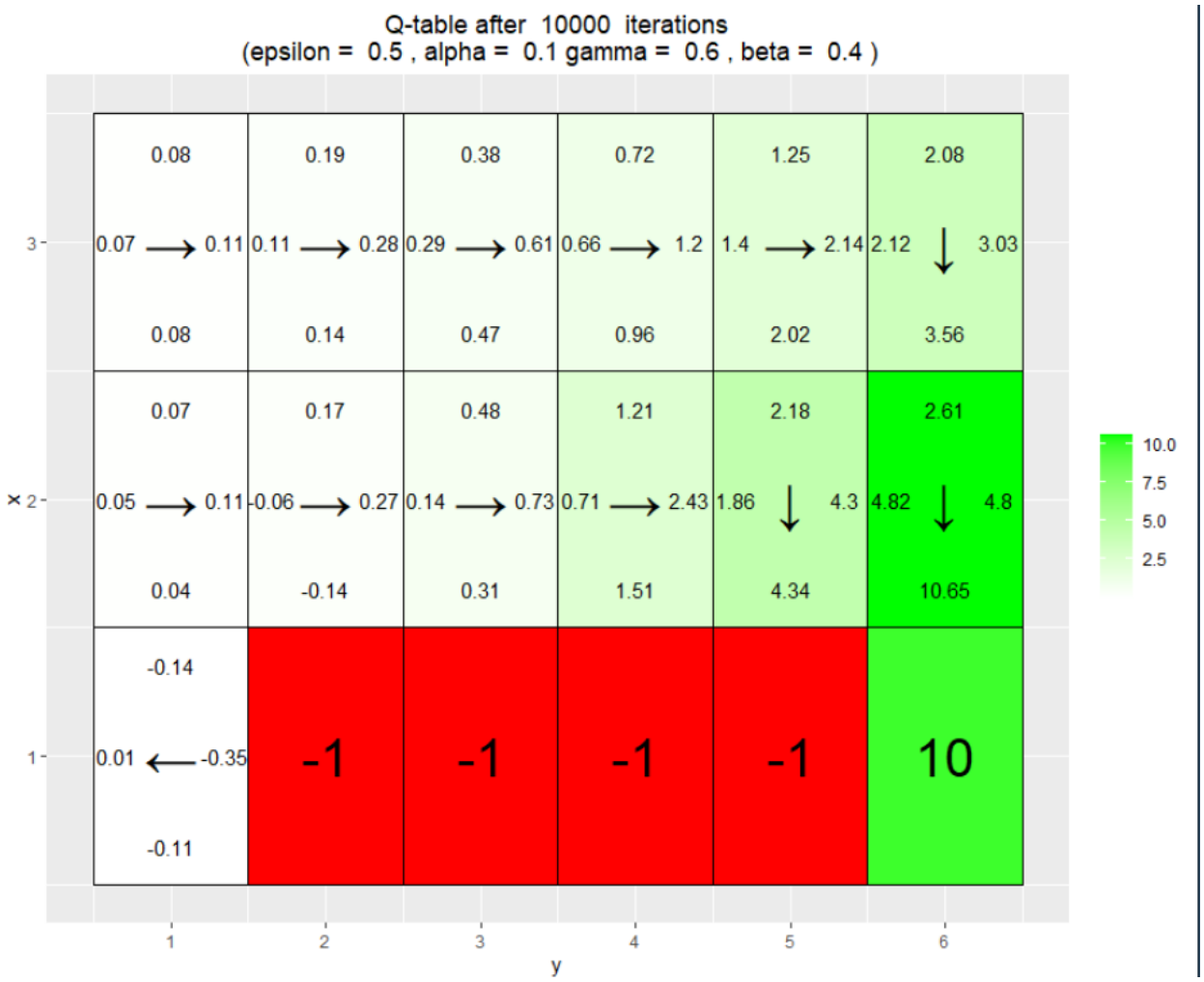
Figure 10: $\gamma = 0.95, \epsilon = 0.5$

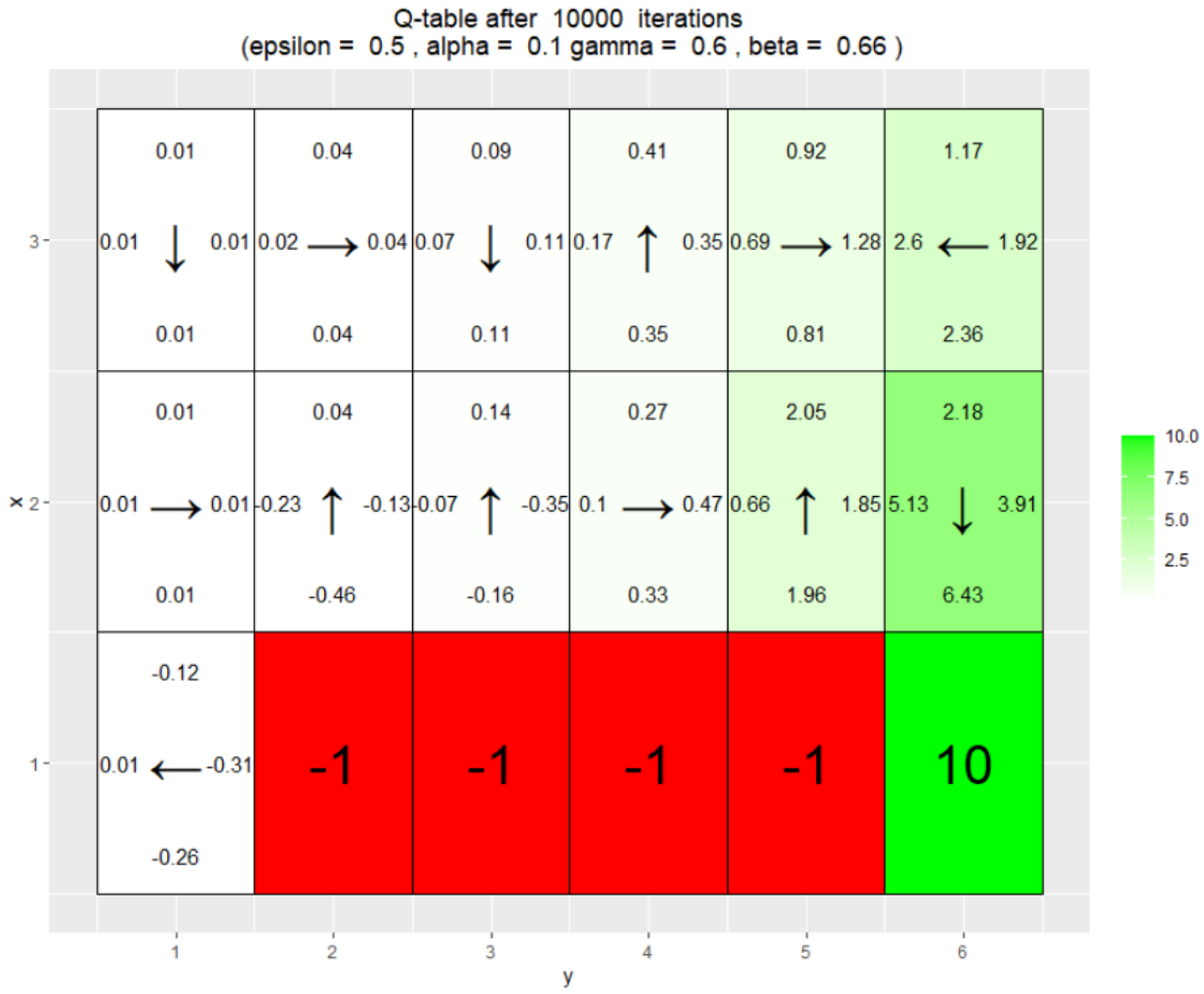
Task 2.4

β is the slipping factor. The slipping factor can make the action "slip" and make the action be something different. The β is the probability of the agent slipping to the side when trying to move with each step. The steps taken for each beta




Q-table after 10000 iterations
(epsilon = 0.5 , alpha = 0.1 gamma = 0.6 , beta = 0.2)
0.24
0.24 → 0.39
0.31
0.4
0.25 → 0.67
0.52
0.69
0.48 → 1.16
1.09
1.26
0.83 → 2.33
2.03
2.32
1.53 → 4.35
3.45
4.05
2.83 ↓ 5.08
7.75
0.26
0.25 → 0.52
0.16
0.47
0.18 → 0.89
0.04
0.84
0.6 → 1.69
0.92
1.38
1.3 → 3.44
2.86
2.59
2.54 → 6.85
5.89
4.39
4.76 ↓ 7.52
13.93
0.21
0.17 ↑ -0.07
0.09
-1
-1
-1
-1
10
x
3
2
1
y
1
2
3
4
5
6
10
5


Q-table after 10000 iterations
(epsilon = 0.5 , alpha = 0.1 gamma = 0.6 , beta = 0.4)
0.08
0.19
0.38
0.72
1.25
2.08
0.07 → 0.11
0.11 → 0.28
0.29 → 0.61
0.66 → 1.2
1.4 → 2.14
2.12 ↓ 3.03
0.08
0.14
0.47
0.96
2.02
3.56
0.07
0.17
0.48
1.21
2.18
2.61
0.05 → 0.11
-0.06 → 0.27
0.14 → 0.73
0.71 → 2.43
1.86 ↓ 4.3
4.82 ↓ 4.8
0.04
-0.14
0.31
1.51
4.34
10.65
-0.14
0.01 ← -0.35
-0.11
-1
-1
-1
-1
10
3
2
1
x
1
2
3
4
5
6
y
10.0
7.5
5.0
2.5


Q-table after 10000 iterations
(epsilon = 0.5 , alpha = 0.1 gamma = 0.6 , beta = 0.66)
0.01
0.01 0.01
0.01
0.04
0.02 0.04
0.04
0.09
0.07 0.11
0.11
0.41
0.17 0.35
0.35
0.92
0.69 1.28
0.81
1.17
2.6 1.92
2.36
0.01
0.01 0.01
0.01
0.04
-0.23 -0.13
-0.46
0.14
-0.07 -0.35
-0.16
0.27
0.1 0.47
0.33
2.05
0.66 1.85
1.96
2.18
5.13 3.91
6.43
-0.12
0.01 -0.31
-0.26
-1
-1
-1
-1
10
3
2
1
x
1
2
3
4
5
6
y
10.0
7.5
5.0
2.5