CS6700

PROGRAMMING ASSIGNMENT- 1

DIPRA BHAGAT(CS21S048)

SUBHAM DAS(CS21S058)

Q-Learning:

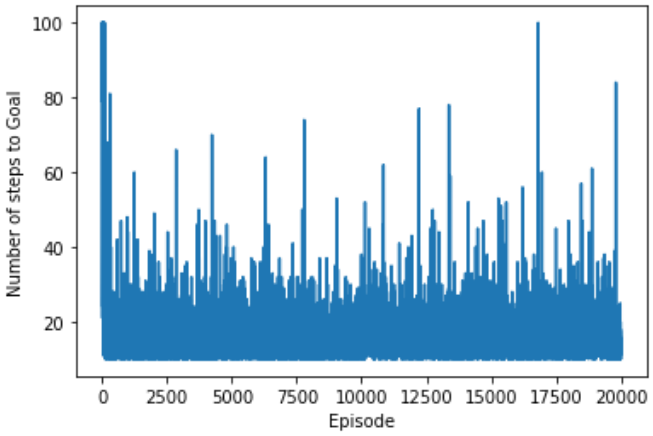
Below are the plots corresponding to each of the 16 combinations. The plots have been generated by taking an average of 3 independently run experiments.

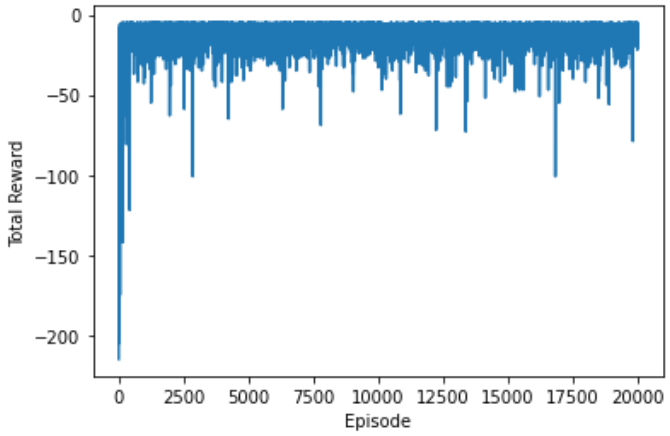
Combination 1:

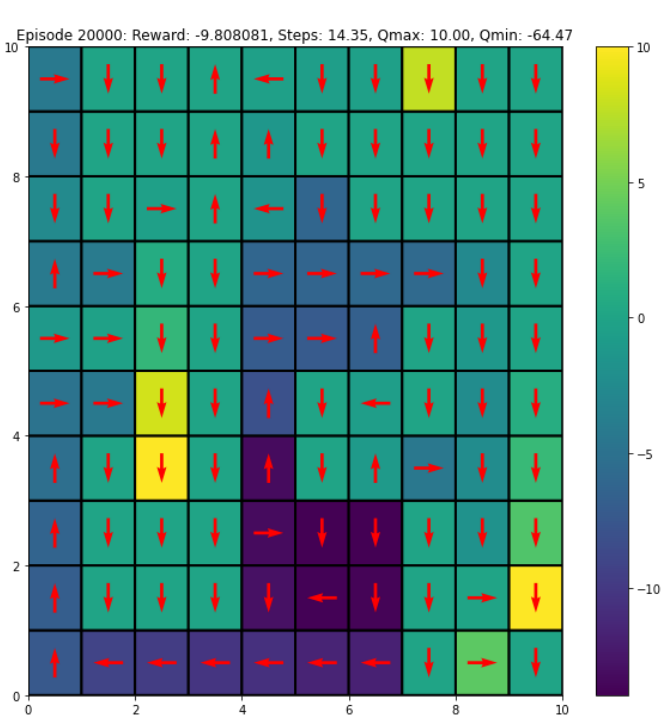
- Policy: Epsilon Greedy
- Start State: [0, 4]
- P = 1
- Wind = True

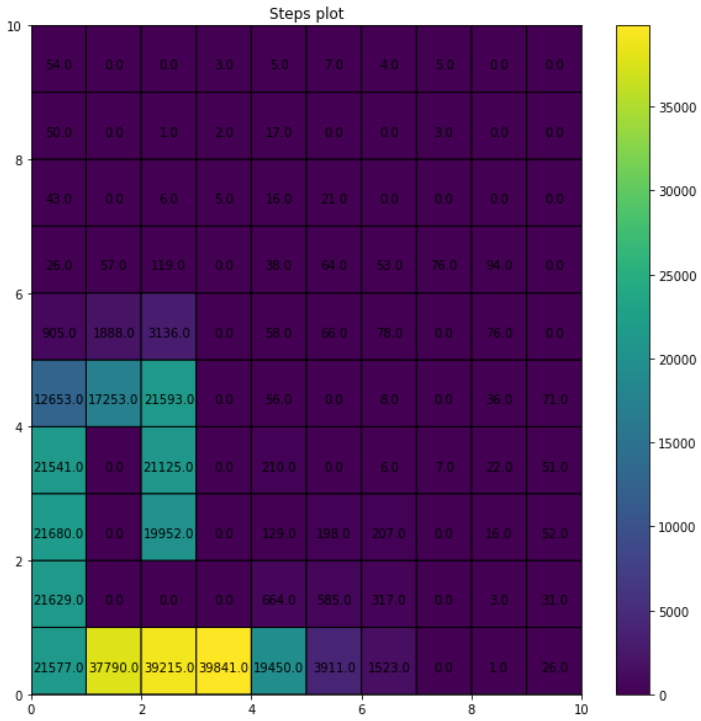
Best chosen hyperparameters:

- ε = 0.07
- α = 0.4
- ϒ = 0.94









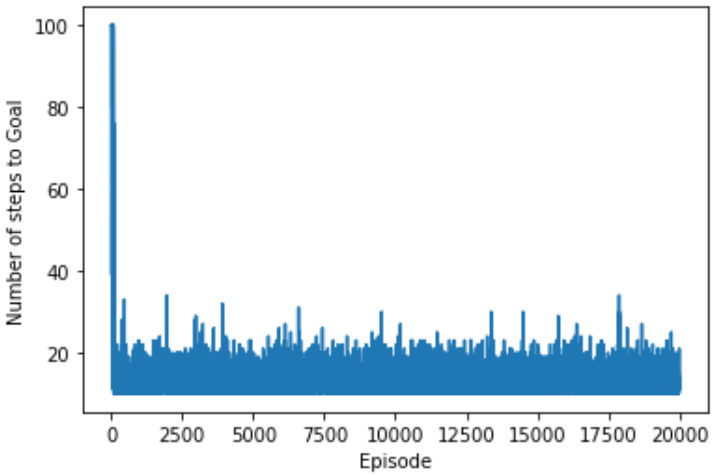
Inferences: Starting from 0,4 we can see from the step heat map that it is reaching towards 2,2 goal state. There are slight fluctuations throughout the step graph indicating the role of the wind in creating unwanted situations while following older and better paths. It immediately corrects itself but again goes astray because of the unwanted factor. In the epsilon greedy policy, we had to determine the value of the epsilon first which determines whether we follow the last best path or explore a new one. In this case for higher epsilon or that is with more exploration, it is more prone to reach the closer restart states and fail the challenge hence giving very bad results. Thus epsilon value needed to be given quite low to manage this. Again for very low epsilon (0.02-0.05), it is also getting worse presumably because of not changing or adapting to a new path after reaching some bad states or obstructions. We are using 0.4 for alpha as anything bigger(0.5-0.7) or lower(0.1-0.3) is giving horrendous results. For Gamma, we are using a really big value as we have to keep looking into the future for possible obstructions and to avoid any restart states once they are reached so the same path is not repeated at any cost. Satisfactory rewards and steps are acquired.

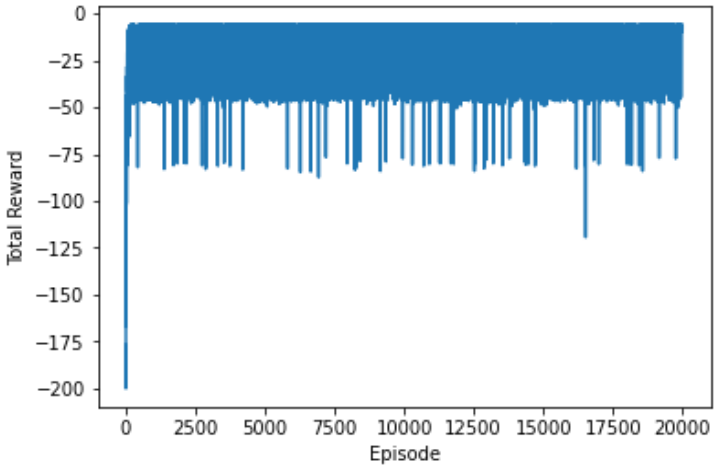
Combination 2:

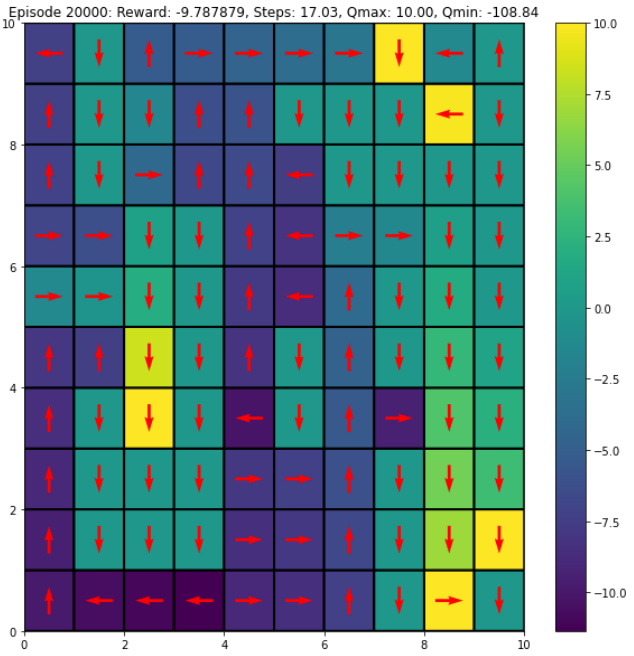
- Policy: Epsilon Greedy
- Start State: [0, 4]
- P = 1
- Wind = False

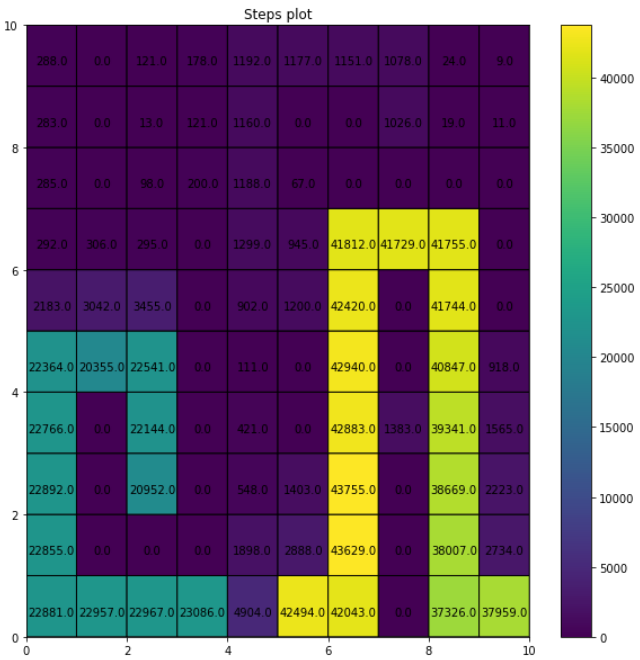
Best chosen hyperparameters:

- ε = 0.07
- α = 0.4
- ϒ = 0.94









Inferences: With the same hyperparameters as its windy counterpart it is giving the best rewards in this case. It can also be seen that the graph here looks much more stable than the windy part with fluctuations of lesser

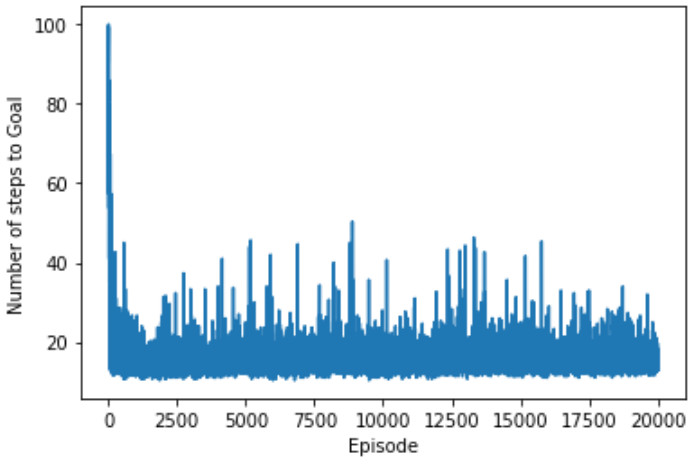
magnitude. Possibly it is adapting way faster in absence of wind and returning to the required path that it needs to take.

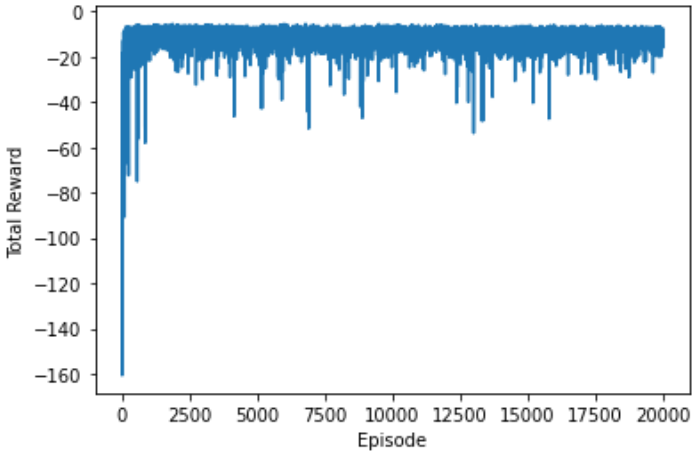
Combination 3:

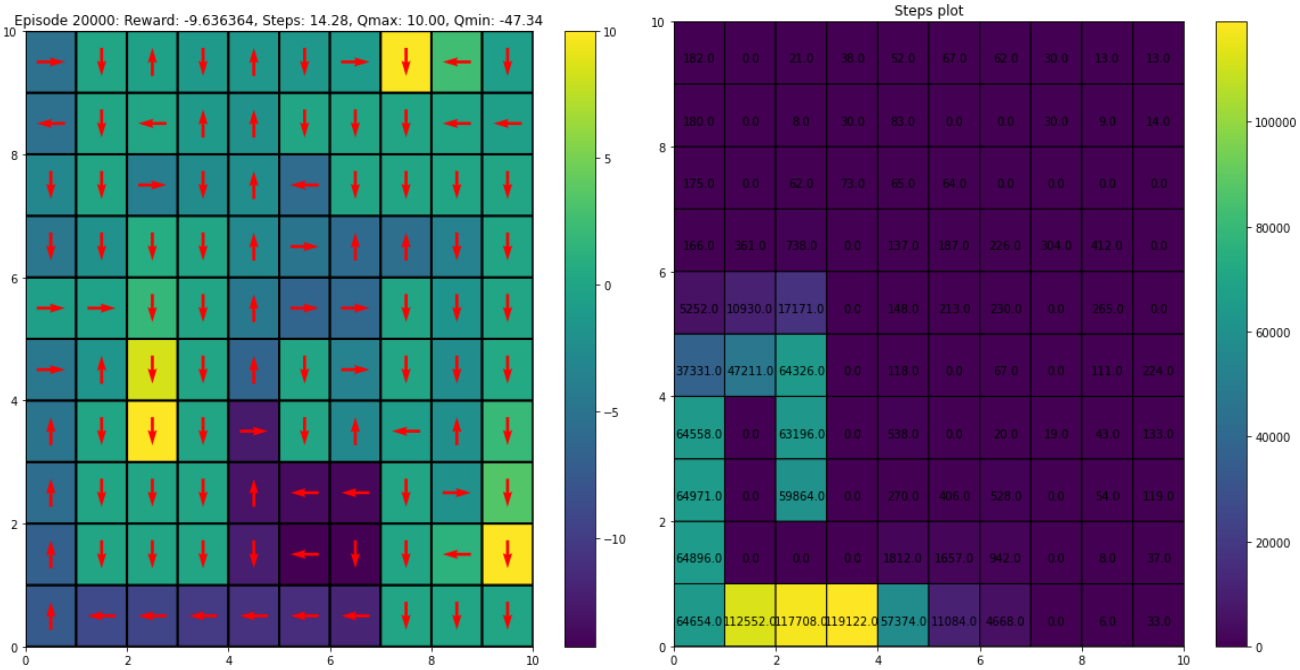
- Policy: Softmax
- Start State: [0, 4]
- P = 1
- Wind = True

Best chosen hyperparameters:

- β = 5
- α = 0.4
- ϒ = 0.94







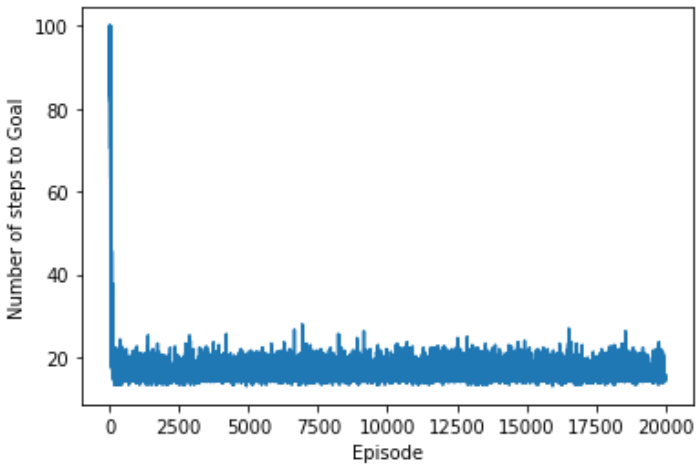
Inferences: Alpha as 0.4 and gamma as 0.94 is the sweet spot for these series of experiments…we can further tune them by using (0.38, 0.41) but we refrained from overfitting. Softmax gives much better graphs for the steps and rewards than the epsilon in this case. Windy fluctuations are there but the stability provided by the softmax policy keeps the fluctuations from getting huge. Graphs are very satisfactory.

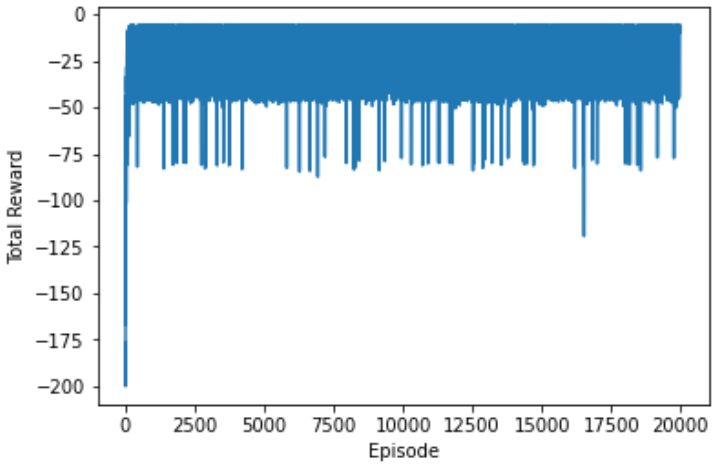
Combination 4:

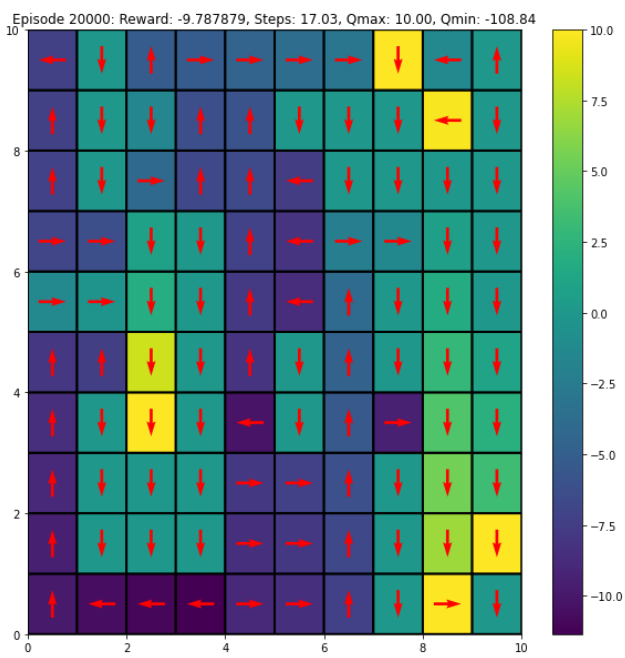
- Policy: Softmax
- Start State: [0, 4]
- P = 1
- Wind = False

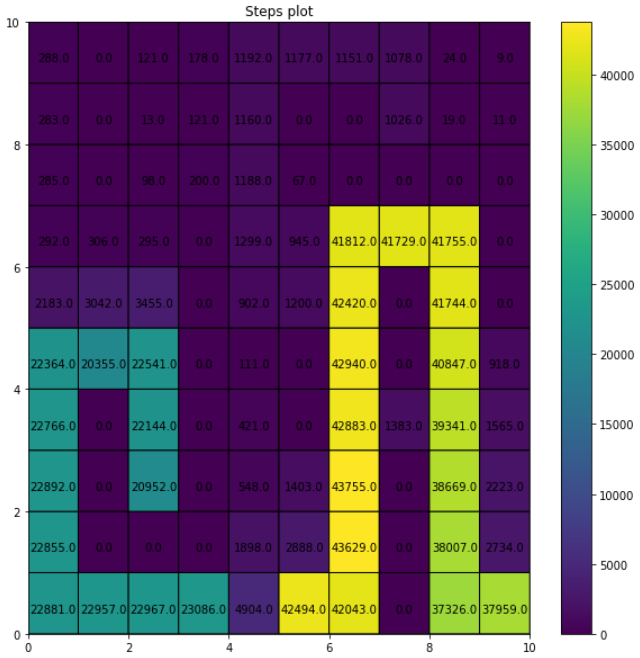
Best chosen hyperparameters:

- $\beta = 5$
- $\alpha = 0.4$
- $\Upsilon = 0.94$









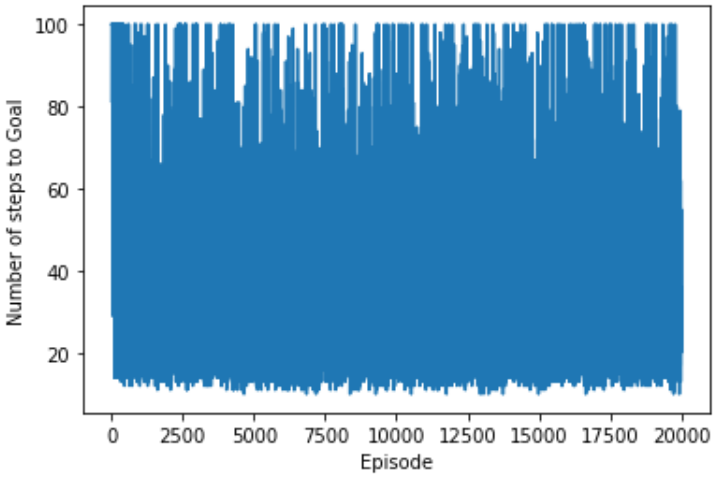
Inferences: Comparing the total reward graphs of combination-3 and 4 we can easily see how the reward is fluctuating for windy and stays stable for non-windy. The number of steps required is also devoid of any irregularities except the ones where it tries to explore instead of committing to the last best-known path. The main part is the Beta hyperparameter which stays best around the range 3 to 8 and gave us the best reward for beta=5. Anything less than 3 is making the rewards relatively high and very very high for beta greater than 30. All in all the results look good.

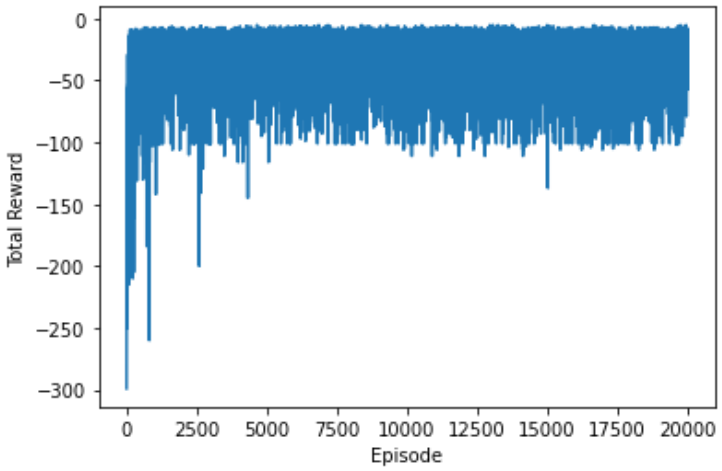
Combination 5:

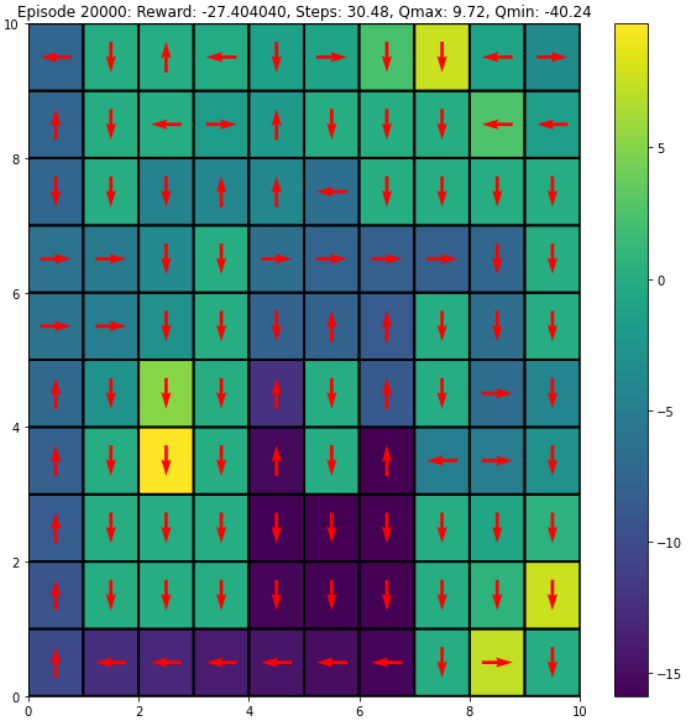
- Policy: Epsilon Greedy
- Start State: [0, 4]
- P = 0.7
- Wind = True

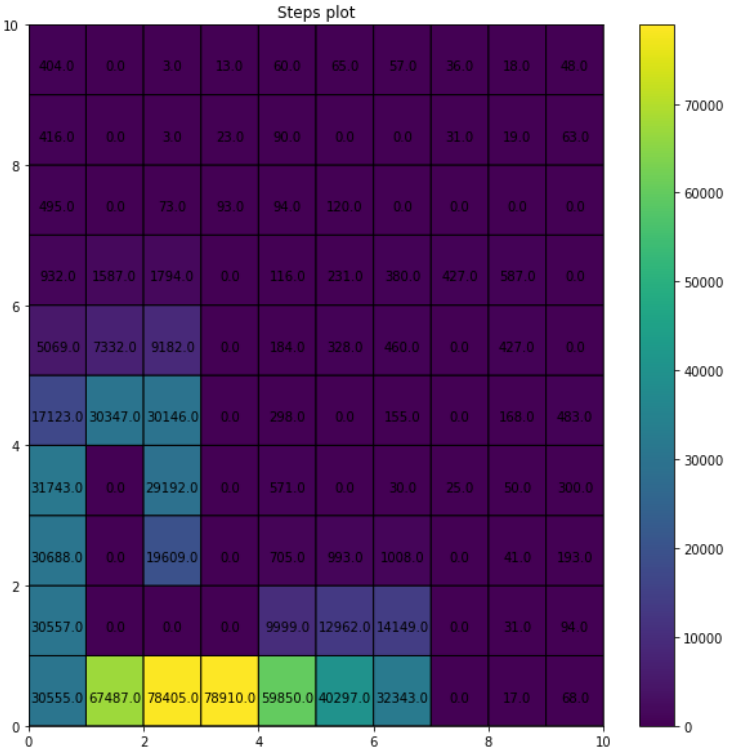
Best chosen hyperparameters:

- ε = 0.063
- α = 0.1
- ϒ = 0.94









Inferences: From the steps heat map we can see that there is more exploration on the right side of the path where it should be going. The probability of 0.7 is making the system way more unpredictable, the windy attribute is making things worse. Combined with all these factors it takes an unusually high amount of steps to reach the goal state after getting off track quite a few times. The only saving grace is the fact that the path follows a boundary

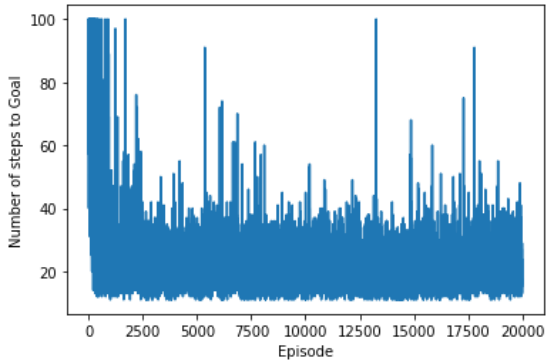
otherwise there would have been nothing to keep it from running off track in search of unknown roads. The range of epsilon became way too constricted and only some values around 0.06 could be used for this case. There might have been better epsilon in higher or lower areas but even after extensive testing, we failed to get any better results. At Least the reward did not fluctuate much after reaching a maximum. For p=0.7 alpha had to be reduced drastically.

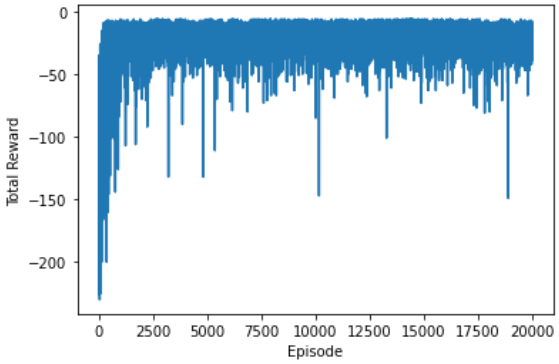
Combination 6:

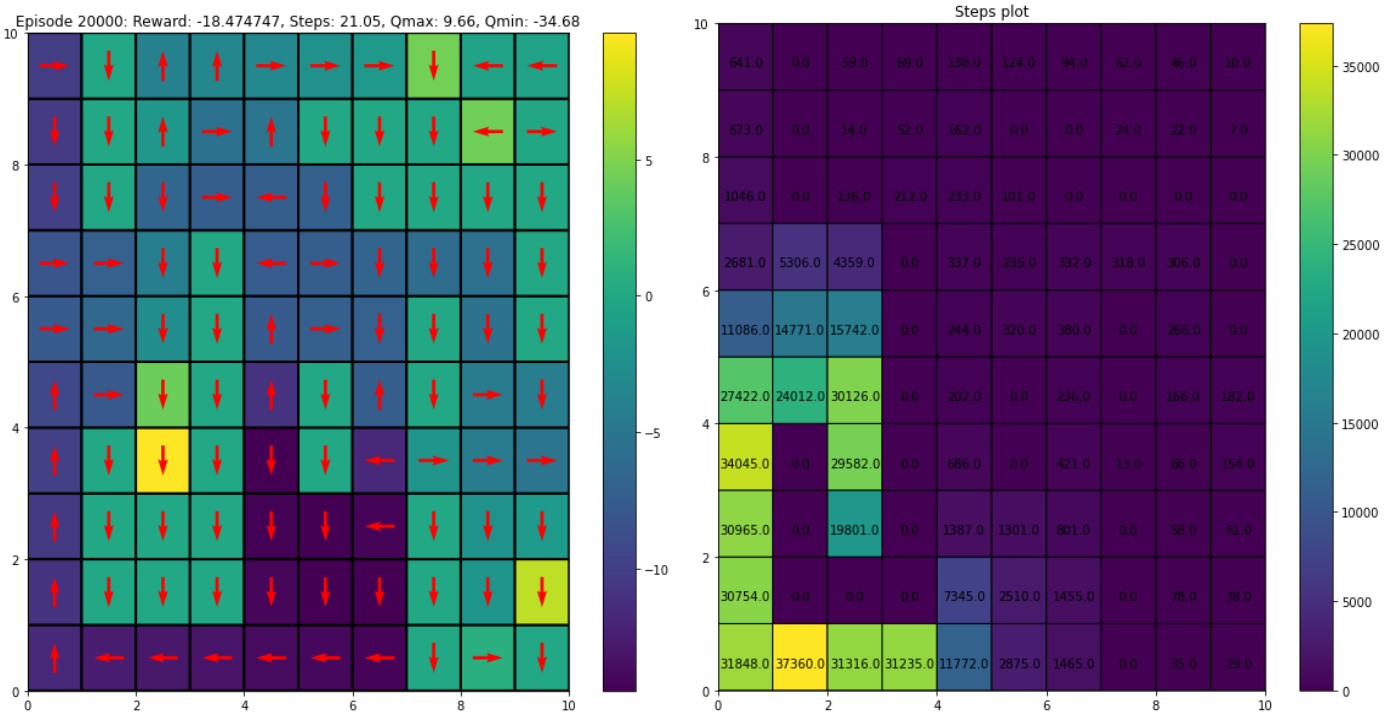
- Policy: Epsilon Greedy
- Start State: [0, 4]
- P = 0.7
- Wind = False

Best chosen hyperparameters:

- ε = 0.063
- α = 0.1
- ϒ = 0.94







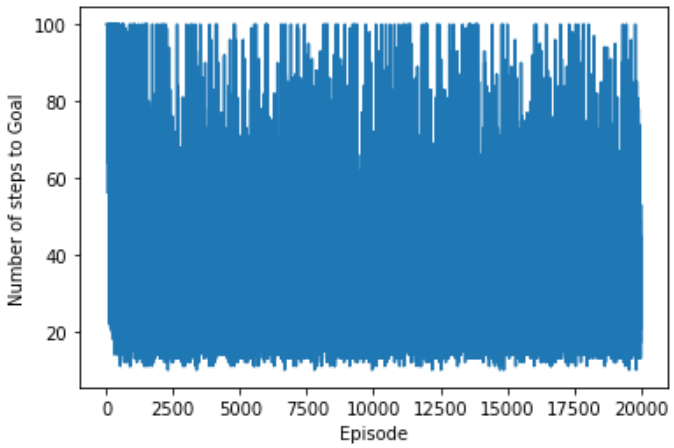
Inferences: The absence of wind made the steps graph way better than the windy version. The only randomness arrived because of the very few instances where the agent decided to move towards a different state not explored earlier, which also gave very low rewards in those cases because of moving through a lot of steps. We followed the same hyperparameters as they were visibly giving better results than with those who have changes in them.

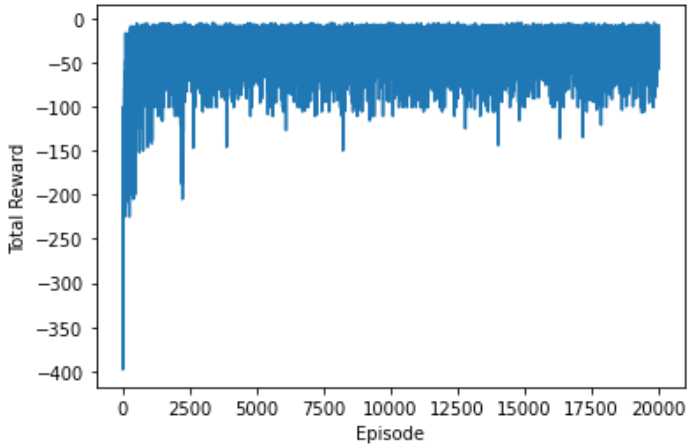
Combination 7:

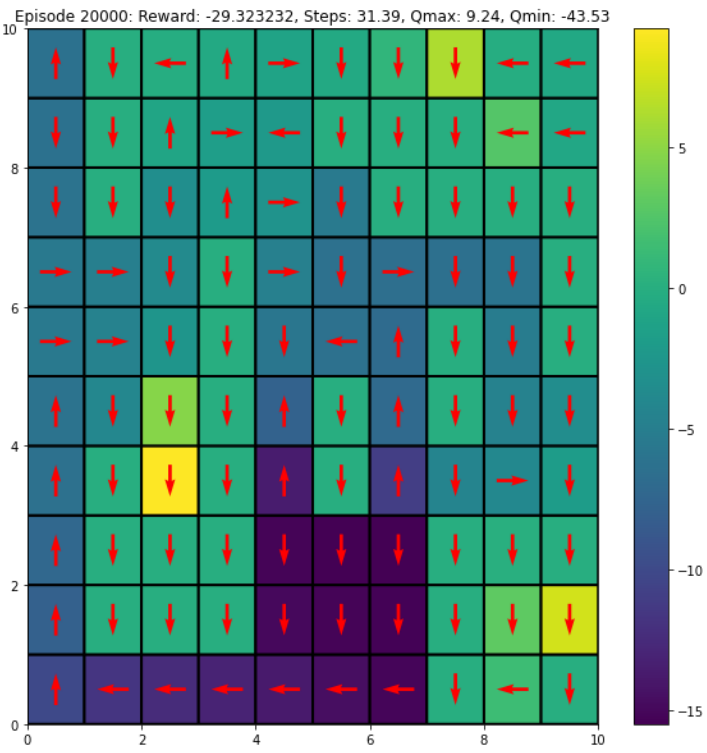
- Policy: Softmax
- Start State: [0, 4]
- P = 0.7
- Wind = True

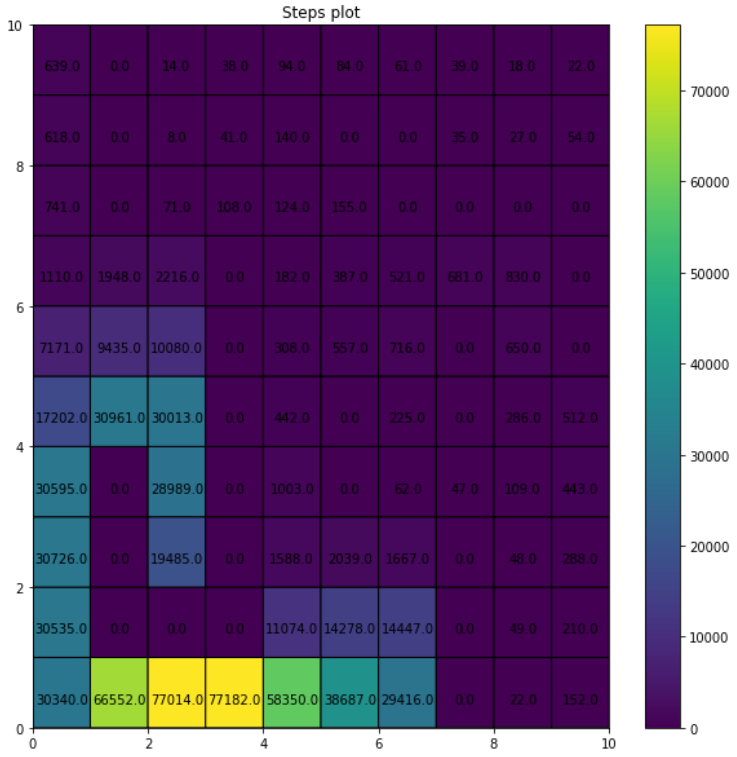
Best chosen hyperparameters:

- $\beta = 3$
- $\alpha = 0.1$
- $\Upsilon = 0.93$









Inferences: Almost no change with its epsilon version. Gamma had to be reduced by only 1 percent to get a closer value with its epsilon version. Softmax in these cases fared worse than the epsilon.

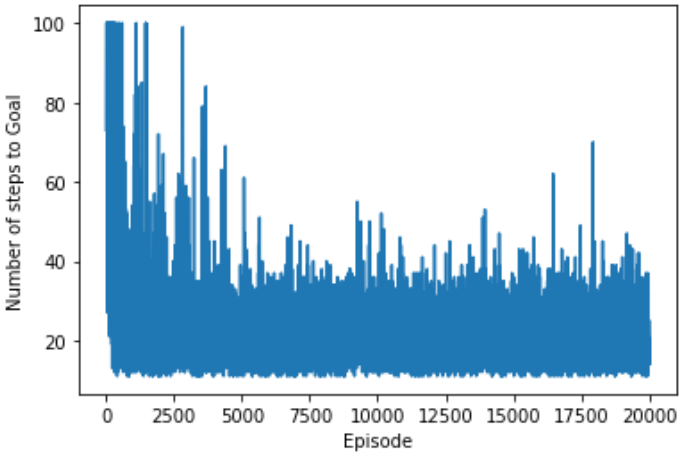
Combination 8:

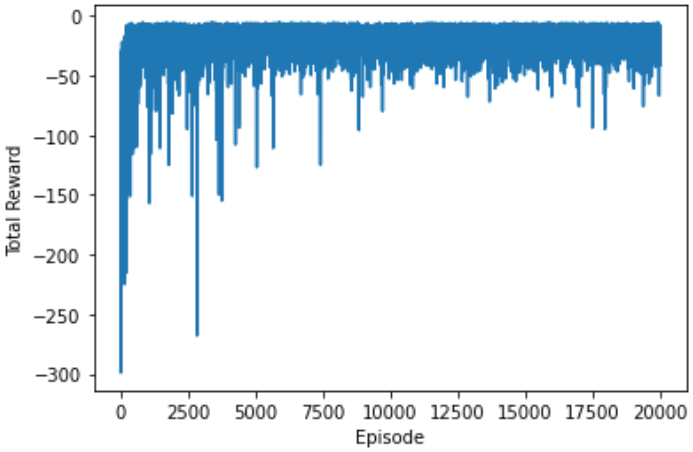
- Policy: Softmax
- Start State: [0, 4]

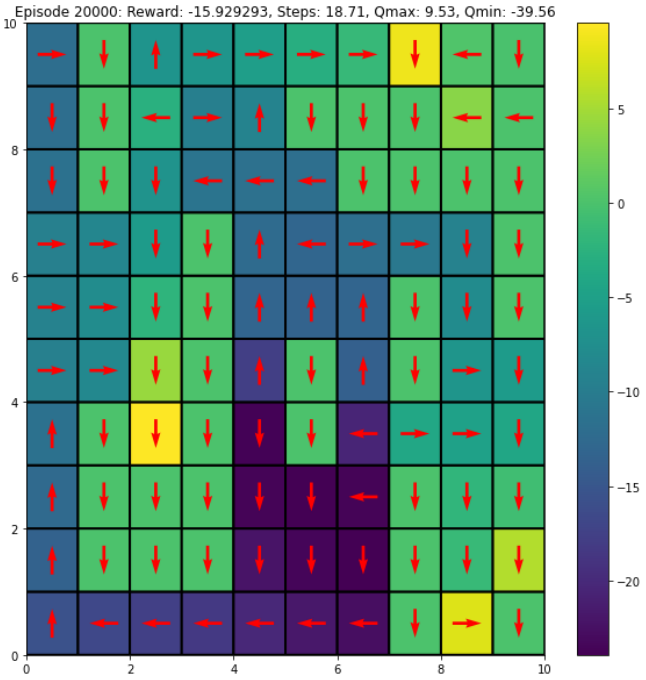
- P = 0.7
- Wind = False

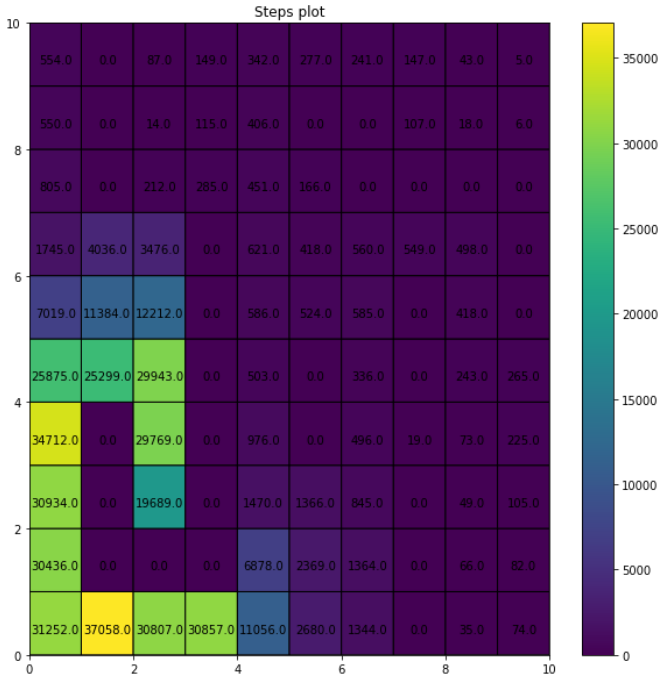
Best chosen hyperparameters:

- β = 3
- α = 0.05
- ϒ = 0.94









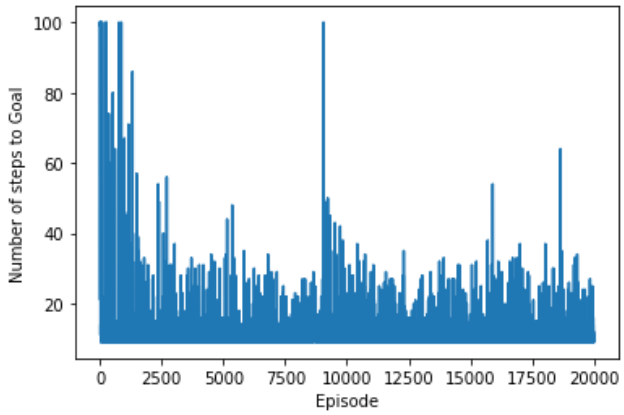
Inferences: Not much difference in the step and reward graph with the epsilon version. Beta had to be reduced from earlier 5 to 3. Alpha had to be reduced to bits as taking too much of the new state values were making the graphs go haywire from 0.1 to 0.05.

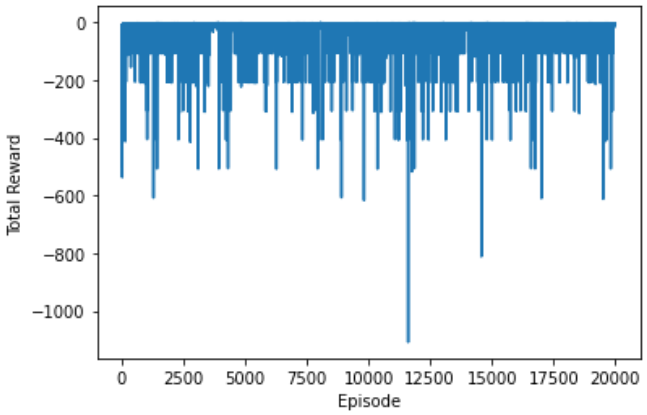
Combination 9:

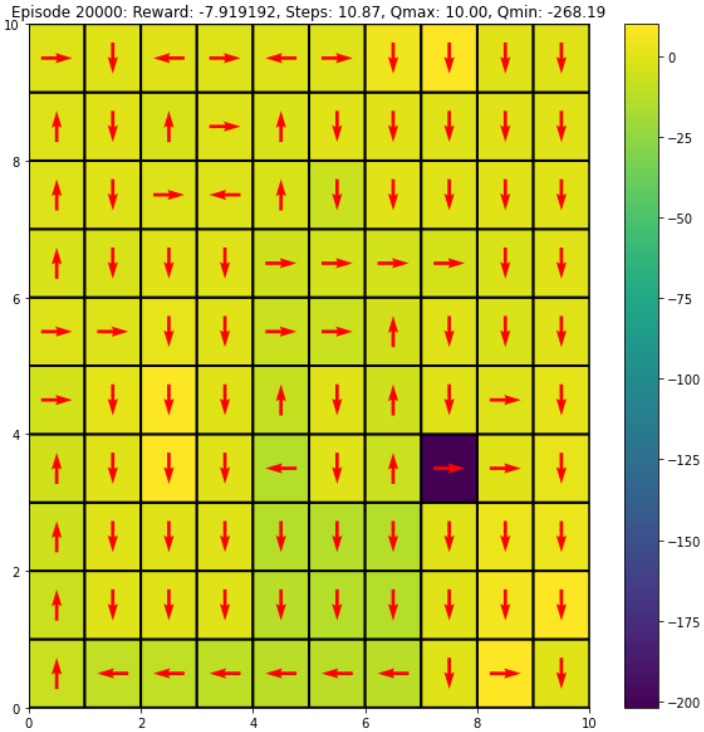
- Policy: Epsilon Greedy
- Start State: [3, 6]
- P = 1
- Wind = True

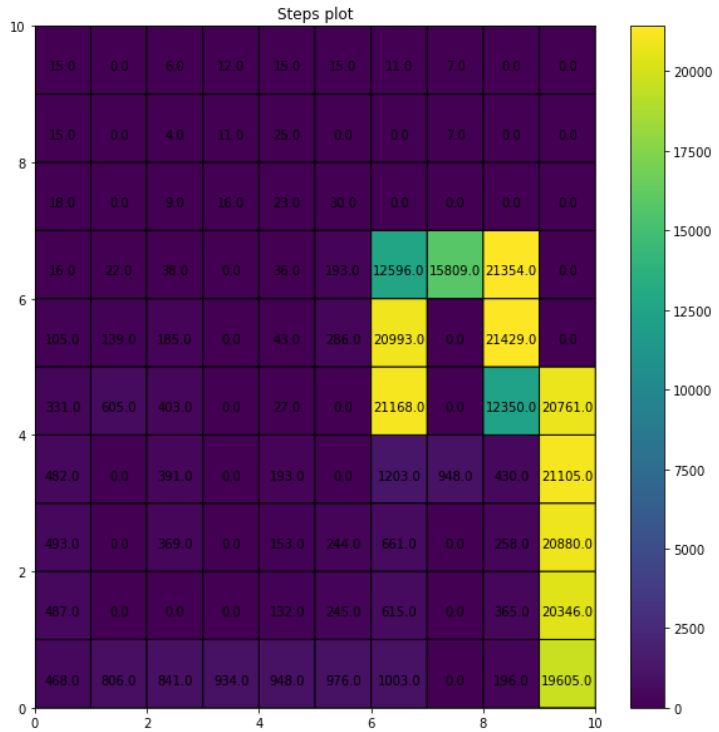
Best chosen hyperparameters:

- ε = 0.063
- α = 0.4
- Ɣ = 0.94









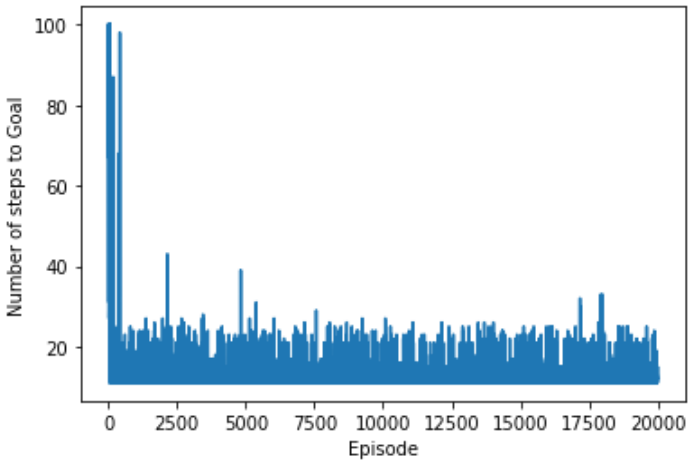
Inference: For the start state 3,6 we are reaching 0,9 goal state. As usual, the graphs tend to get better with p=1 than p=0.7. There is also a faint path from 3,6 to 2,2 showing that there were very few cases where the agent decided to go for unknown territory and also reach a goal state.

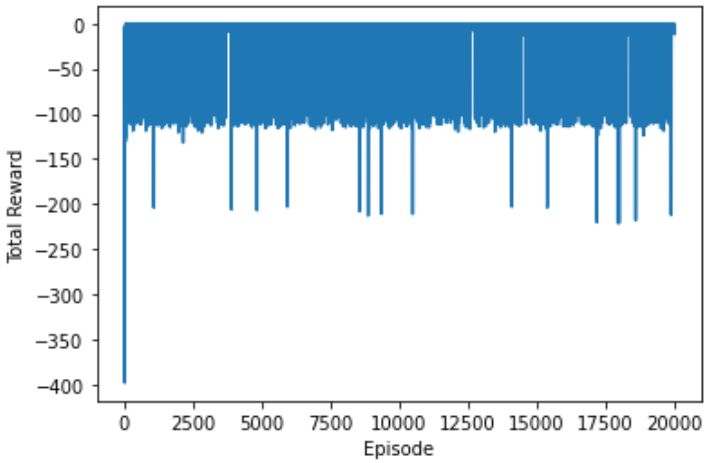
Combination 10:

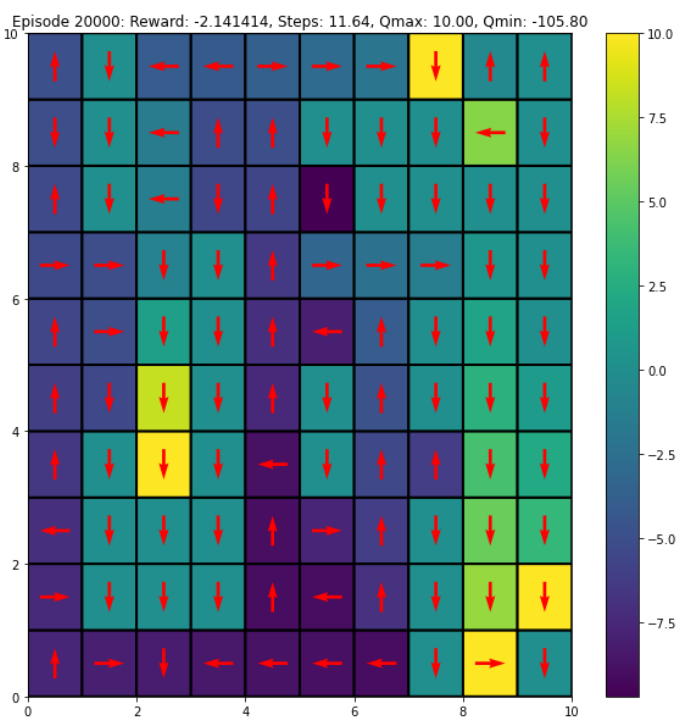
- Policy: Epsilon Greedy
- Start State: [3, 6]
- P = 1
- Wind = False

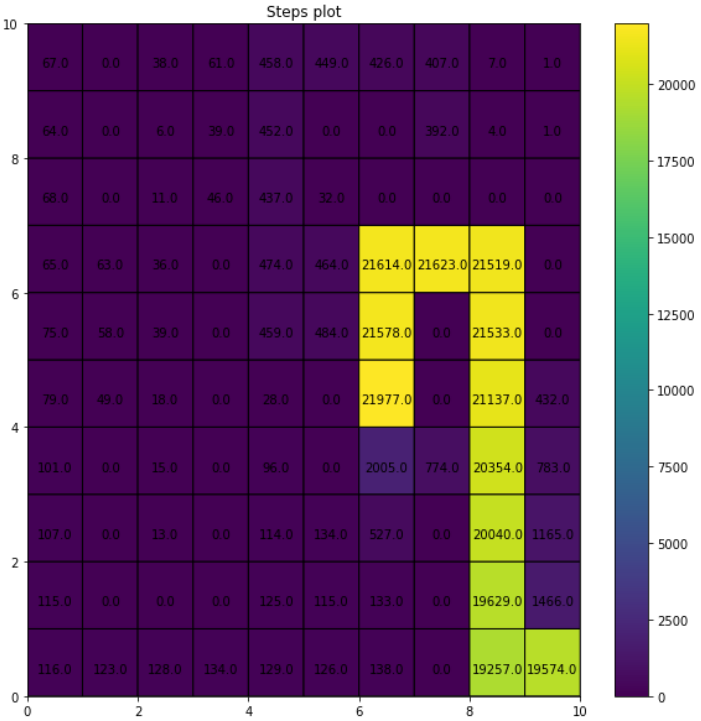
Best chosen hyperparameters:

- ε = 0.073
- α = 0.4
- ϒ = 0.94









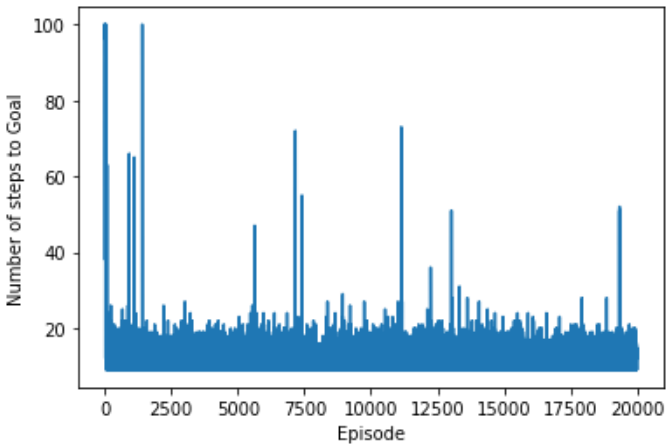
Inferences: Epsilon increased to 0.073 from the windy version. The graph becomes much more stable than the windy one with much lesser fluctuations.

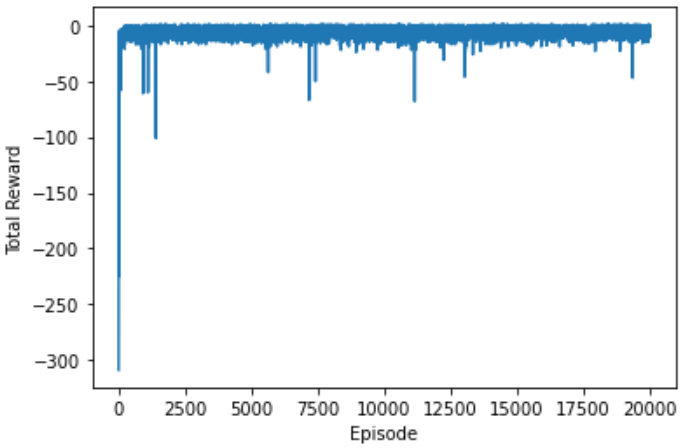
Combination 11:

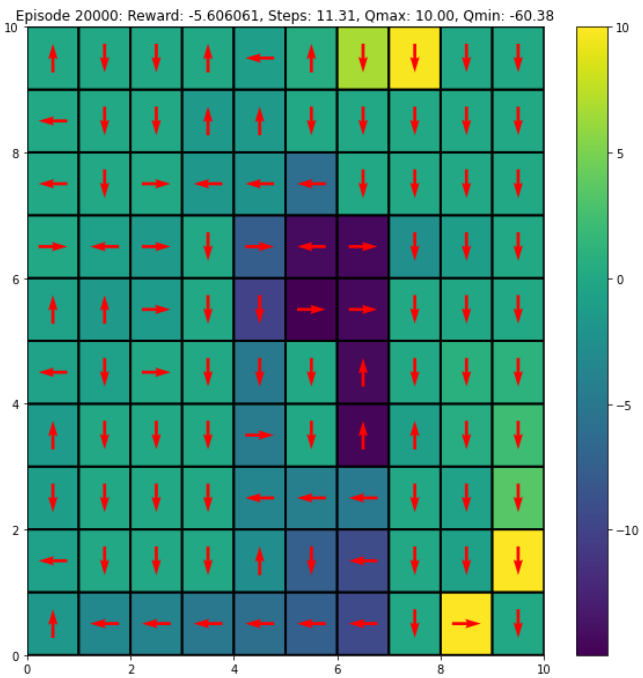
- Policy: Softmax
- Start State: [3, 6]
- P = 1
- Wind = True

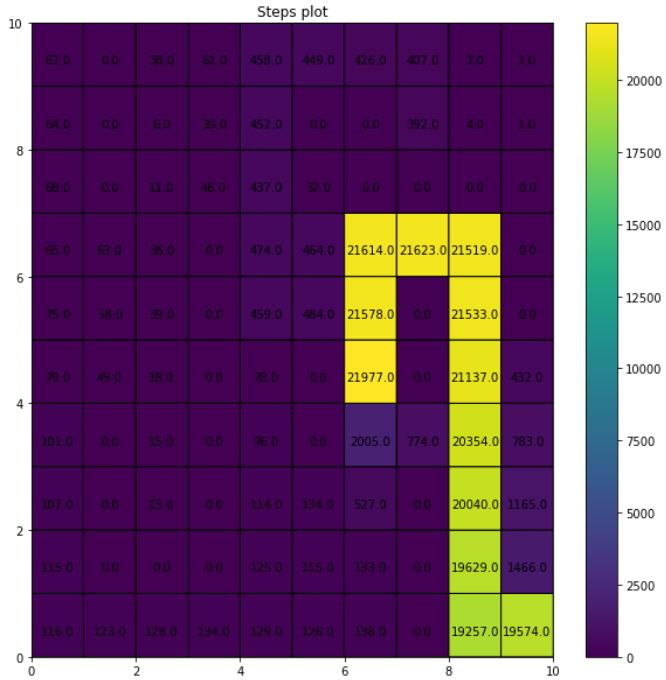
Best chosen hyperparameters:

- $\beta = 5$
- $\alpha = 0.4$
- $\Upsilon = 0.94$









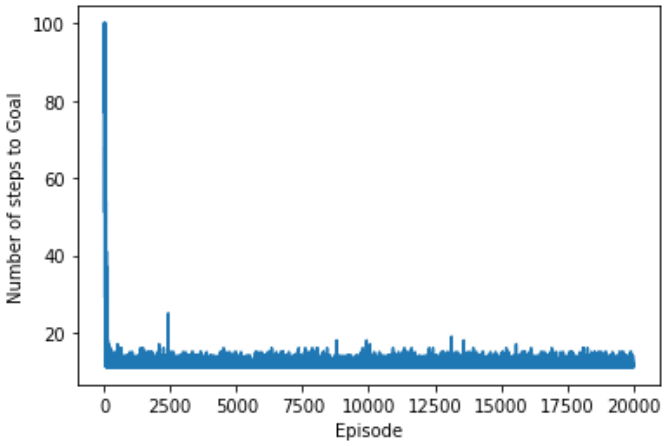
Inference: HUGELY better than epsilon greedy for the same condition when referred to combination 9. Beautiful graph. Only a few fluctuations because of the wind which corrects itself really fast. Since the earlier hyperparameters worked quite well we did not change much.

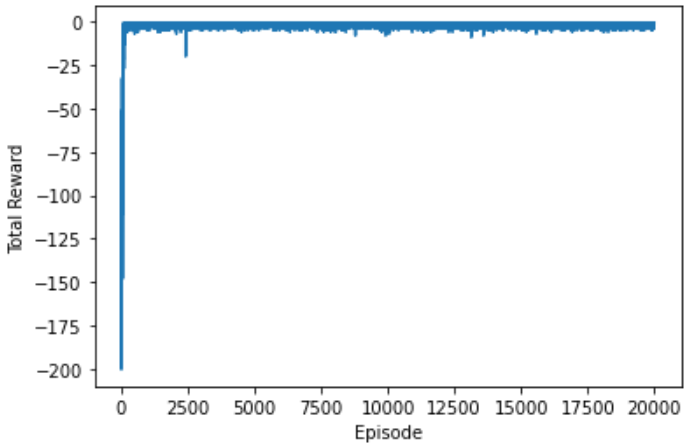
Combination 12:

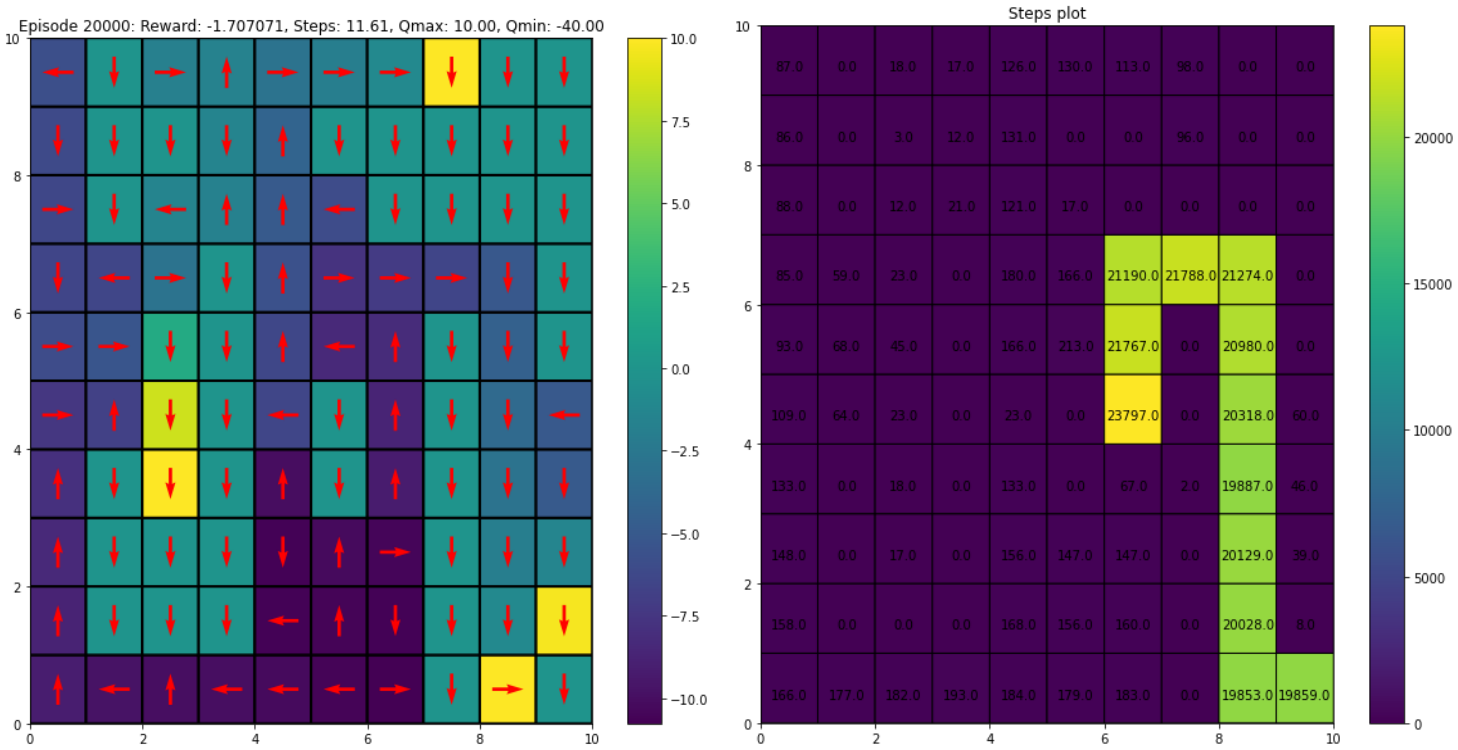
- Policy: Softmax
- Start State: [3, 6]
- P = 1
- Wind = False

Best chosen hyperparameters:

- β = 5
- α = 0.4
- ϒ = 0.94







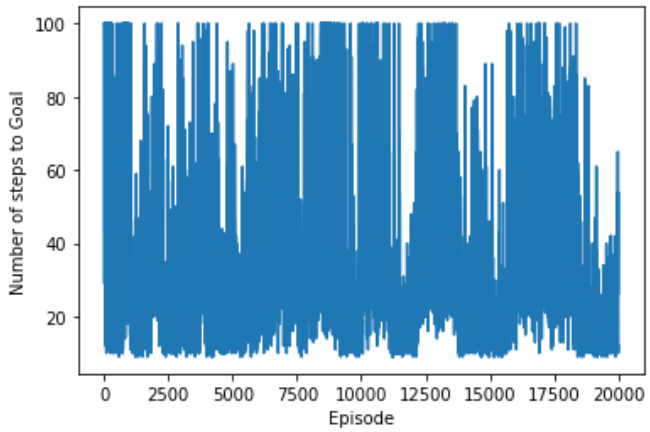
Inferences: The windless version is almost perfect. Converges really fast and stays there. Beats epsilon greedy by a huge margin.

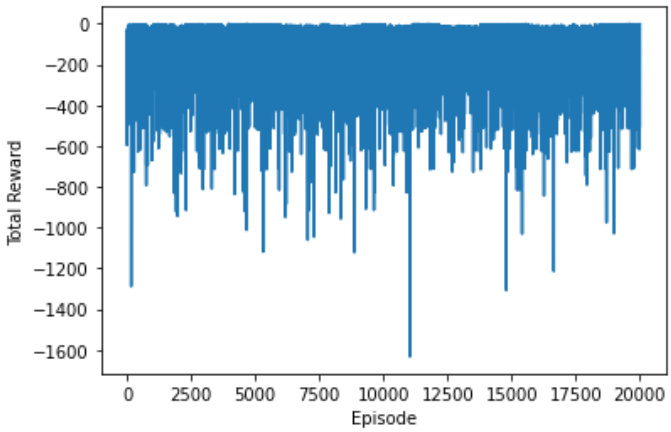
Combination 13:

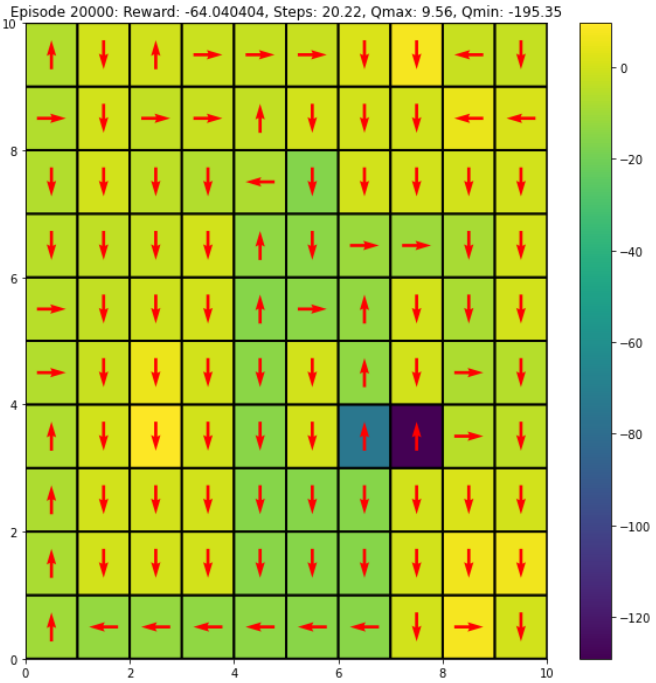
- Policy: Epsilon Greedy
- Start State: [3, 6]
- P = 0.7
- Wind = True

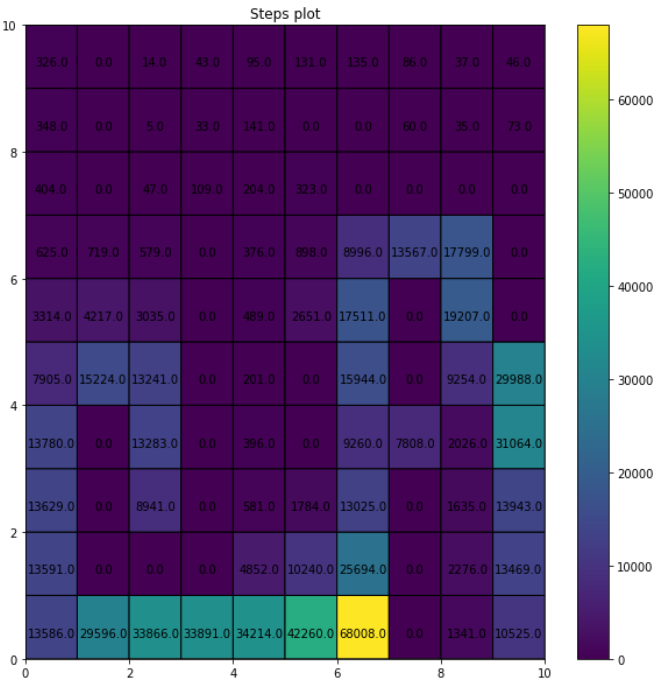
Best chosen hyperparameters:

- ε = 0.03
- α = 0.1
- ϒ = 0.94









Inferences: Here comes the infamous p=0.7 again. Both the goal states 0,9 and 2,2 were reached most of the time because of the non-determinism. Both alpha(0.4 to 0.1) and epsilon(0.063 to 0.03) needed to be hugely changed to even get something productive. One of the lowest rewards we got. Either we couldn't tune the hyperparameters or this combination is not a good fit for this problem.

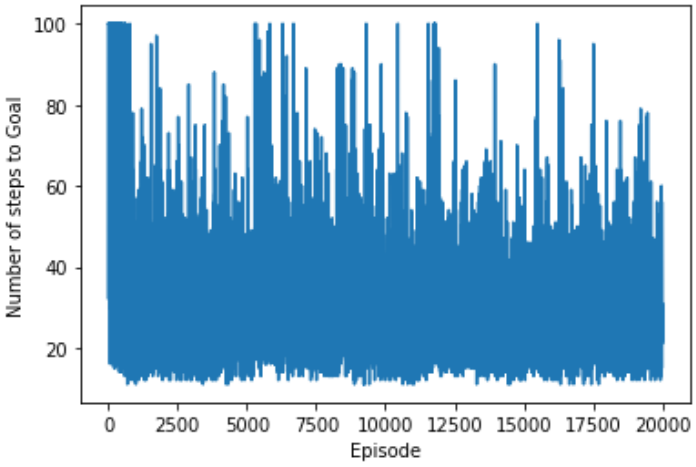
Combination 14:

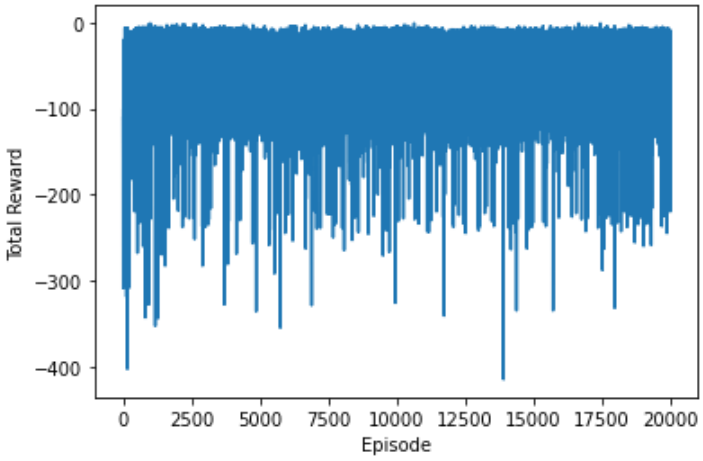
- Policy: Epsilon Greedy
- Start State: [3, 6]
- P = 0.7
- Wind = False

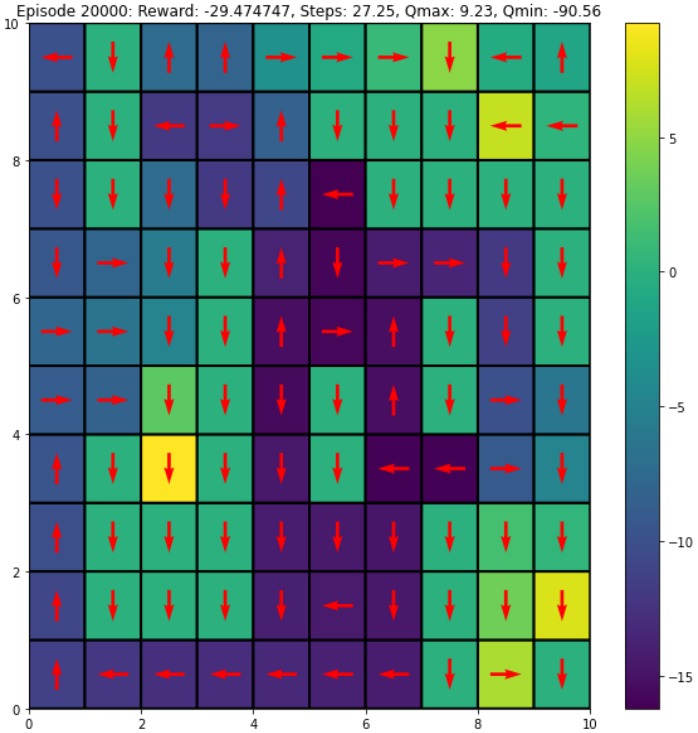
Best chosen hyperparameters:

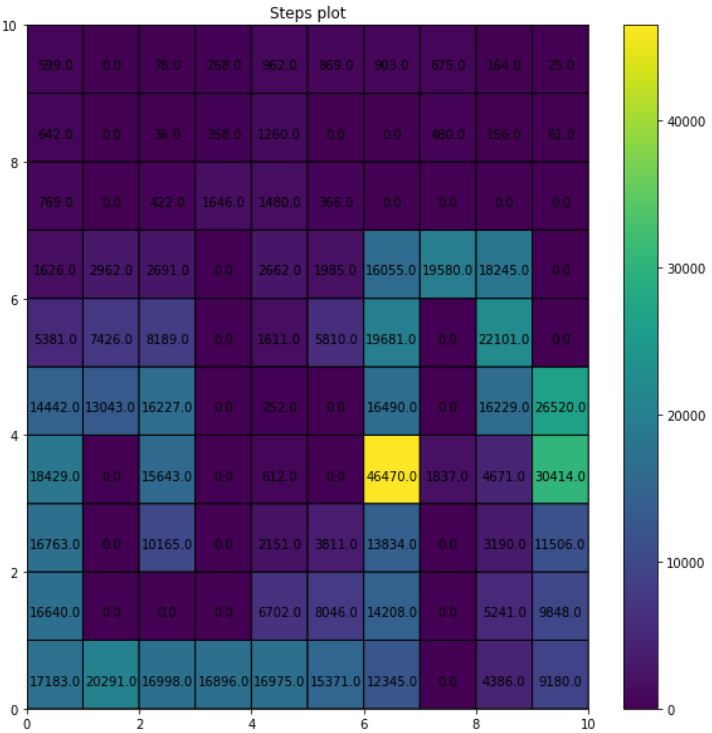
- ε = 0.1
- α = 0.1

- ϒ = 0.94









Inferences: A little better than the windy version. But only by a small margin. Epsilon was giving bad results with its earlier values, so needed to be increased to 0.1 from 0.03,0.063 or 0.073.

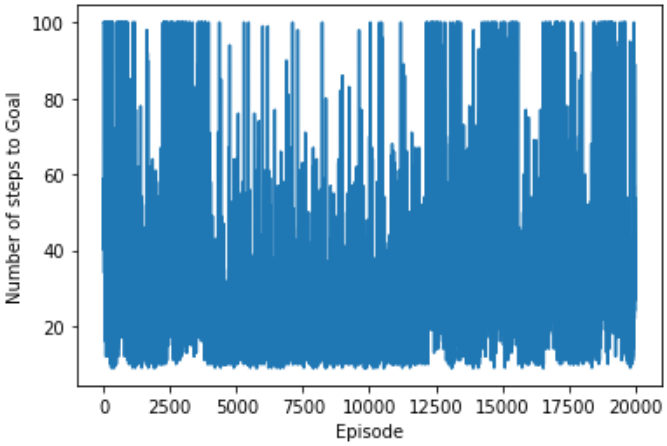
Combination 15:

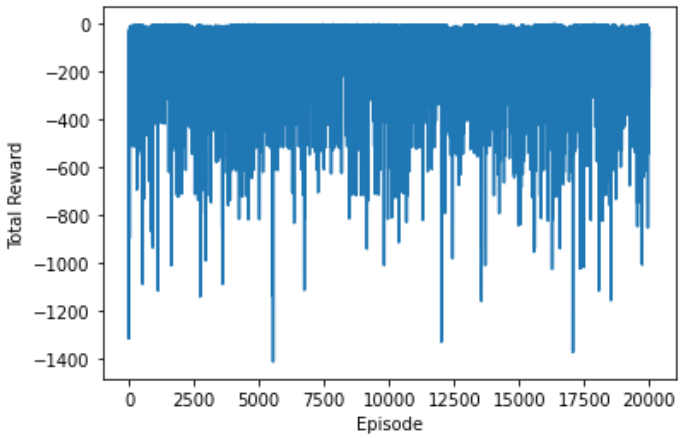
- Policy: Softmax
- Start State: [3, 6]
- P = 0.7
- Wind = True

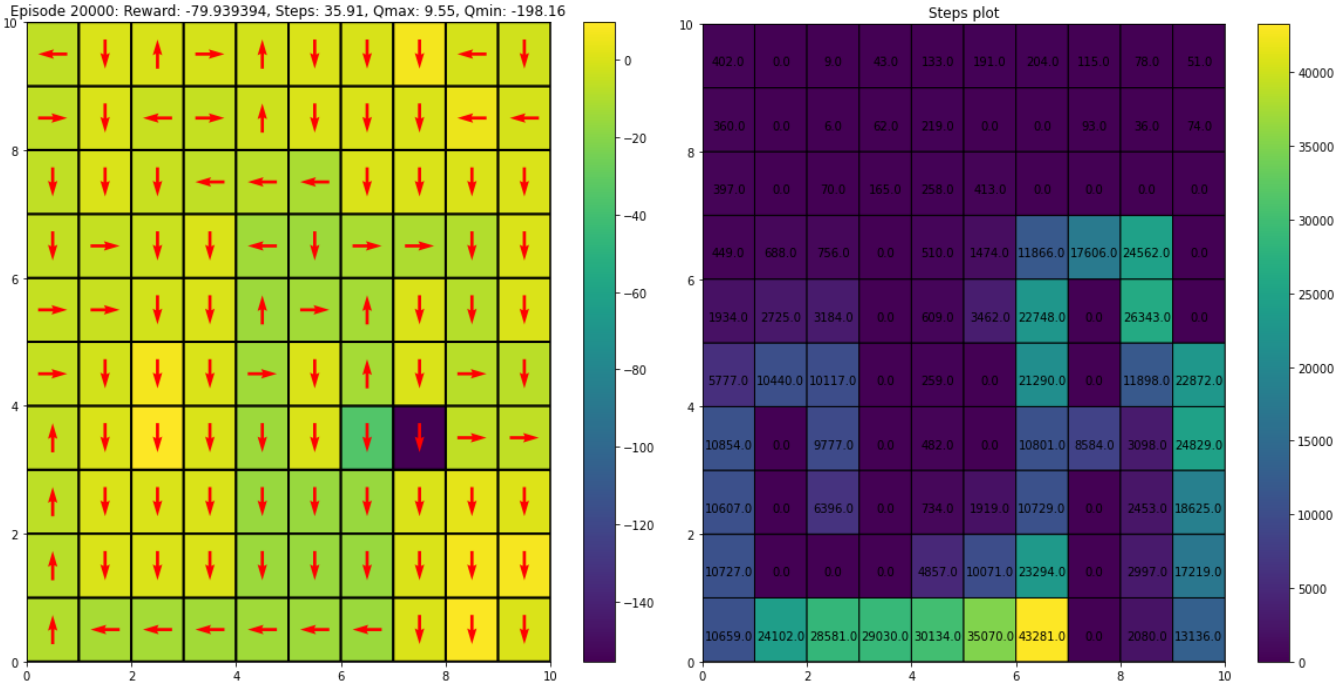
Best chosen hyperparameters:

- β = 0.1

- α = 0.1
- ϒ = 0.94







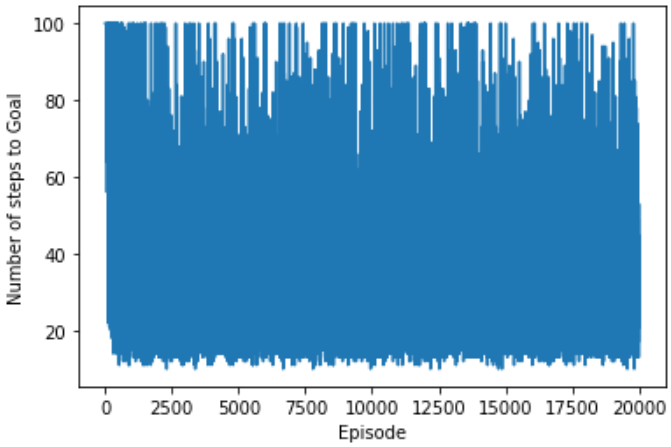
Inferences: Another bad combination. Might be giving a better average in some of the experiments but is not consistent. A wide range of hyperparameters was tested but just might have missed a sweet spot. Beta(0.1 to 30), alpha(0.7 to 0.1) and gamma(0.88 to 0.97). This was the lowest we could get after countless experiments.

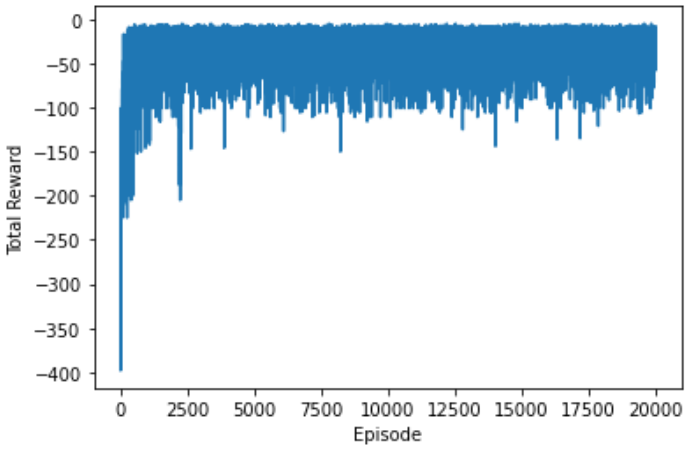
Combination 16:

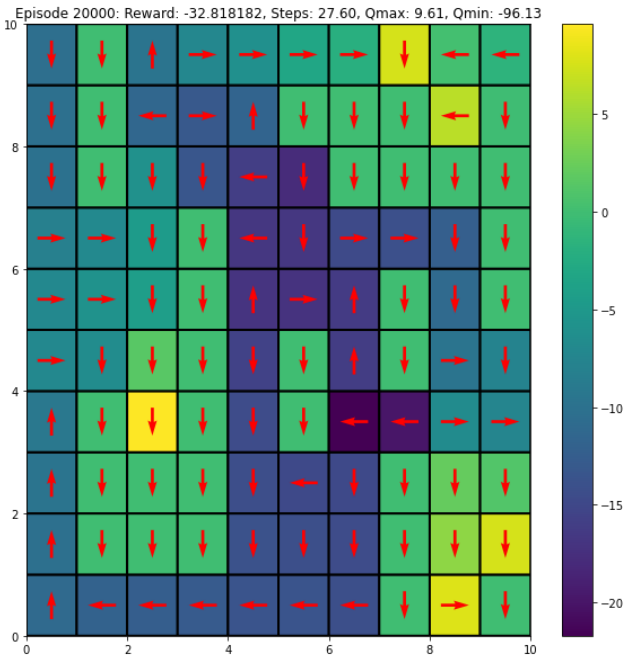
- Policy: Softmax
- Start State: [3, 6]
- P = 0.7
- Wind = False

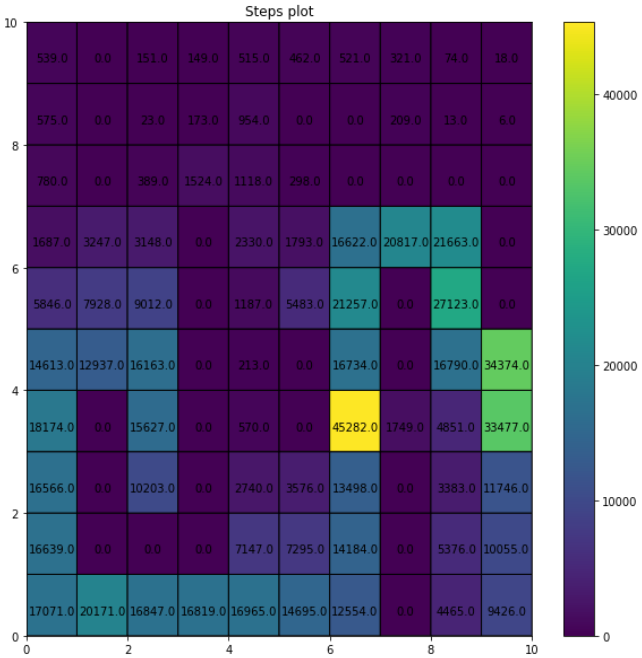
Best chosen hyperparameters:

- β = 4
- α = 0.12
- ϒ = 0.94









Inference: The step graph is very bad but the reward one looks much better than the others. Shows how much the windy version can mess things up on top of the non-deterministic action choosing phenomenon.

SAARSA:

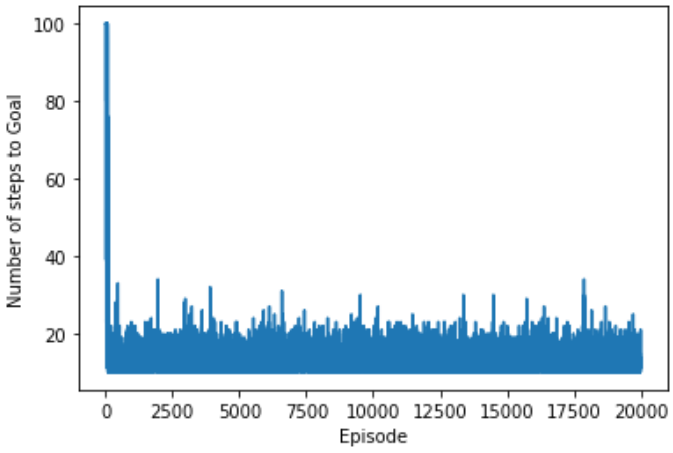
Below are the plots corresponding to each of the 16 combinations. The plots have been generated by taking an average of 3 independently run experiments.

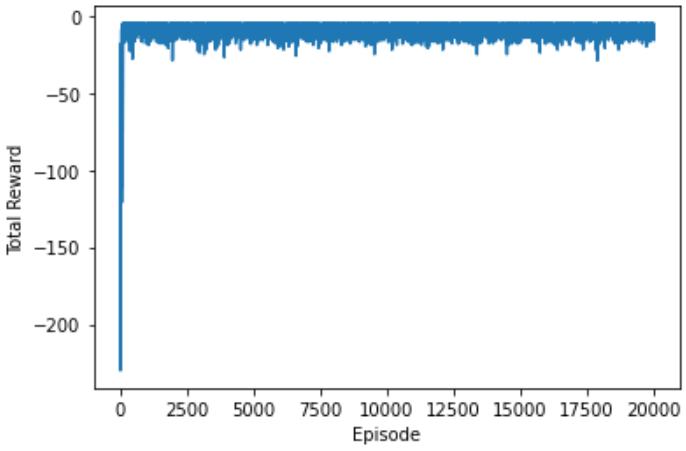
Combination 1:

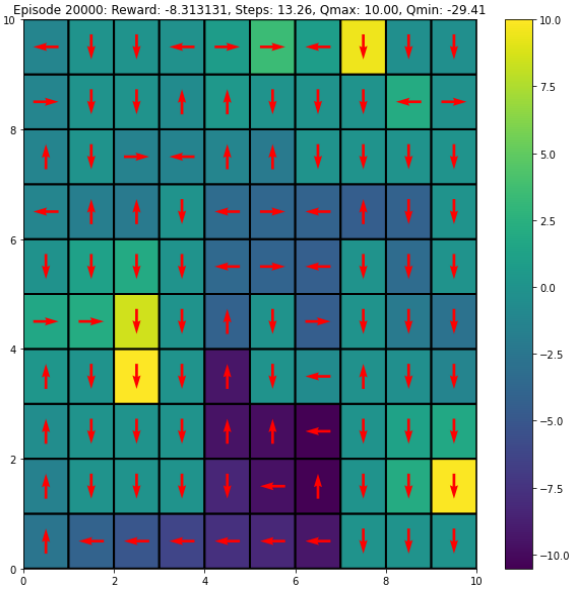
- Policy: Epsilon Greedy
- Start State: [0, 4]
- P = 1
- Wind = True

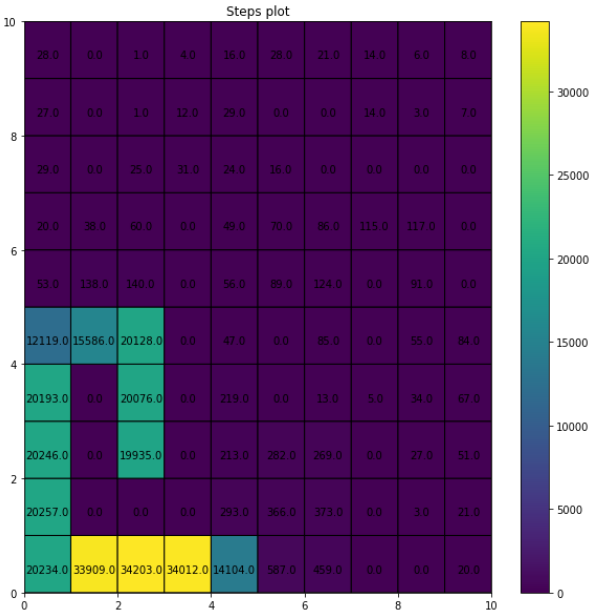
Best chosen hyperparameters:

- ε = 0.01
- α = 0.95
- ϒ = 0.94









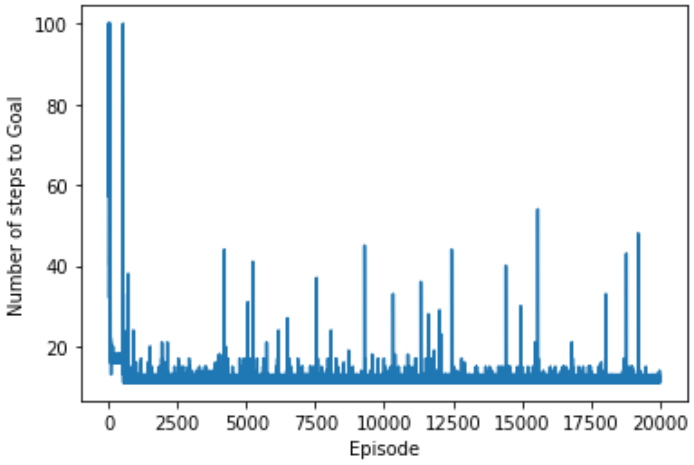
Inference: A usual p=1 works really well here, and exceptionally so when compared to Qlearning under the same conditions. Epsilon had to be reduced to get this while in Qlearning it was much higher. Alpha here is also much higher than only 0.7 or 0.4 in the case of Qlearning. As expected of the best conditions.

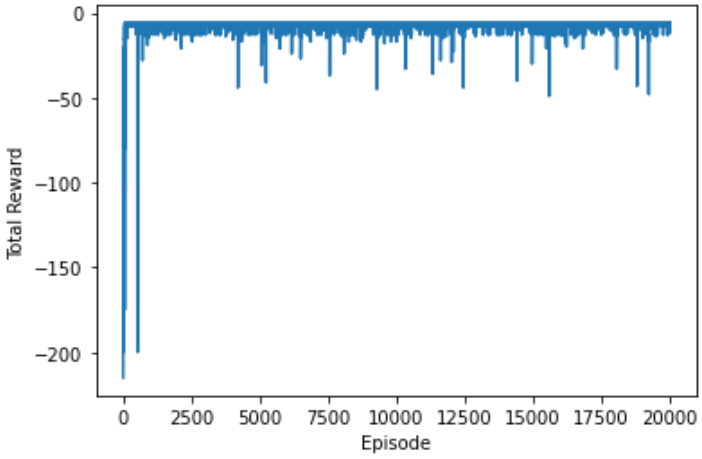
Combination 2:

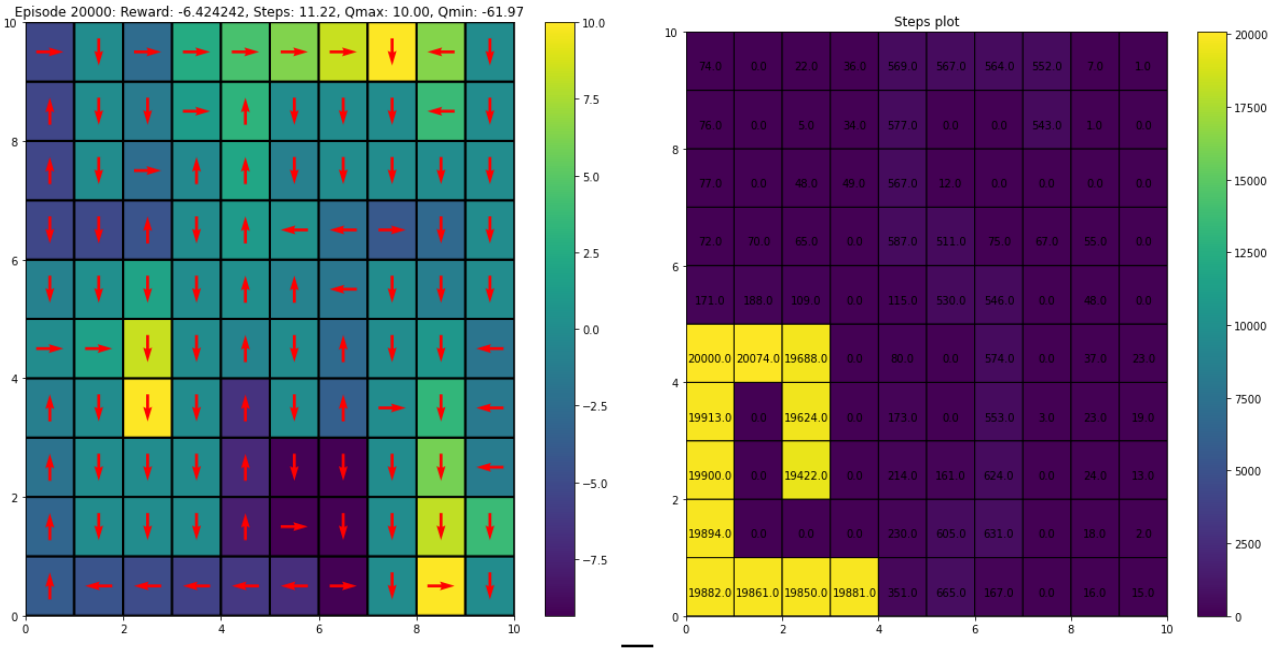
- Policy: Epsilon Greedy
- Start State: [0, 4]
- P = 1
- Wind = False

Best chosen hyperparameters:

- ε = 0.013
- α = 0.38
- ϒ = 0.94







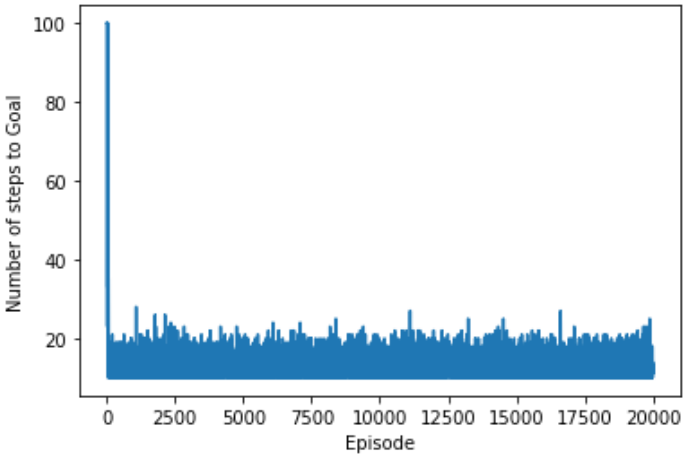
Inference: The absence of wind increased the rewards further and faster convergence. There were a few explorations which were corrected instantly and the range over which the rewards varied stayed very low. Little fine tuning was done after the second decimal places to get the best value.

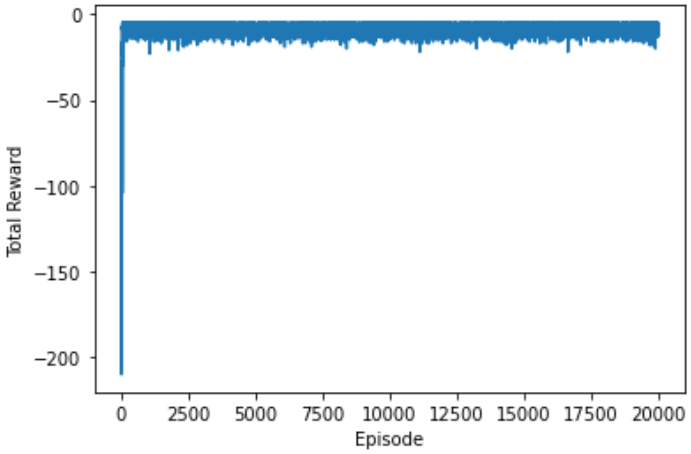
Combination 3:

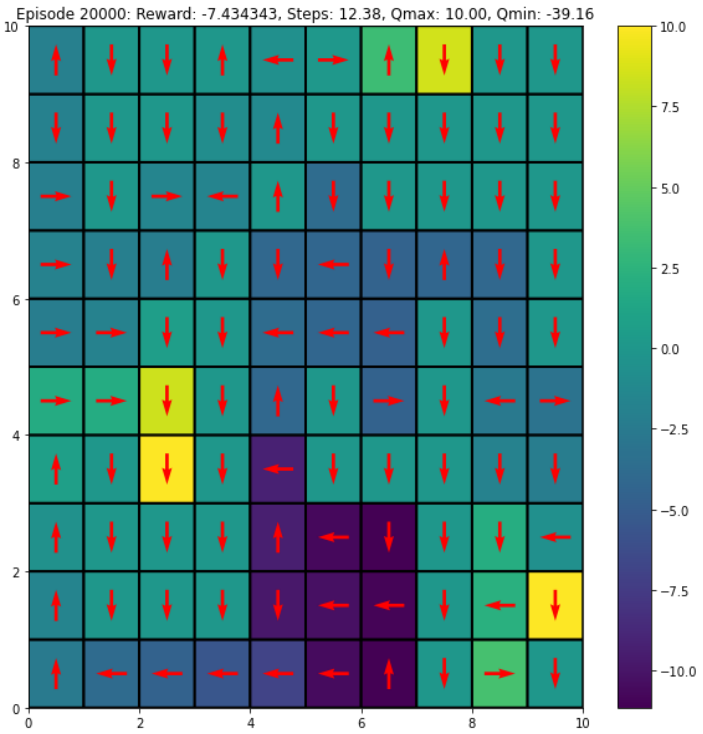
- Policy: Softmax
- Start State: [0, 4]
- P = 1
- Wind = True

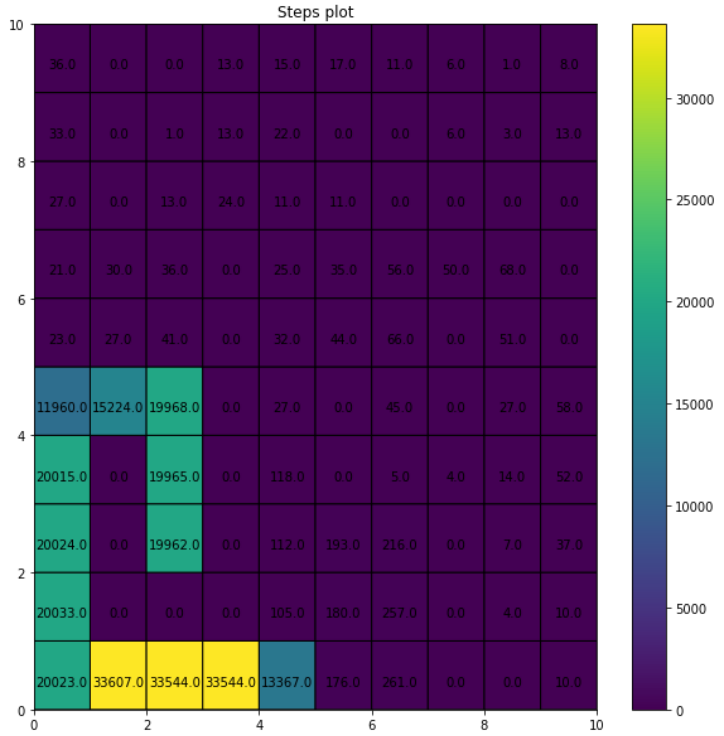
Best chosen hyperparameters:

- β = 5
- α = 0.38
- ϒ = 0.94









Inference: Softmax as usual made the graphs stable with a lesser number of irregularities. The range between the max steps and min steps required stayed comparatively low and there were very few episodes that reached a different goal state other than 2,2. Almost the same parameters give quite a different experience than Qlearning, showing softmax is much more compatible with Saarsa than Qlearning.

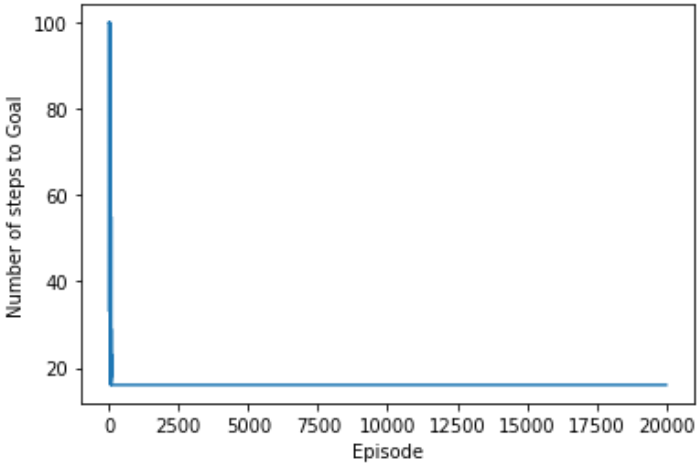
Combination 4:

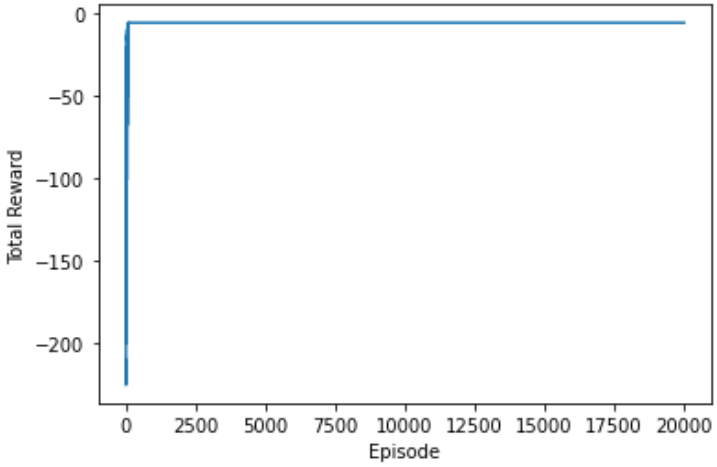
- Policy: Softmax
- Start State: [0, 4]
- P = 1

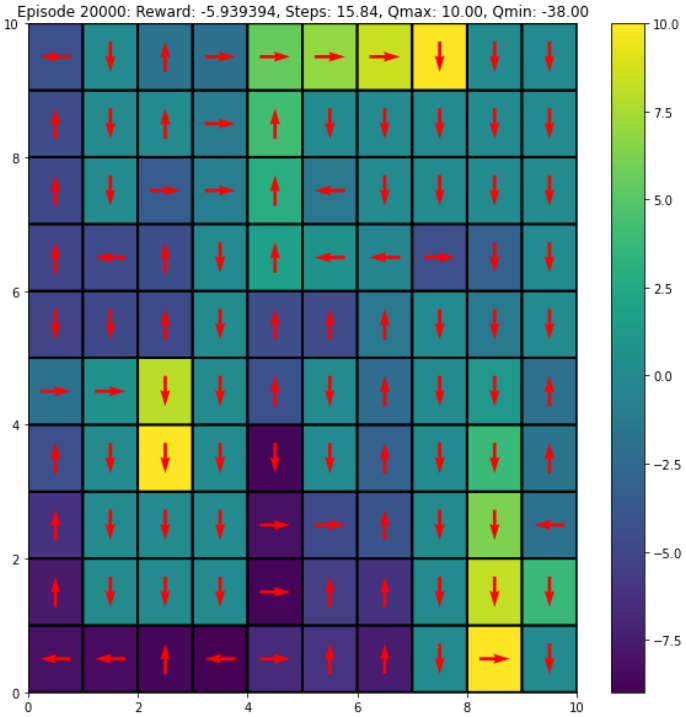
- Wind = False

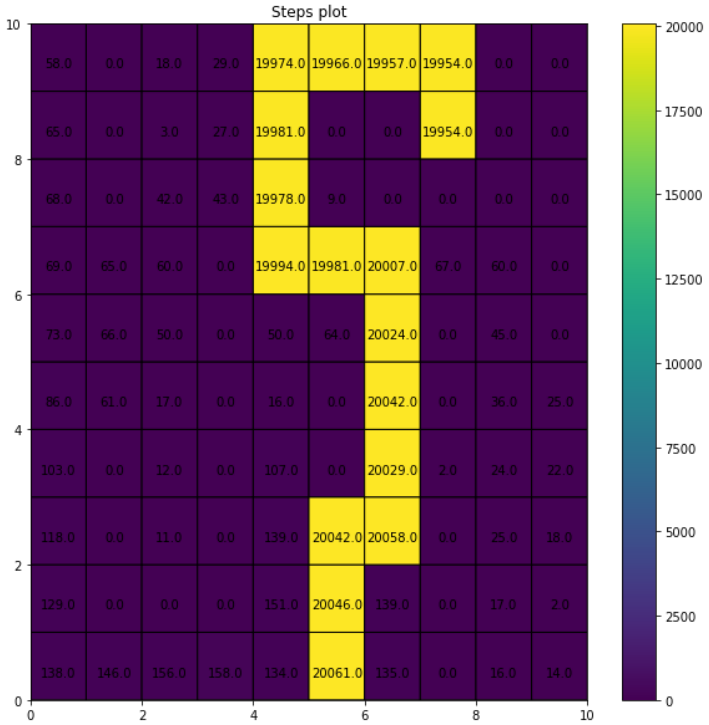
Best chosen hyperparameters:

- β = 5
- α = 0.38
- ϒ = 0.94









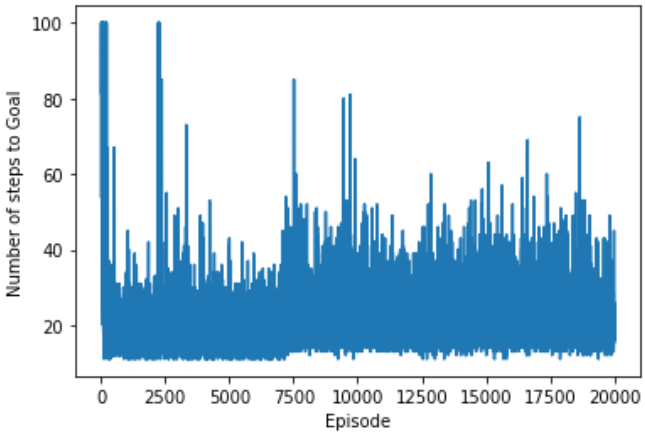
Inferences: Almost perfect graph. Best graph among the lot. With almost the same parameters as Q-learning. Making this combination one of the ideal situations.

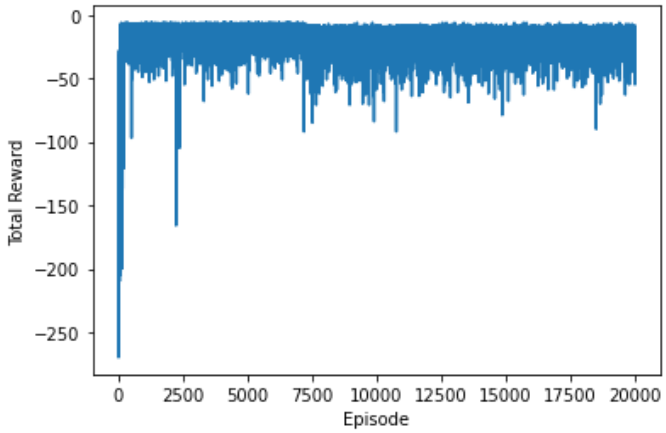
Combination 5:

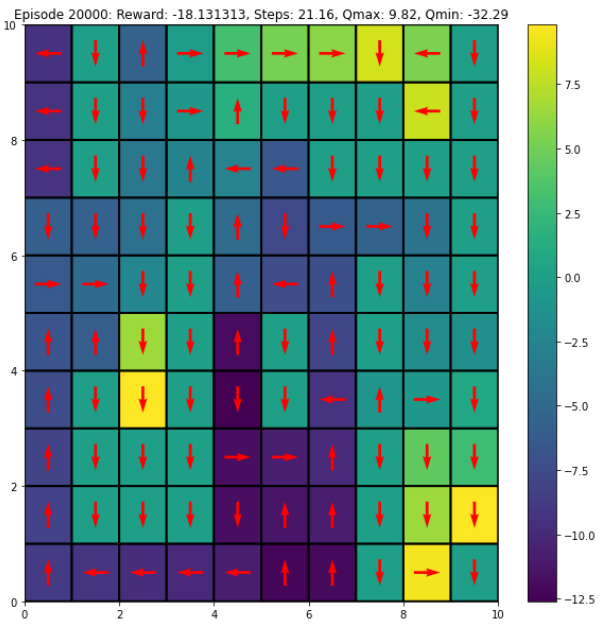
- Policy: Epsilon Greedy
- Start State: [0, 4]
- P = 0.7
- Wind = True

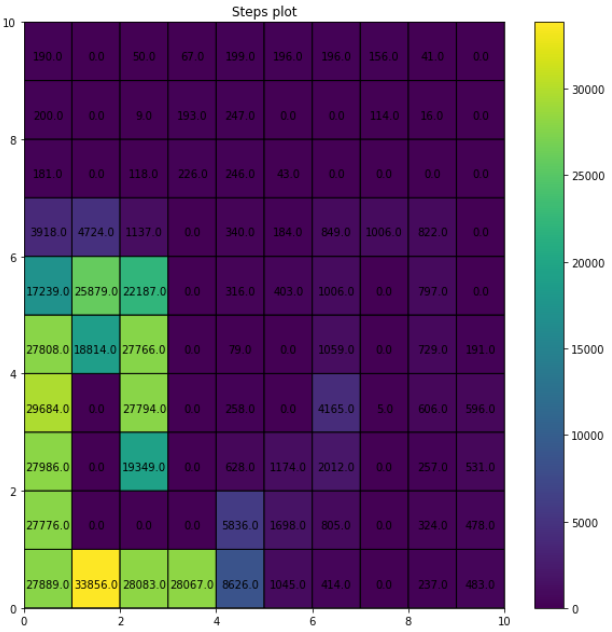
Best chosen hyperparameters:

- ε = 0.01
- α = 0.3
- ϒ = 0.94









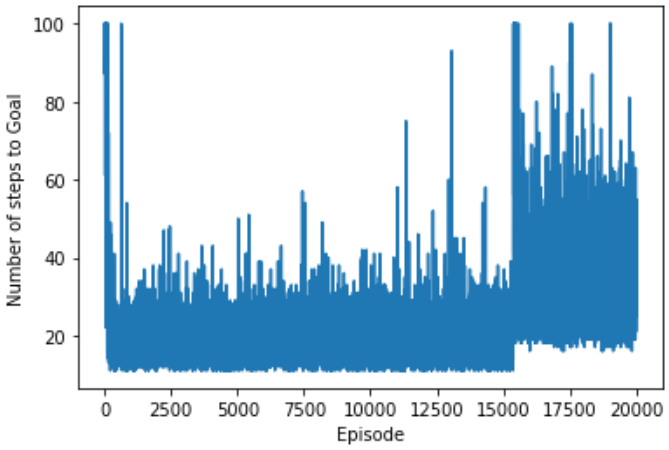
Inferences: Even with p=0.7 this combination gives much better results than Q-learning. The steps graph is thinner and the max reward range is reached much faster. Epsilon had to be changed a lot from 0.063 to 0.1. Reducing alpha also worked for Saarsa.

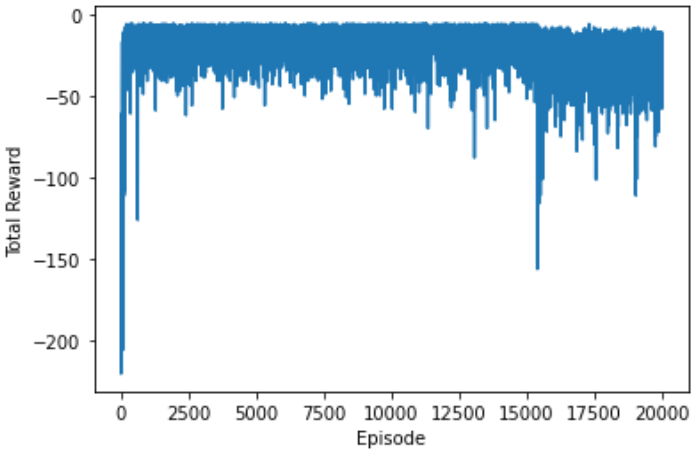
Combination 6:

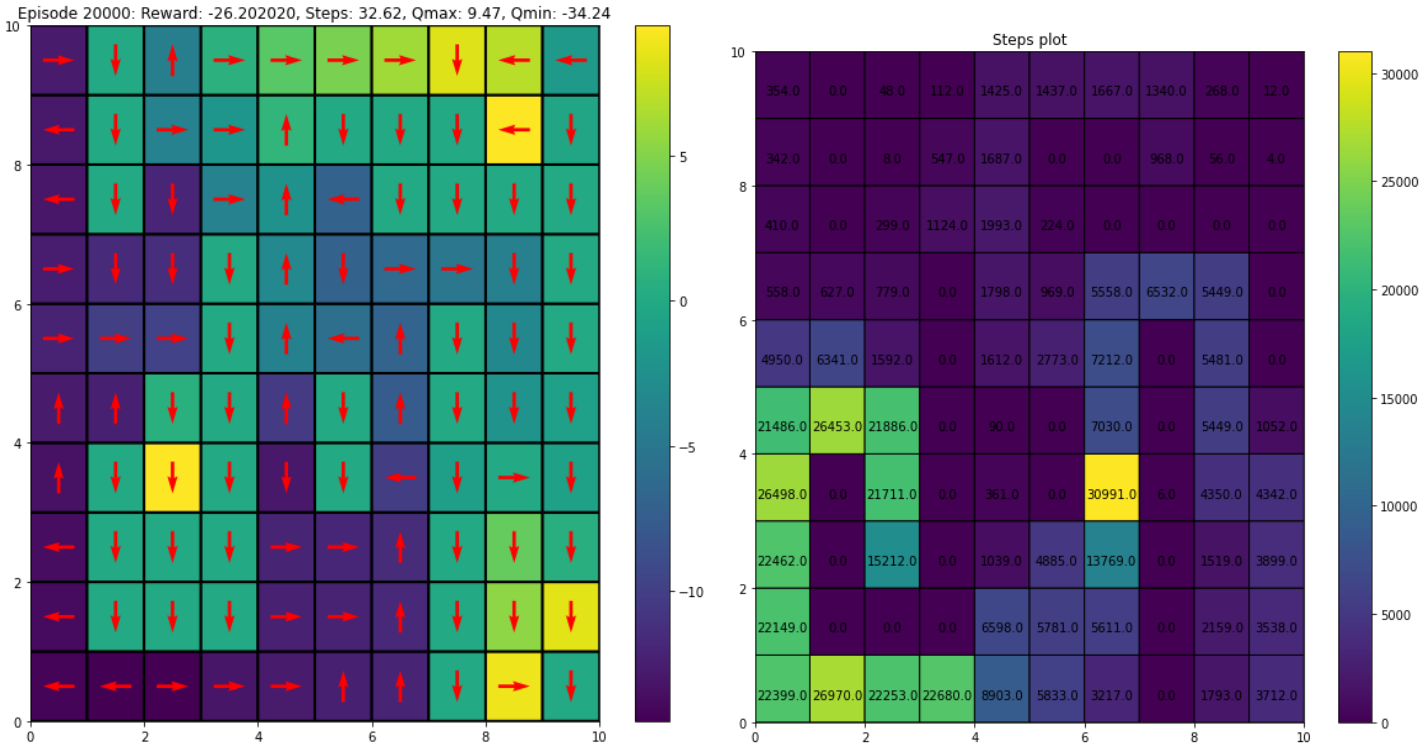
- Policy: Epsilon Greedy
- Start State: [0, 4]
- P = 0.7
- Wind = False

Best chosen hyperparameters:

- ε = 0.01
- α = 0.3
- ϒ = 0.94







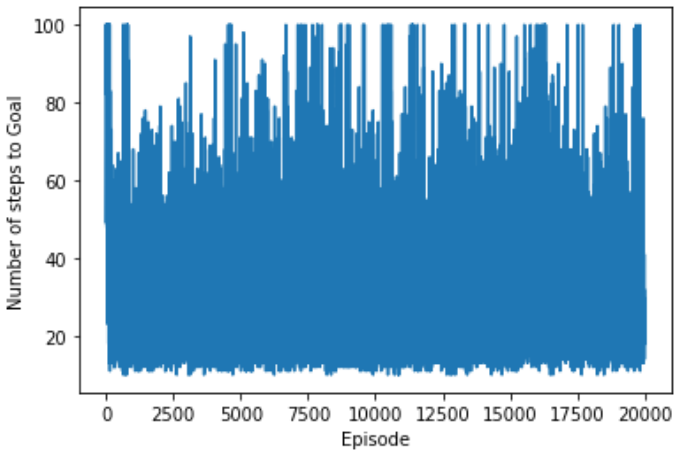
Inferences: With p=0.7 this combination, the results are deteriorating. Getting an average reward of -26.20 and average steps as 32.62. Goal states (2, 2) and (0,9) are reached most of the time.

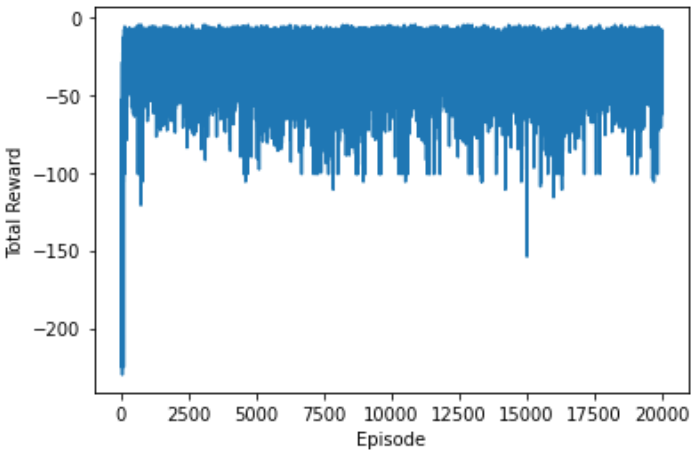
Combination 7:

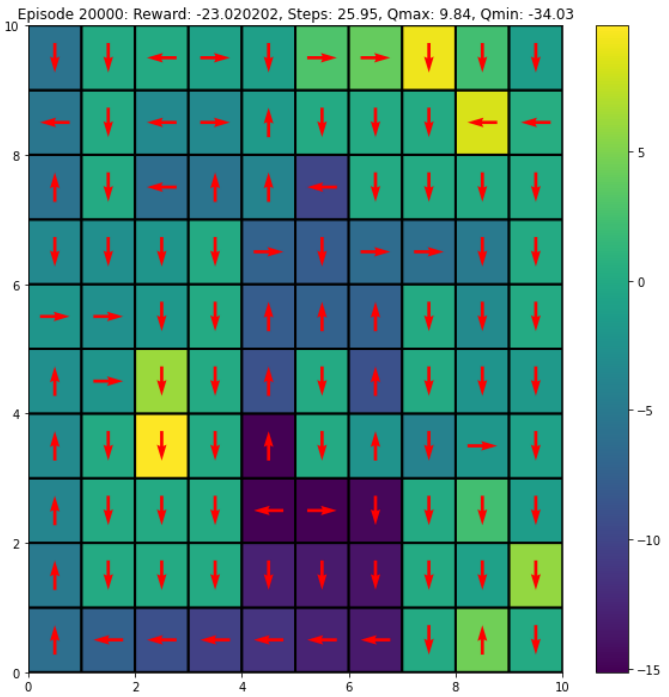
- Policy: Softmax
- Start State: [0, 4]
- P = 0.7
- Wind = True

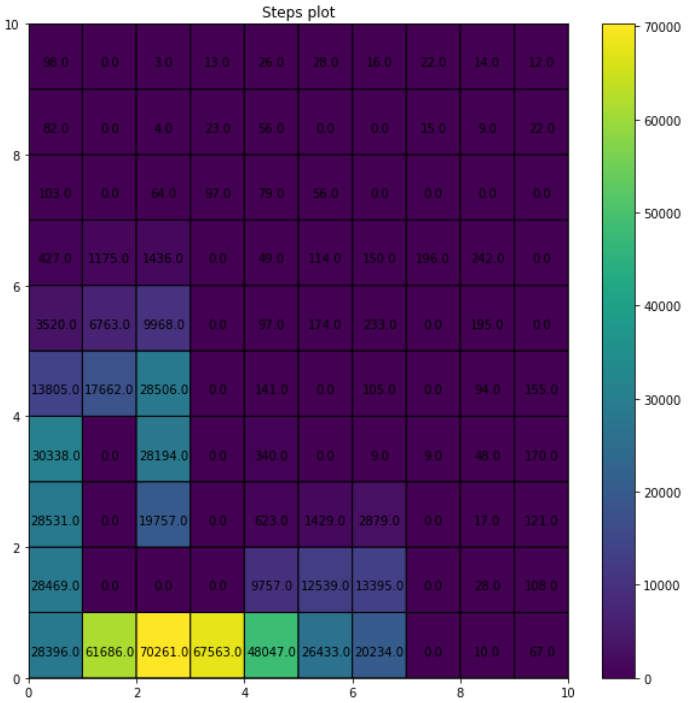
Best chosen hyperparameters:

- β = 5
- α = 0.3
- ϒ = 0.94









Inferences: Tuning this was relatively hard as minute differences were making the results go haywire. After choosing multiple combinations and permutations with previous hyperparameters we deduced it was best to stick with the ones that were being used in SAARSA. Still giving better results than Qlearning. Beta was chosen over a variety of ranges but the first one=5 gave a result we could go with. All in all the presence of wind keeps changing the results in quite an unpredictable manner as in each experiment it might take various steps taking it into account.

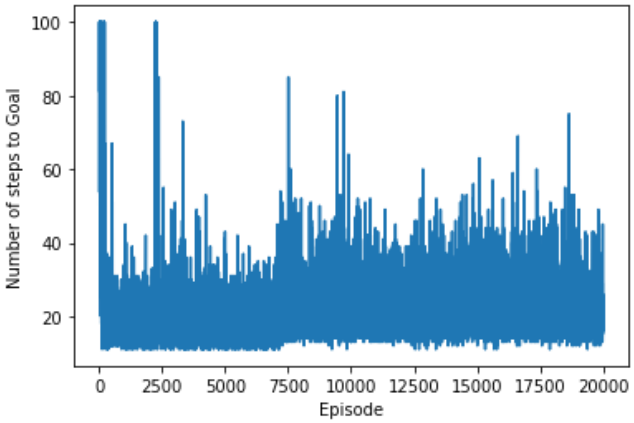
Combination 8:

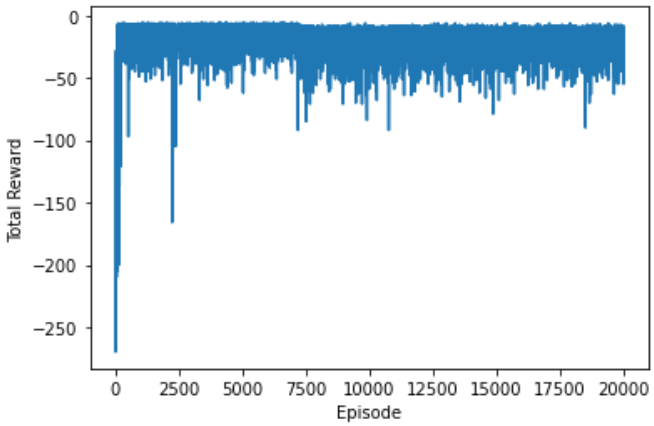
- Policy: Softmax
- Start State: [0, 4]
- P = 0.7

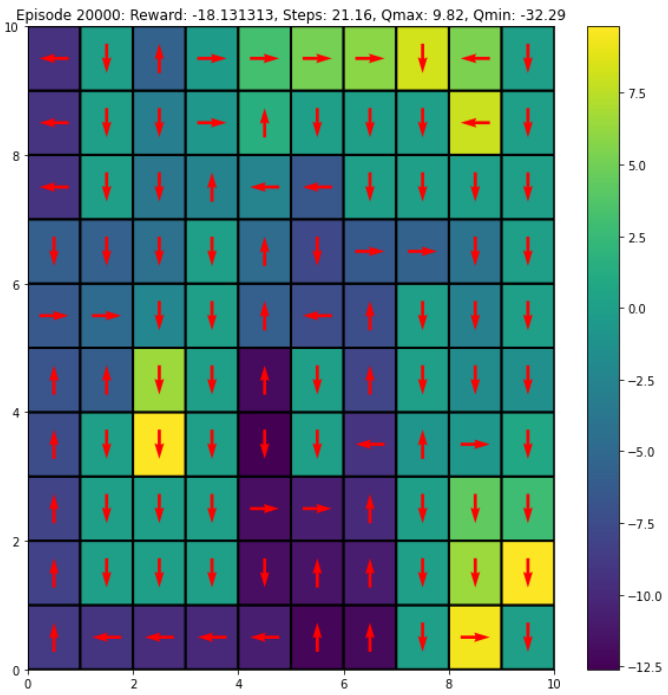
- Wind = False

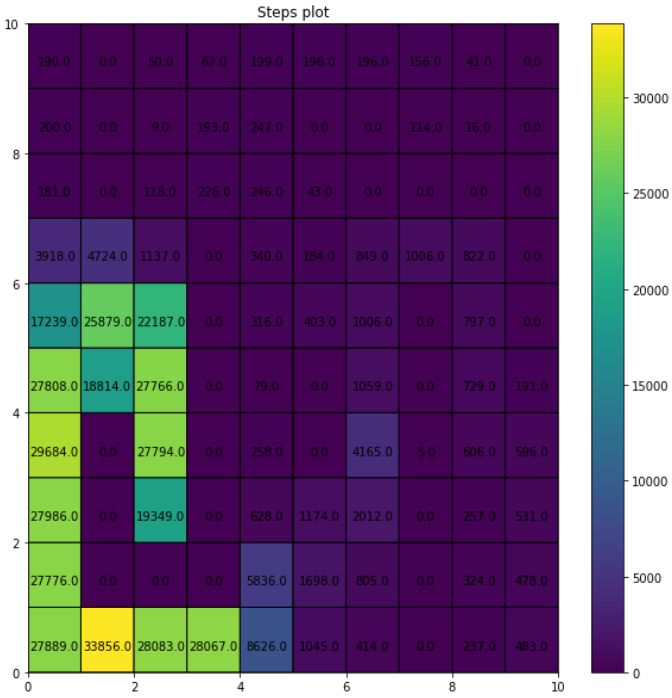
Best chosen hyperparameters:

- β = 5
- α = 0.3
- ϒ = 0.94









Inferences: With **wind = False**, the results are slightly improving. Getting an average reward of -18.13 and average steps as 21.16.

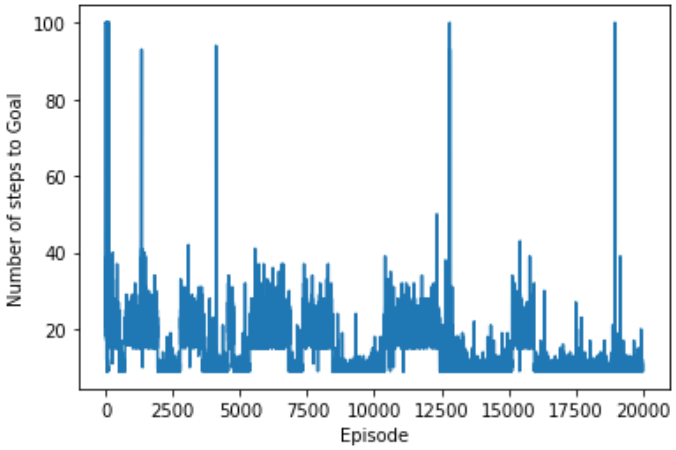
Combination 9:

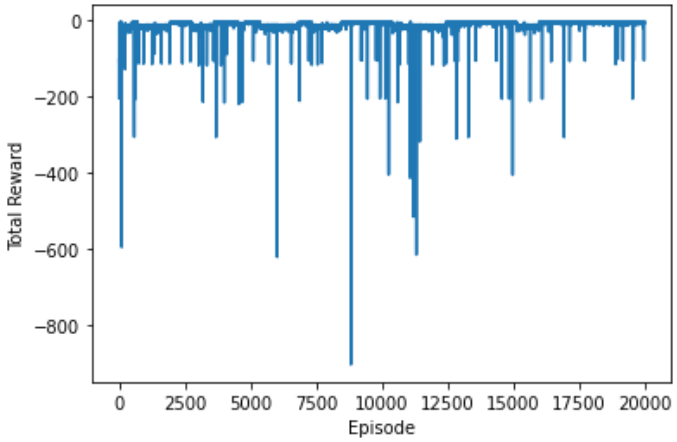
- Policy: Epsilon Greedy
- Start State: [3, 6]
- P = 1

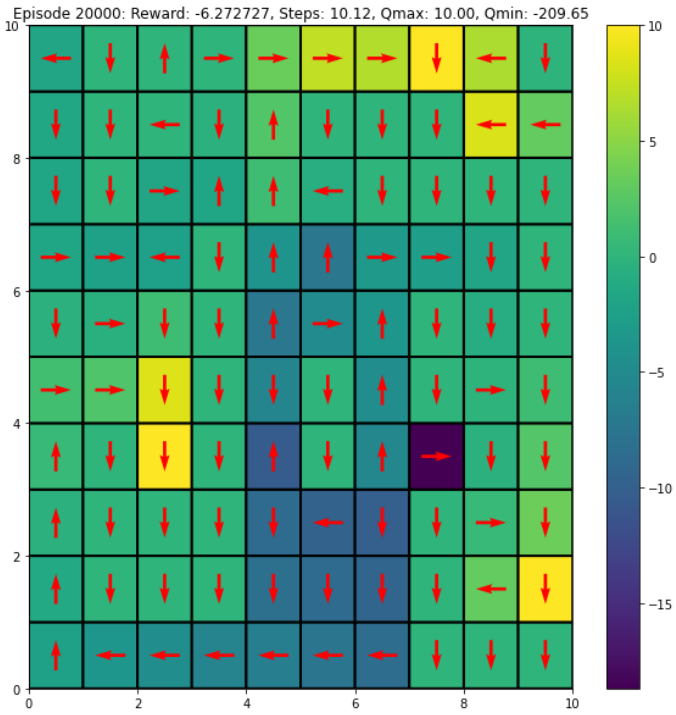
- Wind = True

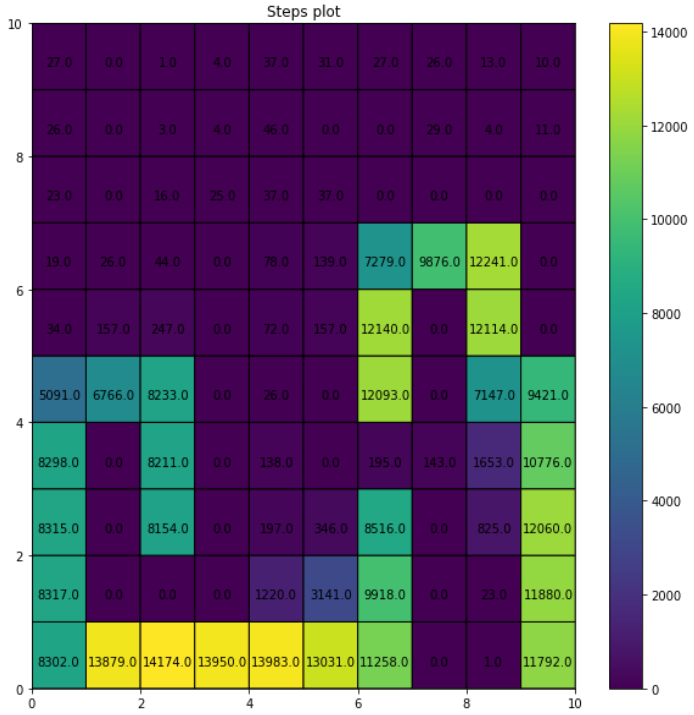
Best chosen hyperparameters:

- ε = 0.01
- α = 0.36
- ϒ = 0.95









Inference: Getting a better graph than Q-Learning with almost the same parameters as Q-learning. Goal states (0,9) and (8, 7) are being reached. The average reward of -6.27 and the average steps is 10.12.

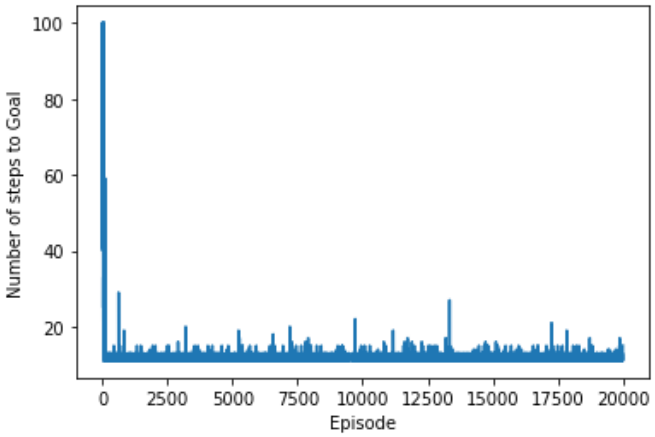
Combination 10:

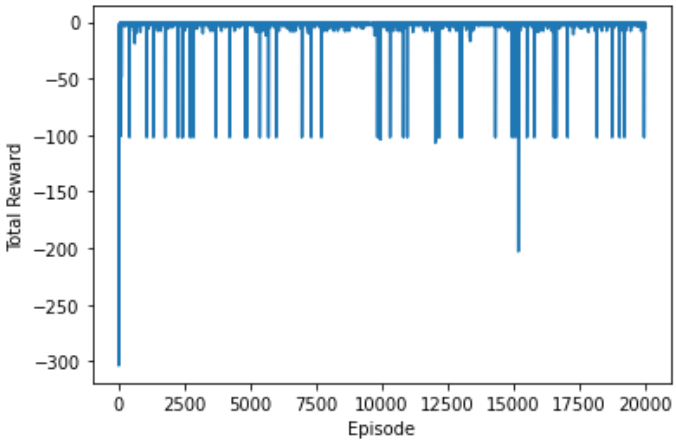
- Policy: Epsilon Greedy

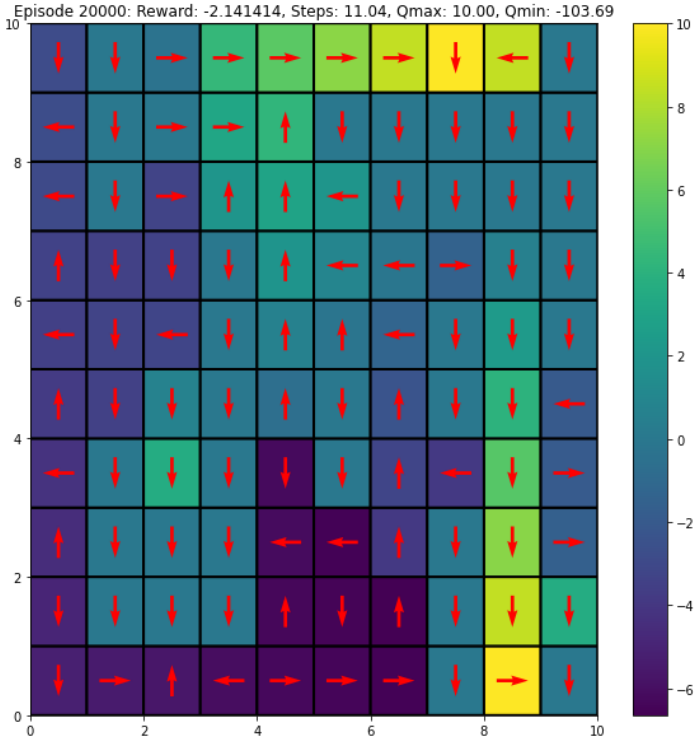
- Start State: [3, 6]
- P = 1
- Wind = False

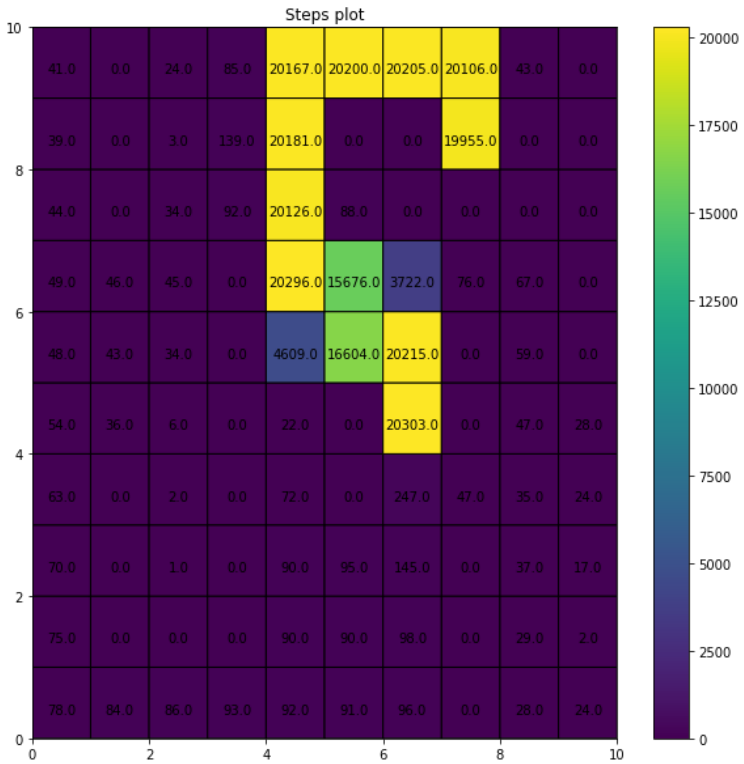
Best chosen hyperparameters:

- ε = 0.01
- α = 0.36
- ϒ = 0.95









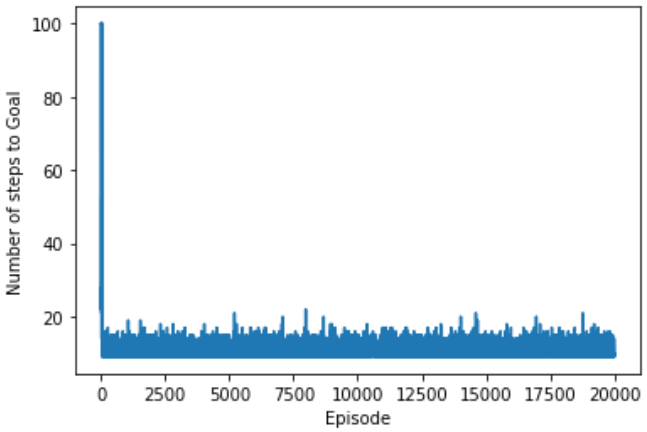
Inference: With **wind = False**, getting a better graph than Q-Learning with almost the same parameters as Q-learning. Goal state (8,7) is being reached. The average reward of -2.14 and the average steps is 11.04.

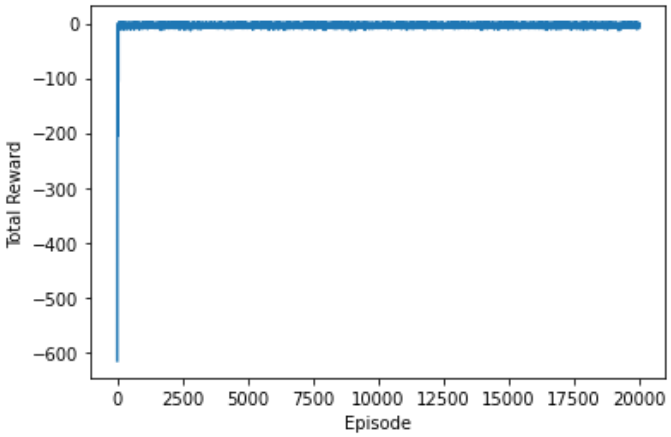
Combination 11:

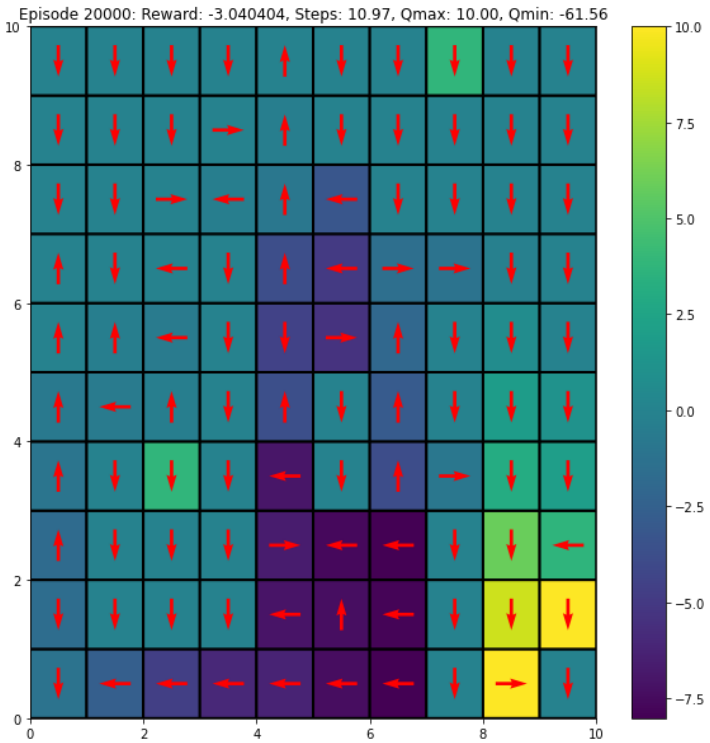
- Policy: Softmax
- Start State: [3, 6]
- P = 1
- Wind = True

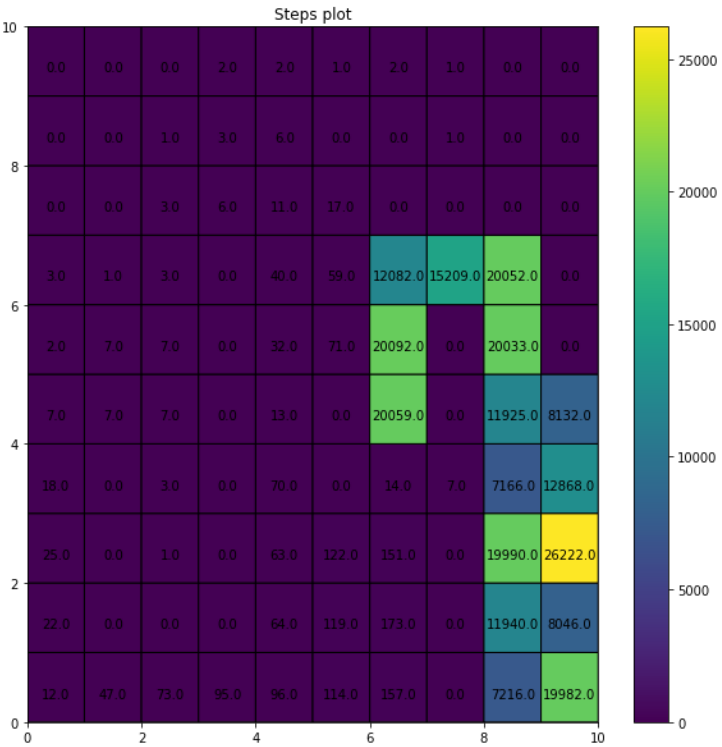
Best chosen hyperparameters:

- β = 5
- α = 0.36
- ϒ = 0.94









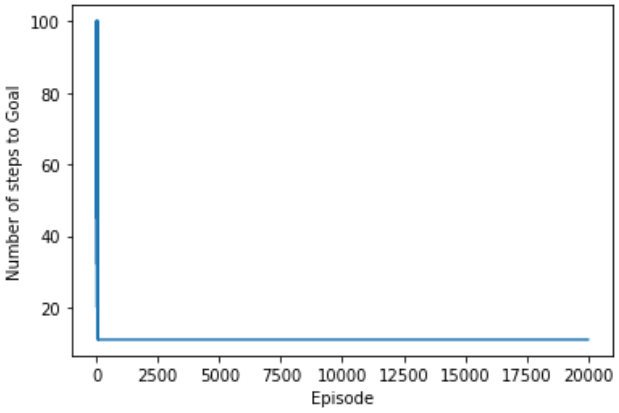
Inference: Got a good graph. With almost the same parameters as Qlearning. Goal state (0,9) is being reached. Average reward of -3.04 and the average steps is 10.97.

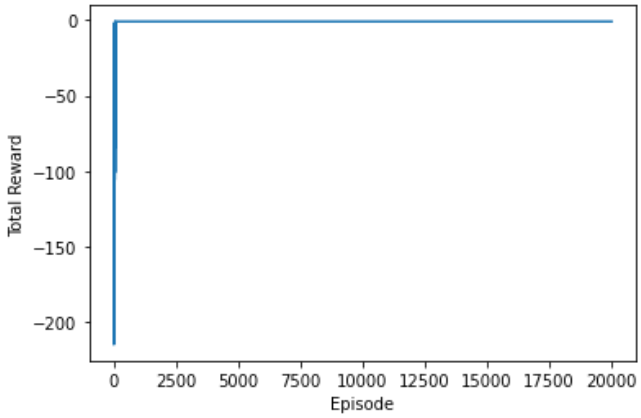
Combination 12:

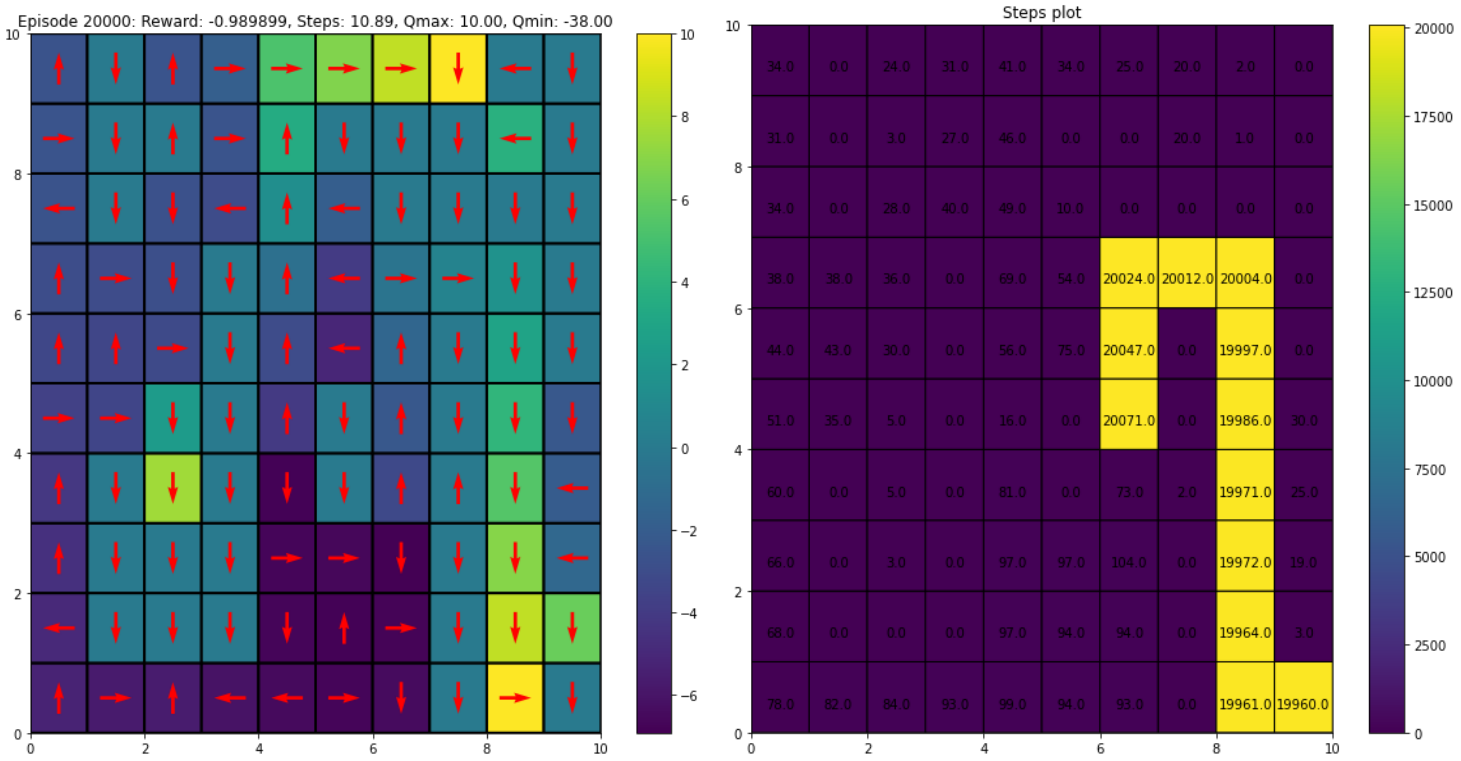
- Policy: Softmax
- Start State: [3, 6]
- P = 1
- Wind = False

Best chosen hyperparameters:

- β = 5
- α = 0.38
- ϒ = 0.94







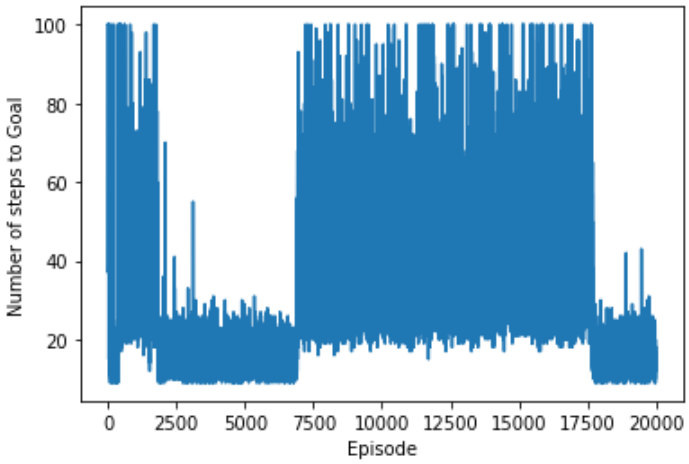
Inference: Almost perfect graph. Best graph among the lot. With almost the same parameters as Q-learning. Goal state (0,9) is being reached.

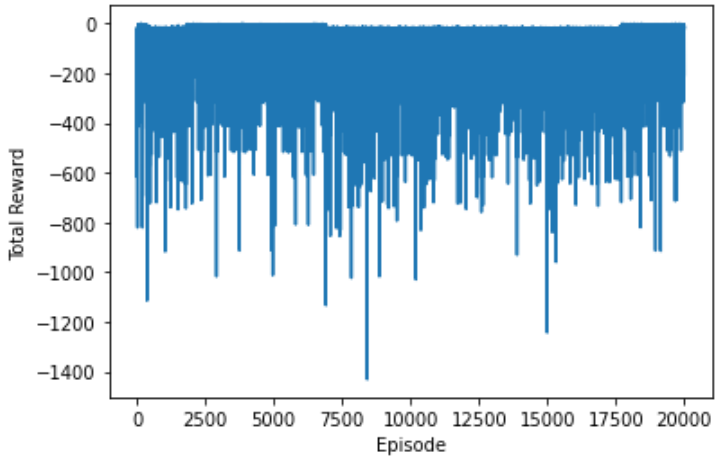
Combination 13:

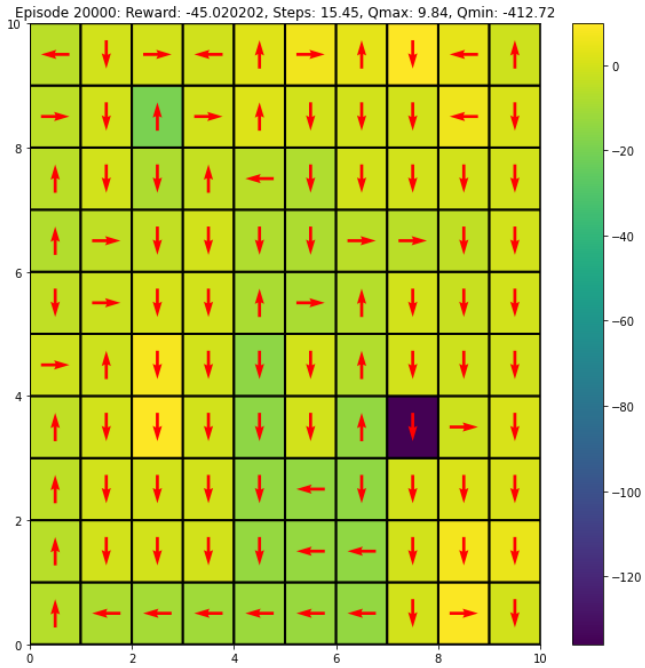
- Policy: Epsilon Greedy
- Start State: [3, 6]
- P = 0.7
- Wind = True

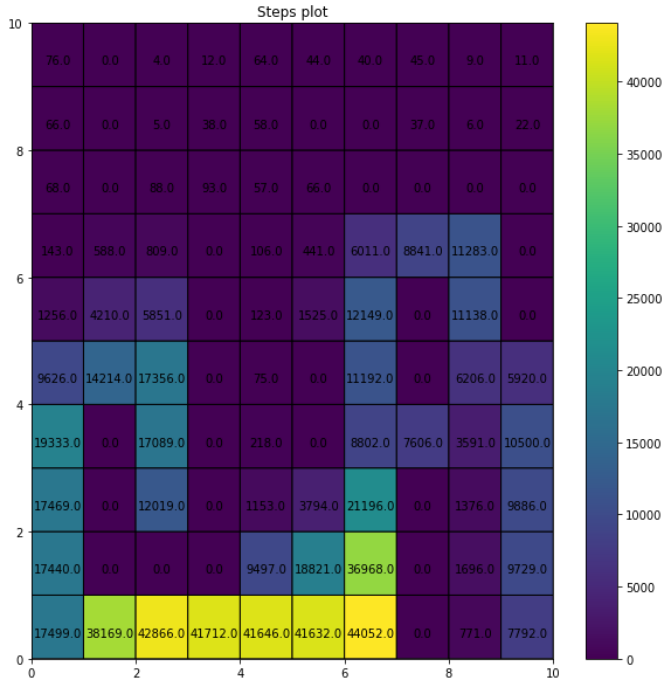
Best chosen hyperparameters:

- ε = 0.01
- α = 0.36
- ϒ = 0.94









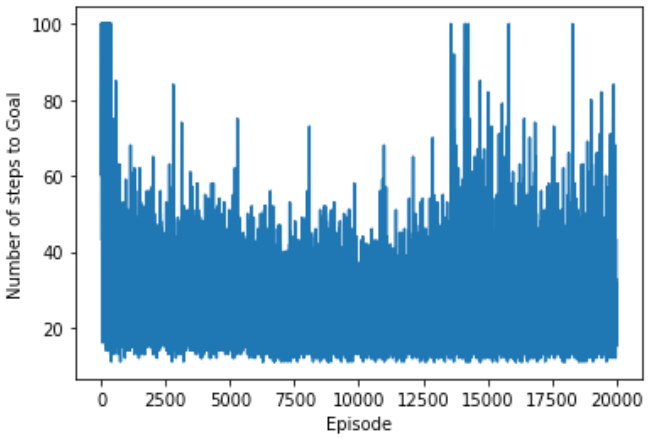
Inference: All the goal states were reached. The best reward we could get is -45.02 with average steps of 15.45.

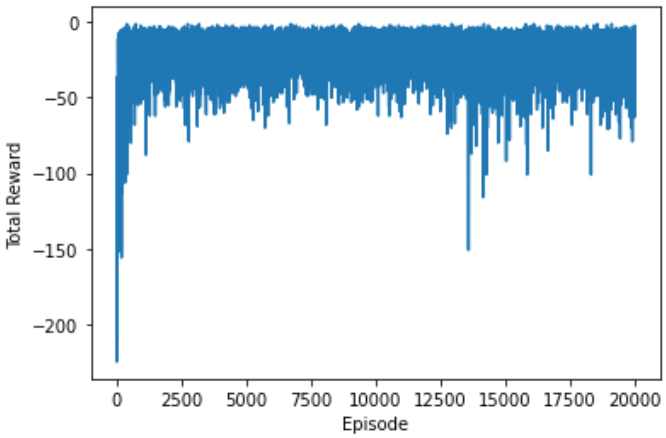
Combination 14:

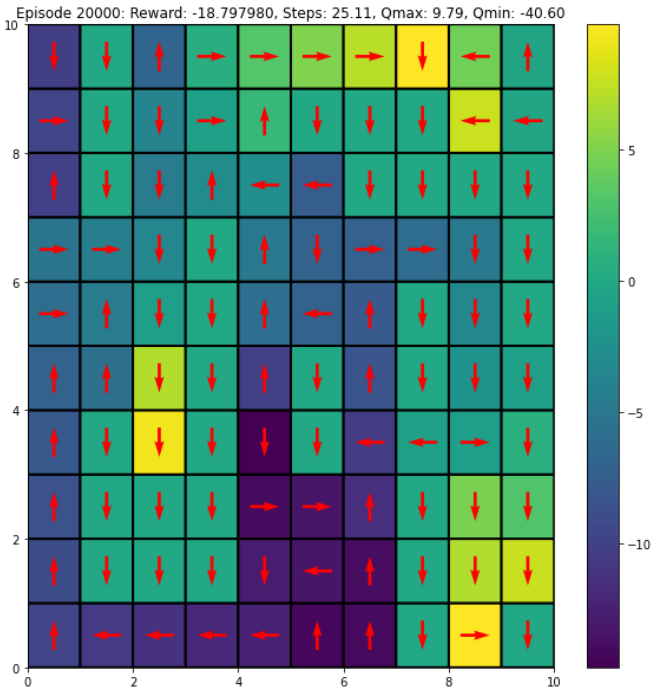
- Policy: Epsilon Greedy
- Start State: [3, 6]
- P = 0.7
- Wind = False

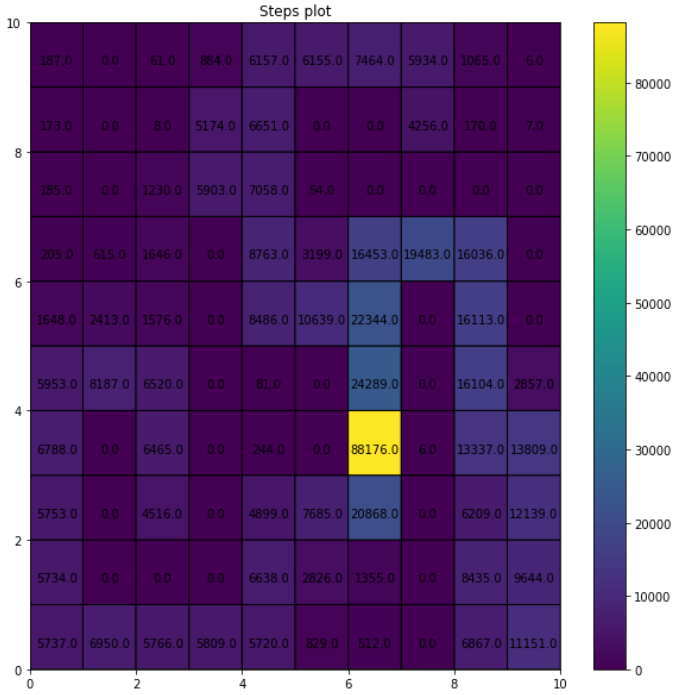
Best chosen hyperparameters:

- ε = 0.01
- α = 0.36
- ϒ = 0.94









Inference: All the goal states were reached most of the time because of non-determinism. Both alpha(0.6 to 0.1) and beta(0.1 to 6) needed to be hugely changed to even get something productive. The best reward we could get is -18.79 with average steps of 25.11.

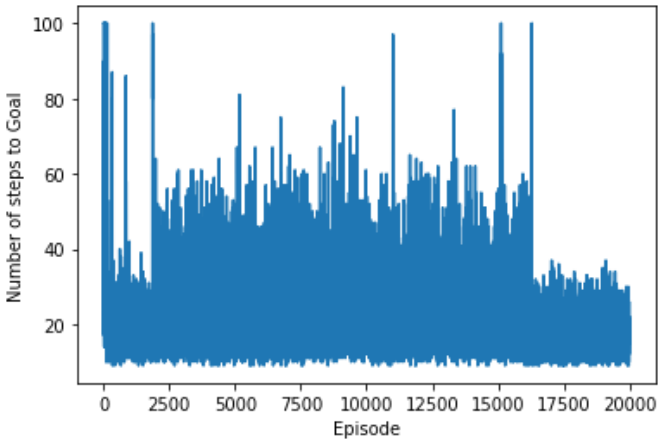
Combination 15:

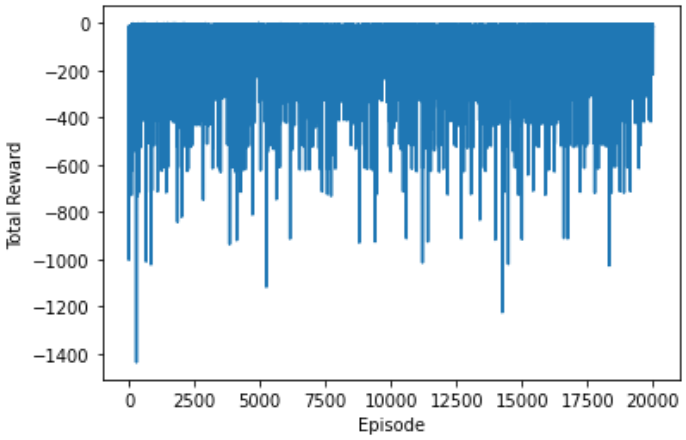
- Policy: Softmax
- Start State: [3, 6]
- P = 0.7
- Wind = True

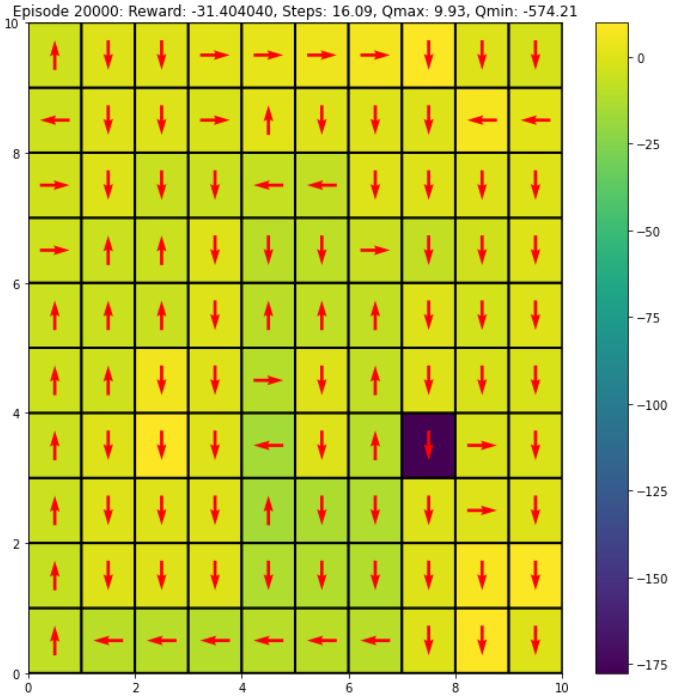
Best chosen hyperparameters:

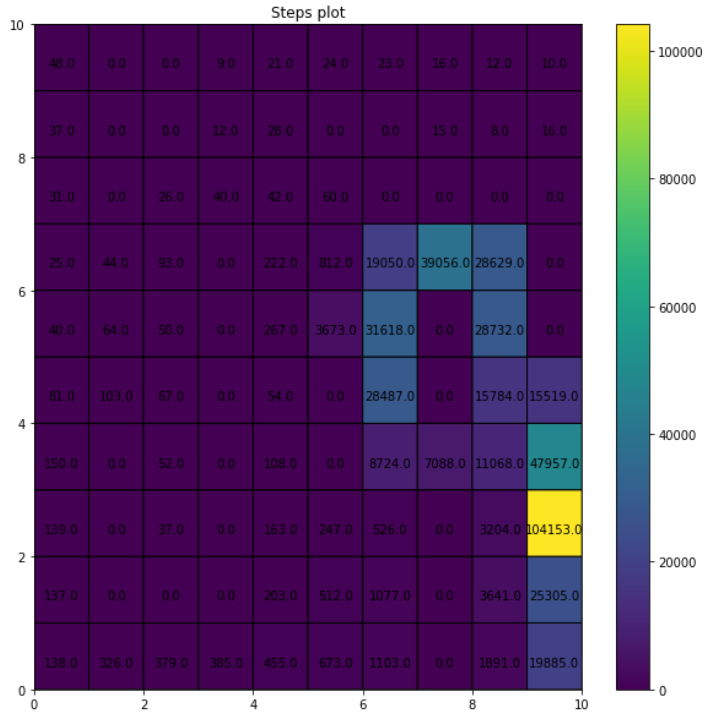
- $\beta = 5$
- $\alpha = 0.57$

- Υ = 0.94









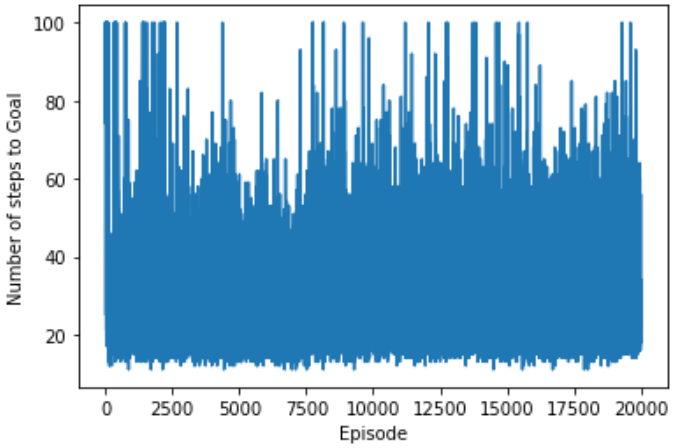
Inference: The goal state (0,9) was reached most of the times. Both alpha(0.6 to 0.1) and beta(0.1 to 6) needed to be hugely changed to even get something productive. The best reward we could get is -31.4 with average steps of 16.09.

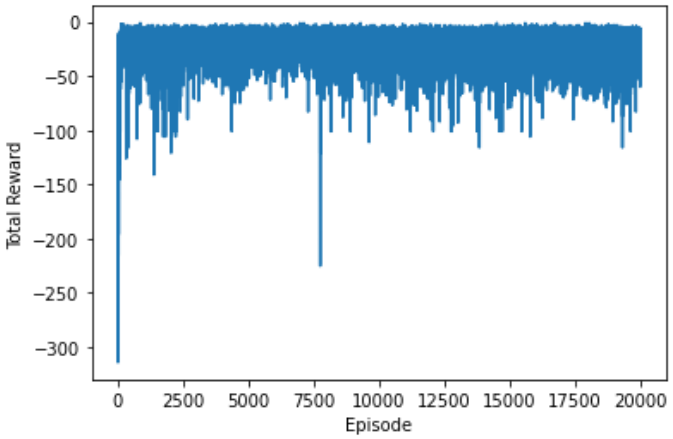
Combination 16:

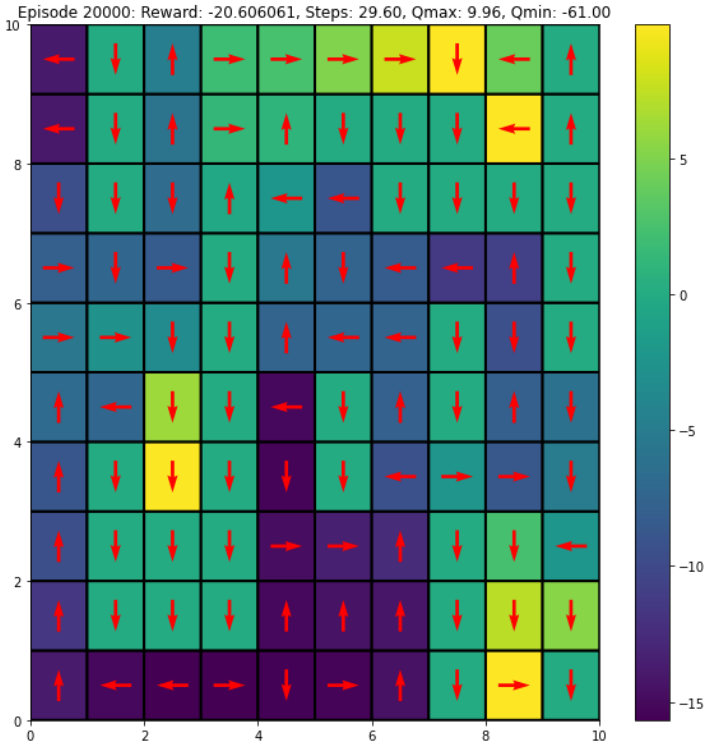
- Policy: Softmax
- Start State: [3, 6]
- P = 0.7
- Wind = False

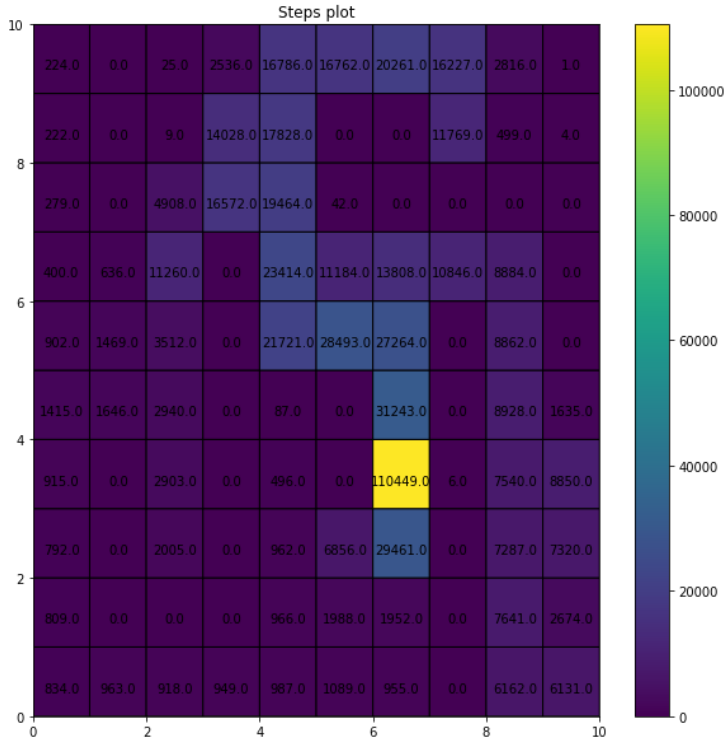
Best chosen hyperparameters:

- β = 5
- α = 0.57
- ϒ = 0.94









Inference: Got a little better result with **Wind = False.** Average reward is -20.60 and the average number of steps is 29.60. All the three goal states are reached but state (8, 7) was reached a maximum number of times.