Name: Vikrant Singh Jamwal
Student ID: 23104534

CT1534 Reinforcement Learning

Assignment 2 (Frozen Lake)

Results

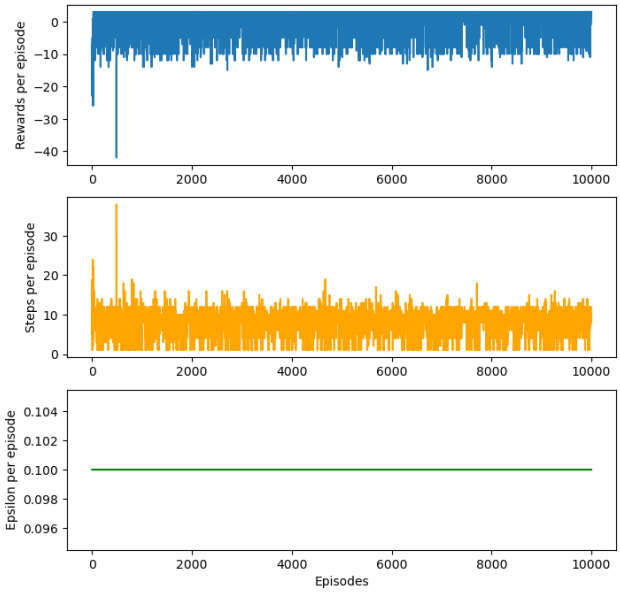
Combination-1

- Hyperparameters
  - Learning rate (alpha) = 0.5
  - Discount rate (gamma) = 0.9
  - Epsilon = 0.10 (Constant)

- Action Values (Yellow – START, Blue – HOLES, Green – GOAL)

-0.434	0.629	1.81	3.122	4.58
0.0	1.81	3.122	0.0	6.2
1.583	3.122	4.58	6.2	8.0
-1.945	0.0	6.2	8.0	10.0
-1.902	-1.855	0.0	9.995	0.0

- Plots for total rewards, steps taken and epsilon value per episode.



- Observations
  - Due to constant epsilon, we can see the exploration in learning till the last episode. We cannot visualize the highest reward or minimum steps required to reach the goal.
  - As the Exploration is high, we can see that every path is considered well as the action value of each divergence in path is similar. But, due to constant epsilon, exploiting the most optimal policy is hindered, hence action values are not accurately distributed.

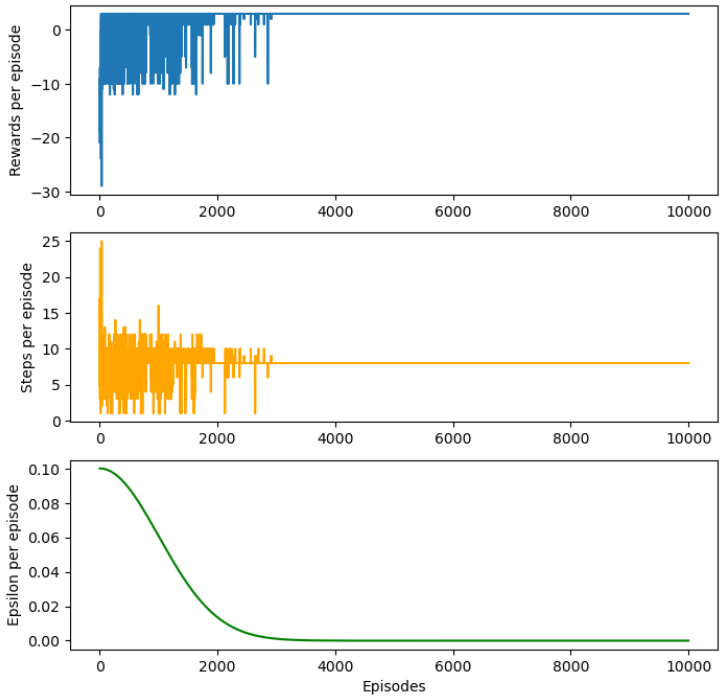
Combination-2

- Hyperparameters
  - Learning rate (alpha) = 0.5
  - Discount rate (gamma) = 0.9
  - Epsilon = 0.10 (with Decay rate = 0.01)
  - Decay is proportional to the episode count and decay rate.

- Action Values (Yellow – START, Blue – HOLES, Green – GOAL)

-0.434	0.629	1.81	-2.004	1.506
0.0	1.81	3.122	0.0	6.098
-1.68	3.122	4.58	6.2	8.0
-1.482	0.0	6.2	8.0	10.0
-1.561	-1.426	0.0	10.0	0.0

- Plots for total rewards, steps taken and epsilon value per episode.



- Observations
  - After introducing Decay of exploration factor (epsilon), we can see that the rewards and step taken are converging. Hence, we can see the maximum reward possible and minimum steps taken to reach the goal.
  - But as the epsilon decayed to 0 in around 3000 episodes, we can see in the table that the exploration stopped way too early and hence not all best paths are weighted equally as only one best path is exploited.

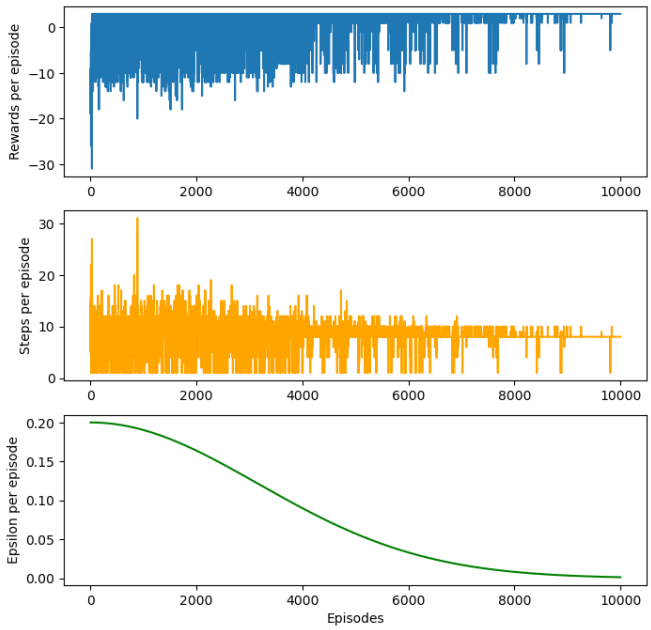
Combination-3

- Hyperparameters
  - Learning rate (alpha) = 0.5
  - Discount rate (gamma) = 0.951 (Increased from 0.9)
  - Epsilon = 0.20 (Increased from 0.10) (with Decay rate = 0.001 (Earlier 0.01))
  - Decay is proportional to the episode count and decay rate.

- Action Values (Yellow – START, Blue – HOLES, Green – GOAL)

0.984	2.086	3.245	4.464	5.745
0.0	3.245	4.464	0.0	7.093
3.245	4.464	5.745	7.093	8.51
1.208	0.0	7.093	8.51	10.0
-1.928	-1.928	0.0	10.0	0.0

- Plots for total rewards, steps taken and epsilon value per episode.



- Observations

  - For the improvement from previous combinations of hyperparameters, increased the Discount factor (gamma) to give more emphasis to the future rewards. Also increased the epsilon to make the starting episodes more exploratory. Decreased the epsilon decay rate to avoid early stoppage of exploration.
  - From the above changes, we can see accurate action values with similar divergence weights for all possible best paths. Also, we can visualize convergence of total rewards and steps taken to reach the goal. Hence, these parameters worked better than above combinations.