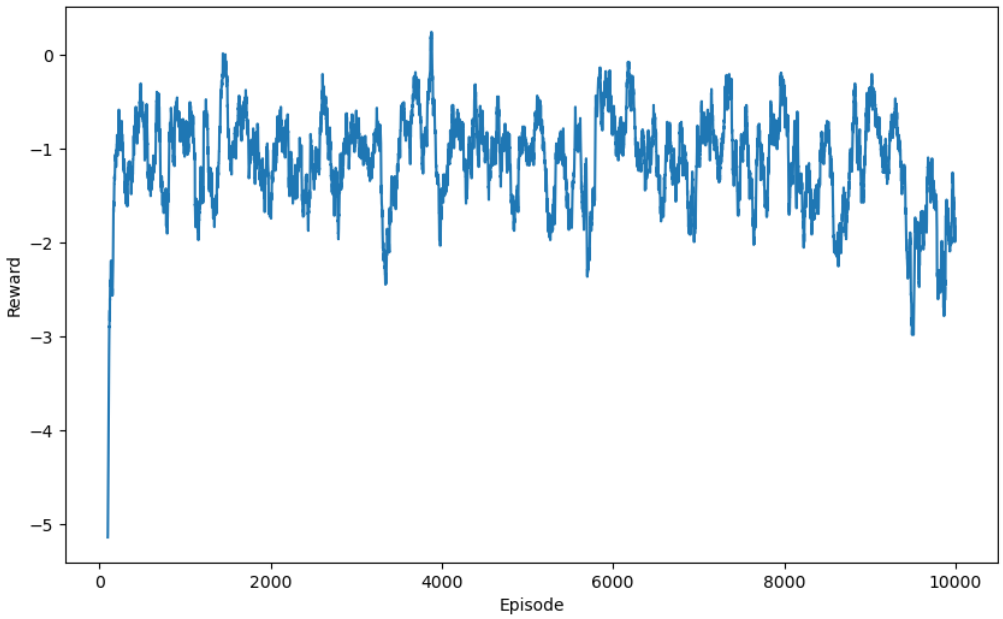
CT-5130 Assignment 2 report

Caolan McDonagh – 21249929

Results

i) Basic hyperparameters: Alpha = 0.5, Gamma = 0.9, Epsilon = 0.10

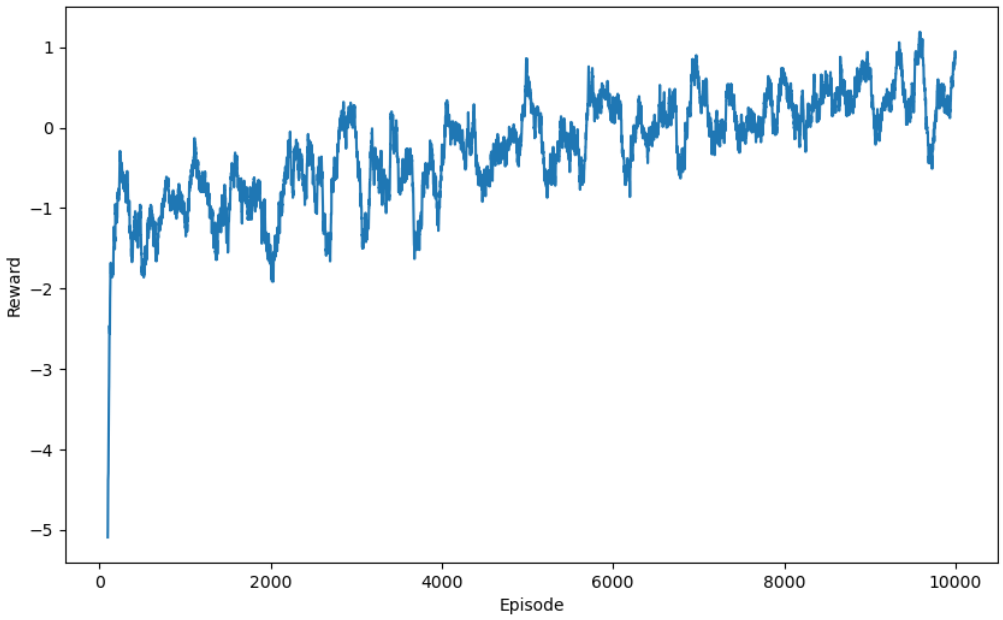


Final Q-Values:				
-1.209	-0.784	0.689	2.051	4.028
-5	1.008	2.389	-5	6.092
-3.036	2.615	3.826	5.467	7.979
-3.748	-5	5.726	7.972	10
-3.454	-2.628	-5	10	10

The parameters of (i) resulted in an agent that did learn a reasonable path to the goal, but as seen in the graph it fluctuates and begins to fall off slightly towards the last few thousand episodes, implying it has reached convergence and isn't learning any further. Comparing this to further results, this could be an indication of suboptimal hyperparameters.

The Q-value results shows that the agent has learned an ideal path towards the goal, avoiding the holes.

ii) Epsilon Decay: Alpha = 0.5, Gamma = 0.9, Epsilon = 0.10 (where epsilon reaches 0.0 by the 10,000th episode via linear decay.)

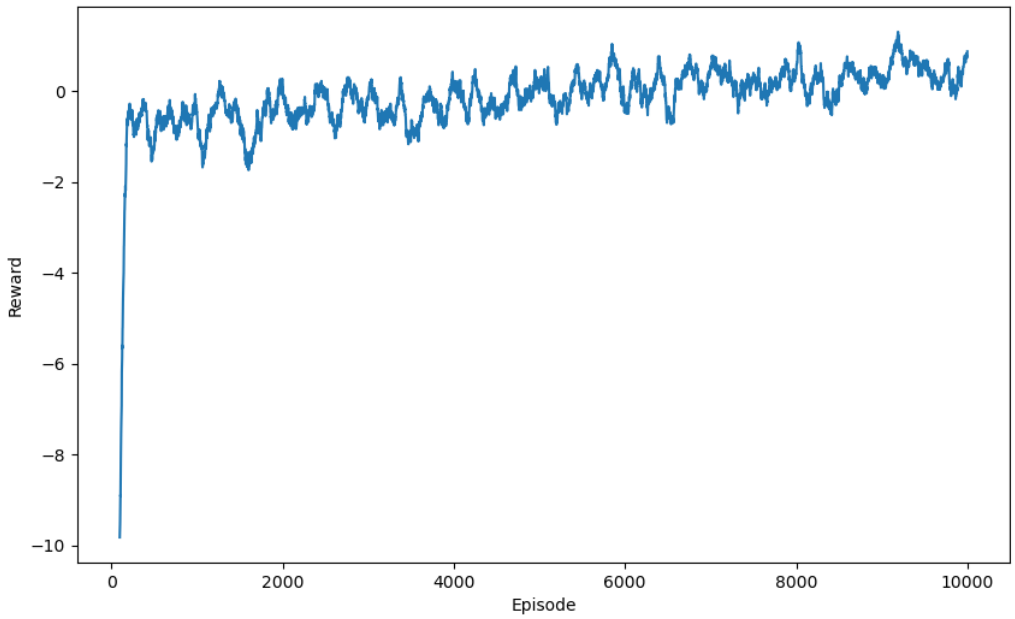


Final Q-Values:				
-0.434	0.629	1.81	3.122	4.58
-5	1.744	3.122	-5	6.2
-0.684	1.896	4.58	6.2	8
-2.652	-5	6.2	8	10
-2.578	-2.806	-5	10	10

(ii) looks more promising and all that was changed was the addition of epsilon decay. This allows the agent to explore more freely early on, but over time it will explore less and stick to optimal paths, as seen by the increasing reward over time, reaching higher peaks over time. The convergence looks to also have been fixed by the decaying as it does not trail off towards the end and keeps rising.

With epsilon decay, Q-value results are different, providing higher values (and so higher expected reward). This is due to the epsilon falling off over time, so the agent is inclined to stick to the known paths and put a higher reward towards these paths.

iii) My own hyperparameters: Alpha = 0.1, Gamma = 0.95, Epsilon = 0.1 (with Epsilon decay)



Final Q-Values:				
0.949	2.053	3.213	1.681	-0.828
-5	2.995	4.435	-5	3.711
0.931	4.35	5.721	7.075	8.5
-0.998	-5	6.835	8.5	10
-0.85	-0.668	-5	10	10

The graph for (iii) shows a slower ramp up in learning, not really beginning to hit the peak until ~3000 episodes in. This hyperparameter selection looks to look work the best (not shown as the y axis isn't fully displayed) as the reward overtime is higher than (i) and (ii) and trends towards keeping above 0, closer to 1 inside the last 1000 episodes or so.

The hyperparameters for (iii) are different than in (i) and (ii), with a slightly higher discount rate to put more emphasis on long term rewards in the agents learning and a much lower learning rate, allowing the agent to monitor much more steps and keep them in mind when learning.

This resulted in higher rewards overall and the Q-values are higher with a clearer single path preferred, compared to (i) and (ii) that had two paths with similar Q-values.