Reinforcement Learning Assignment #1 Report

Hazique Quazi
mquazi@buffalo.edu
UB Person #: 50416875
Github Repo for Submissions **(@ub-rl added as collaborator)**

Abstract

This report details and demonstrates our learning of Reinforcement Learning concepts such as creating an agent and environment, running the agent for a given number of timesteps and visualizing the change of states and rewards given to the agent by the environment in response to a given action. Below I describe the execution of the agent in Deterministic and Stochastic Environment.

1. Deterministic Environment

The deterministic environment can be described as follows:

Action Set

$$A = \{Left = 0, Top = 1, Right = 2, Bottom = 3\}$$

State Set

$$S = \{0\ to\ 15\ Discrete\ States\}$$

Reward Set

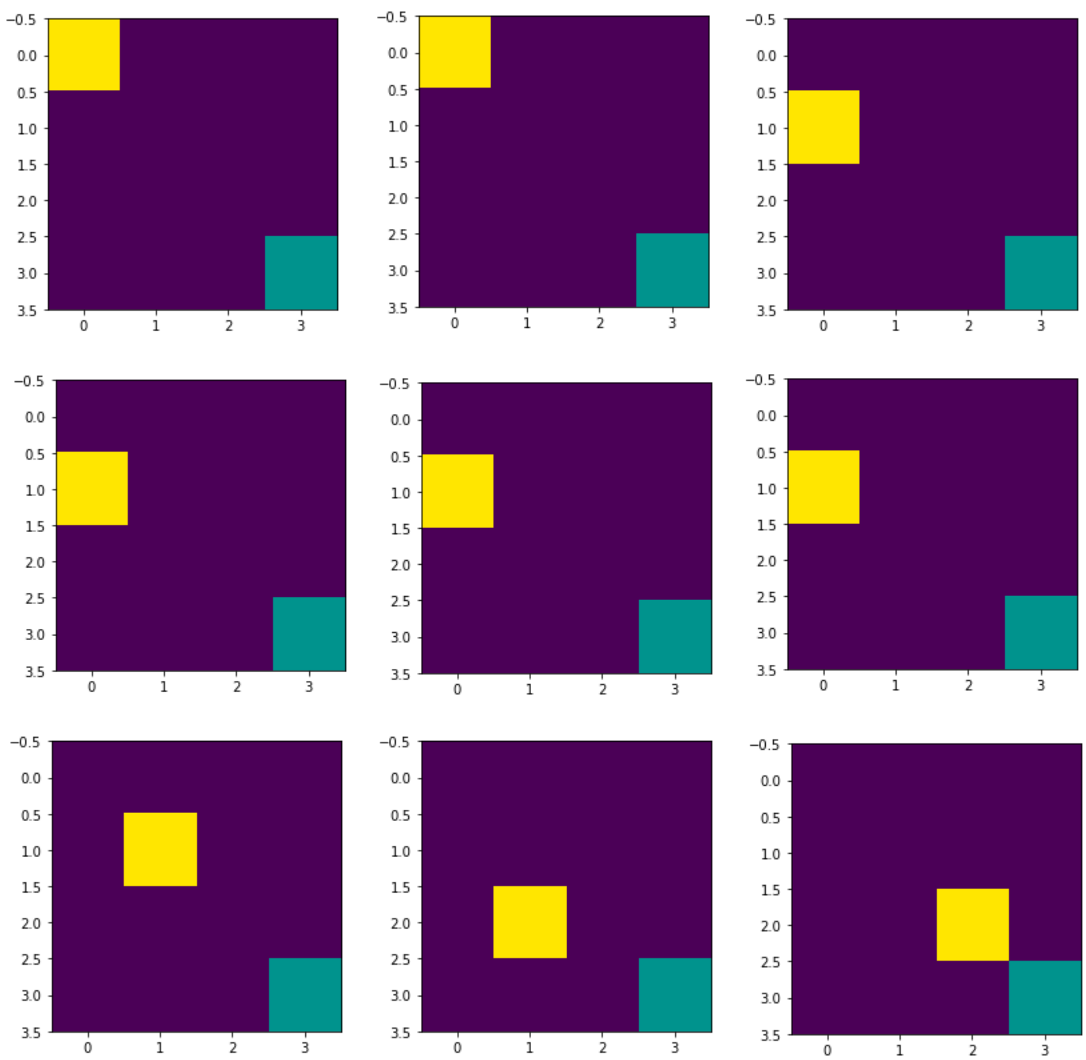
$$R = \{-1, 0, 1, 2\}$$

The reward received by the agent can be better represented as a 4x4 matrix

```
([[0, 1, 1, 1],
[1, -1, -1, 1],
[1, -1, -1, 1],
[1, 1, 1, 2]])
```

The reward 0 corresponds to the starting position of the agent whereas the 2 corresponds to the end position.

Visualization for Deterministic Environment



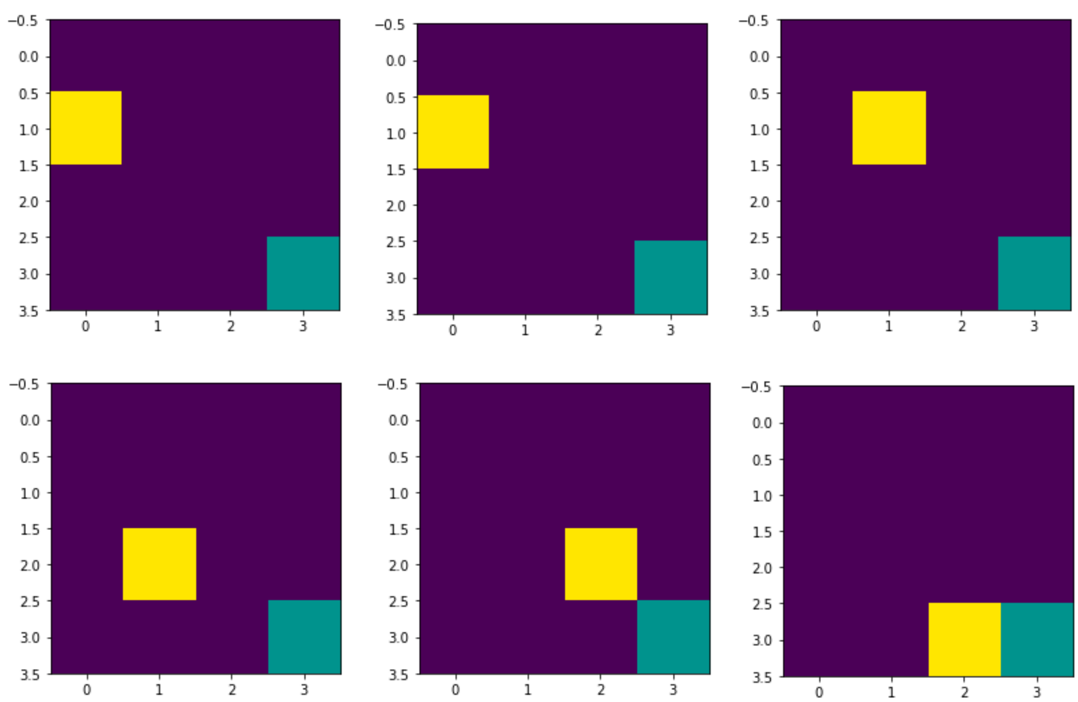
TimeStep	Action	Reward	Total Rewards	Done
0	0	0	0	False
1	0	0	0	False
2	2	1	1	False
3	1	1	2	False
4	1	1	3	False
5	1	1	4	False
6	3	-1	3	False
7	2	-1	2	False
8	3	-1	1	False
9	0	-1	0	True

2. Stochastic Environment

The Stochastic environment has the same Action Set, State Set & Reward Set as the Deterministic environment. It however adds stochasticity to the action the agent can take at any time step. I have defined the stochastic environment by letting an agent take an action probabilistically according to the following logic:

1. Assign random probabilities to actions (left = up = 0 to10%, right = down = 11-90%)
2. Check if direction with maximum probability has a probability value greater than 30%
   a. If False, the agent goes left instead of right OR the agent goes up instead of down.

Visualization for Stochastic Environment

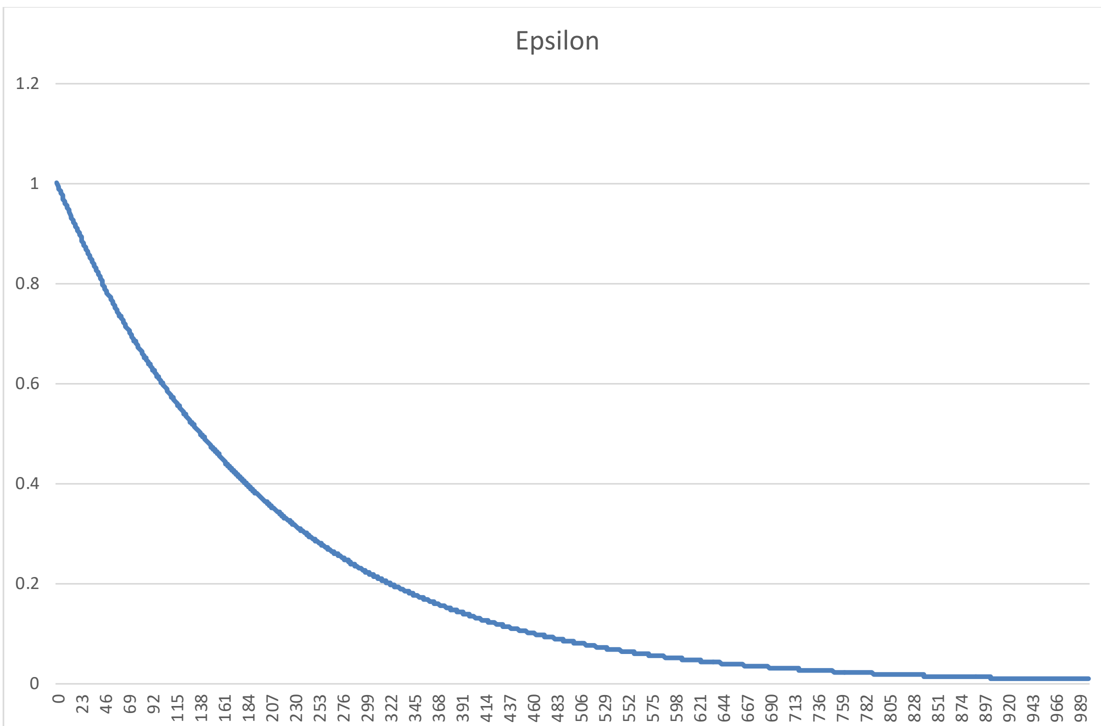


TimeStep	Action	Reward	Total Rewards	Done
0	2	1	1	False
1	1	1	2	False
2	3	-1	1	False
3	2	-1	0	False
4	3	-1	-1	False
5	2	1	0	False
6	3	2	2	True

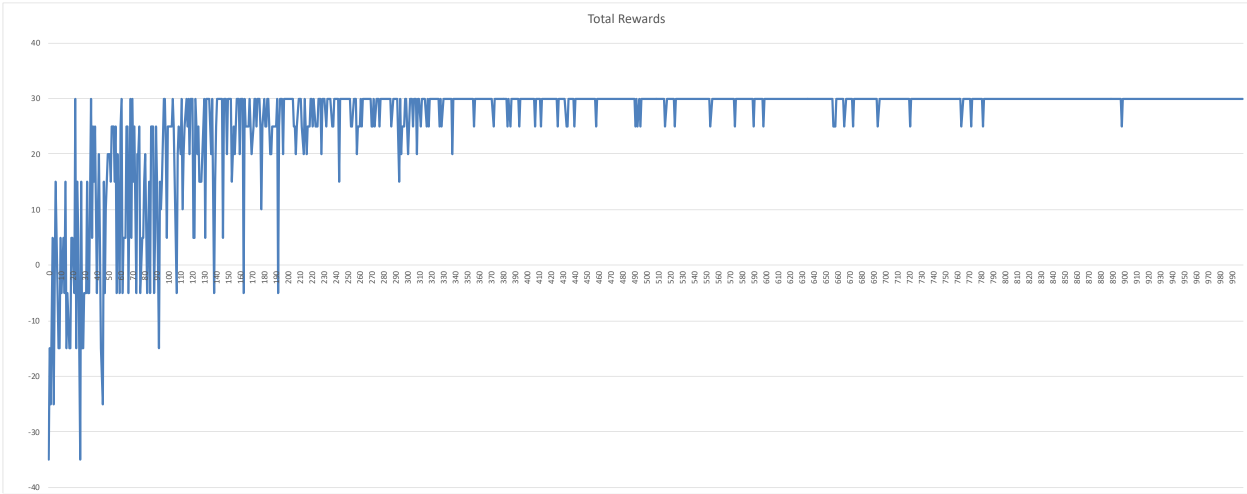
3. Training on Q Learn

Deterministic Environment

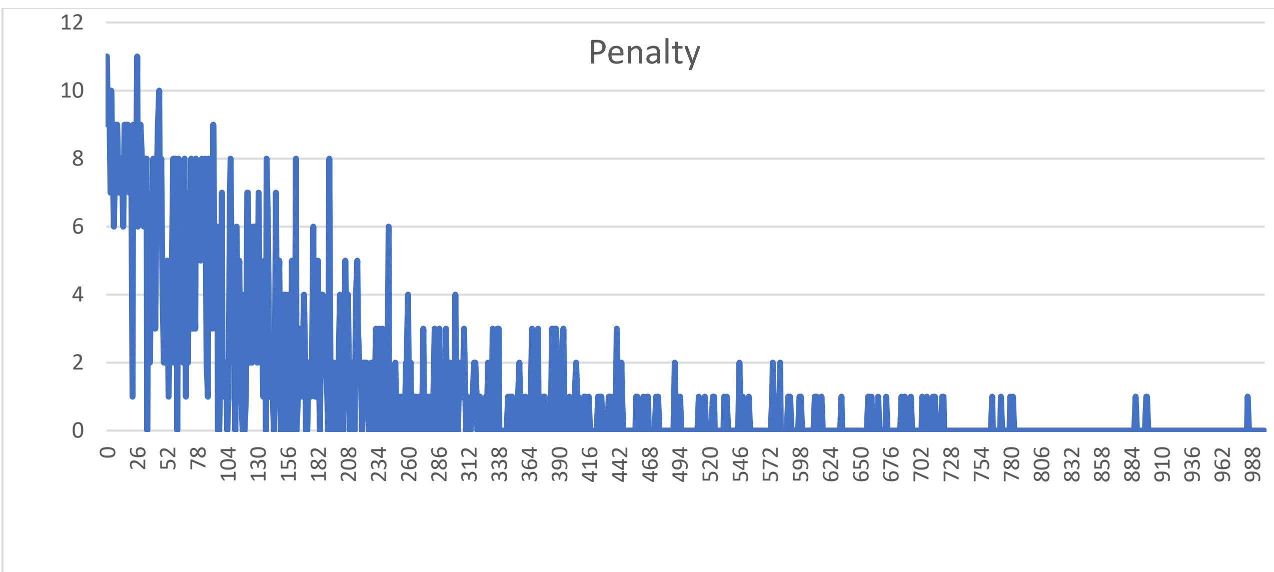
Epsilon Decay – The epsilon value decays at a rate of 0.995 per episode for 1000 episodes



Total Rewards during Training



Total penalties during Training



Inference:

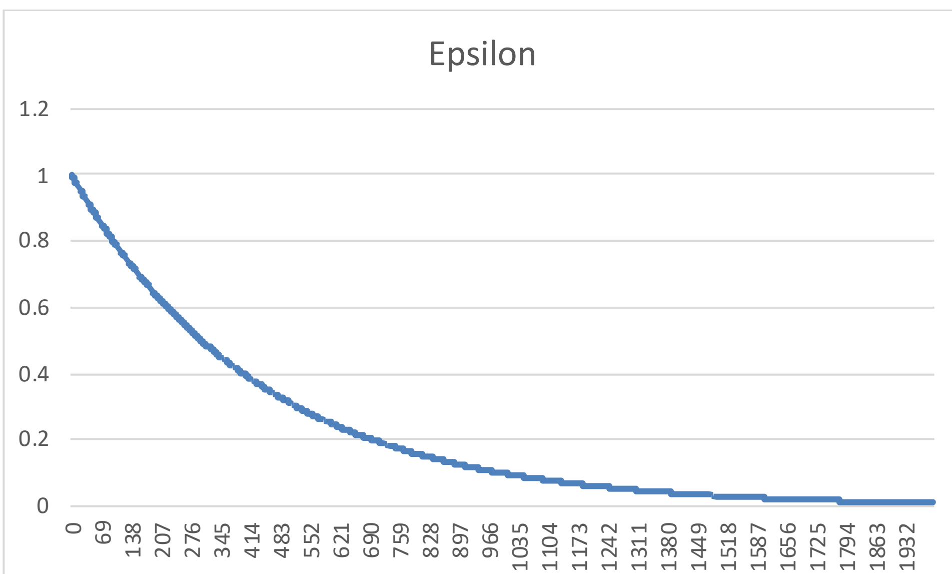
As the number of penalties go down and the number of rewards collected saturates at a timesteps = 6, we can say that the agent has learned the optimal path.

Evaluation Results:

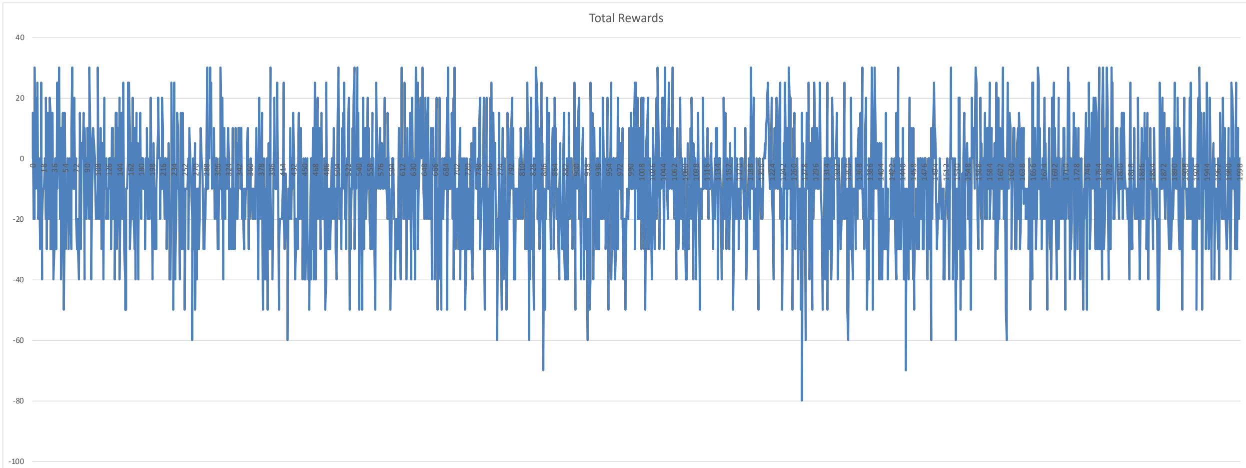
Since the agent earns constant reward in 6 timesteps while successfully reaching the goal position, our agent has in fact learned the optimal path.

Stochastic Environment:

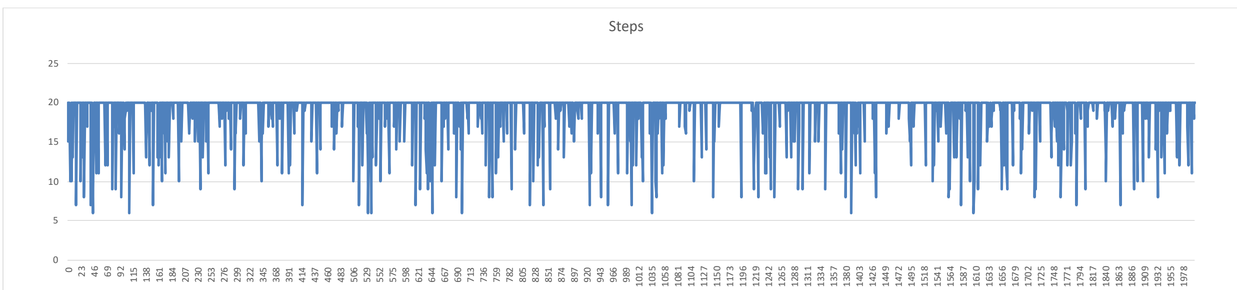
Epsilon Decay rate is 0.9977 per step for 2000 steps

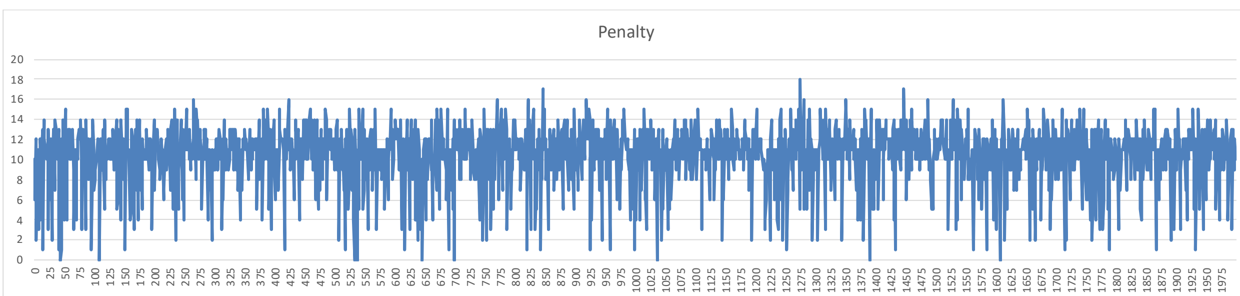


The stochastic environment presents a challenge for the agent. As a result, the agent collects a wide range of rewards during the training cycle.



Concomitantly, the agent ends up taking a greater number of steps and penalties during the training.





Evaluation Results:

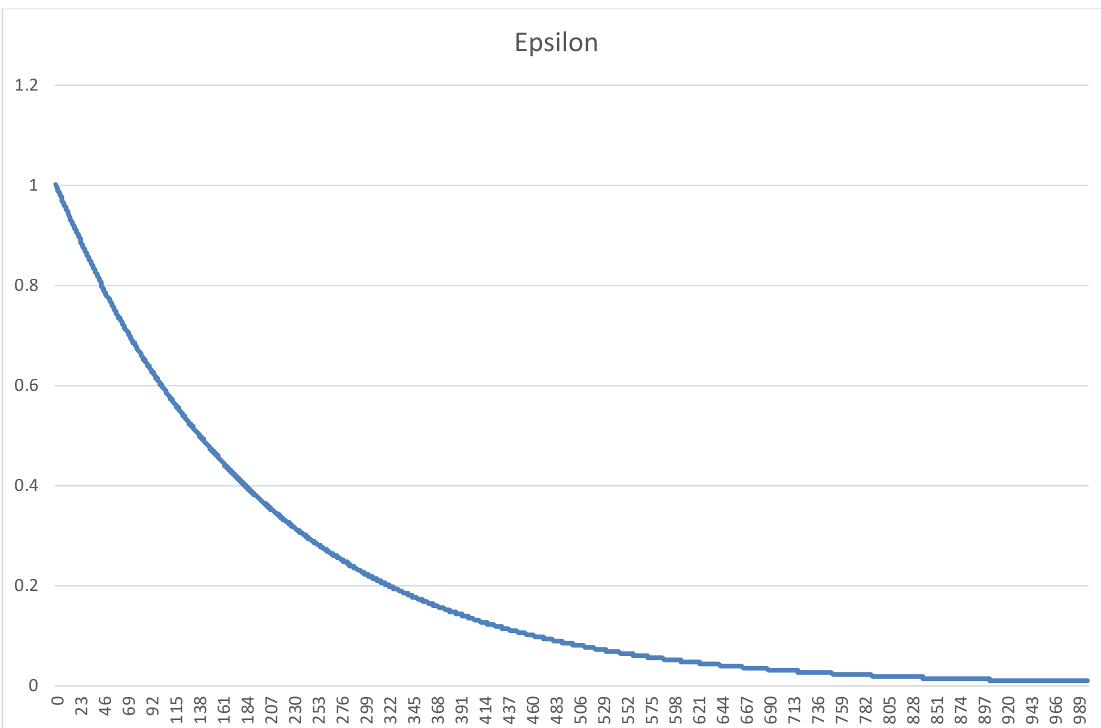
As a result of the environment being stochastic, the agent is not able to execute the optimal policy. It seldom reaches the goal position.

Episode #	otal reward	Penalty	Steps
1	-30	13	20
2	0	10	20
3	-10	11	20
4	-15	10	17
5	-20	12	20
6	-40	14	20
7	20	8	20
8	-10	11	20
9	-30	13	20
10	30	2	10

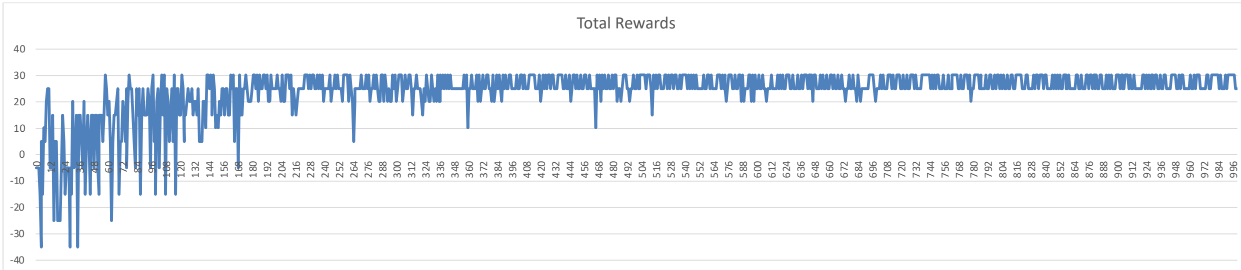
4. Training on SARSA

Deterministic Environment

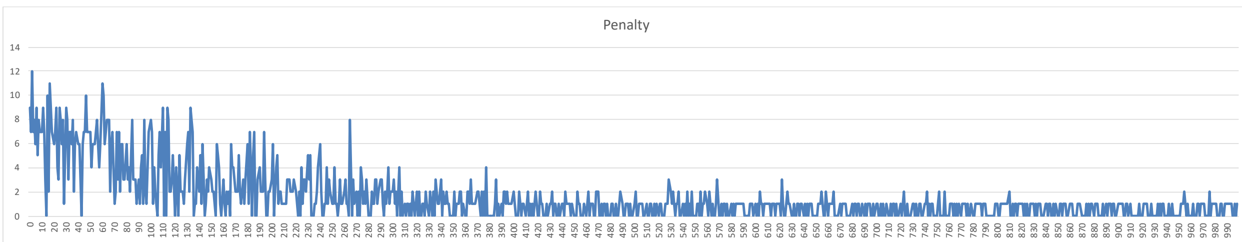
Epsilon Decay – The epsilon value decays at a rate of 0.995 per episode for 1000 episodes



Total Rewards during Training



Total Penalties



Inference:

As the number of penalties go down and the number of rewards collected saturates at minimum timesteps 6 or 7, we can say that the agent has learned the optimal path using SARSA algorithm.

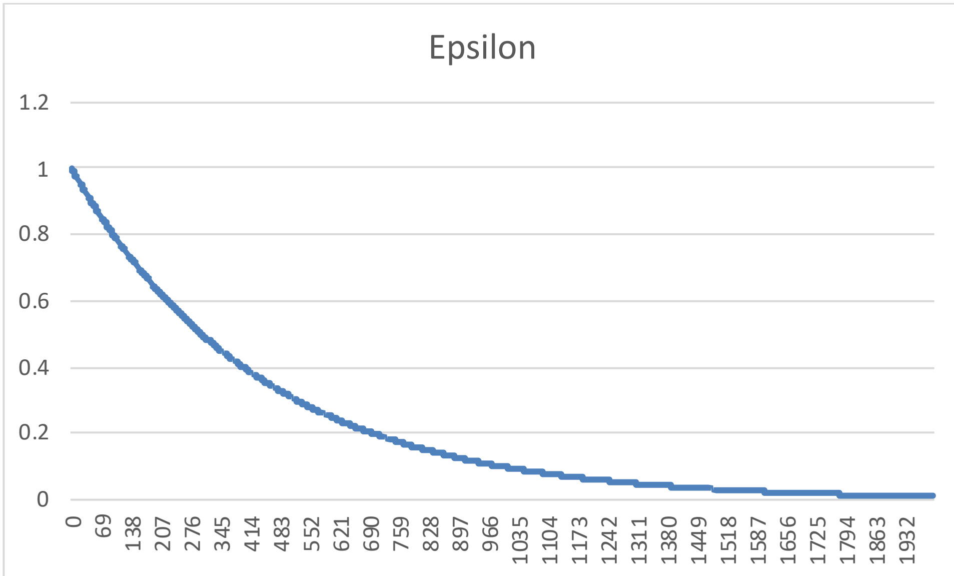
Evaluation Result:

Since the agent earns constant reward in 6 timesteps while successfully reaching the goal position, our agent has in fact learned the optimal path.

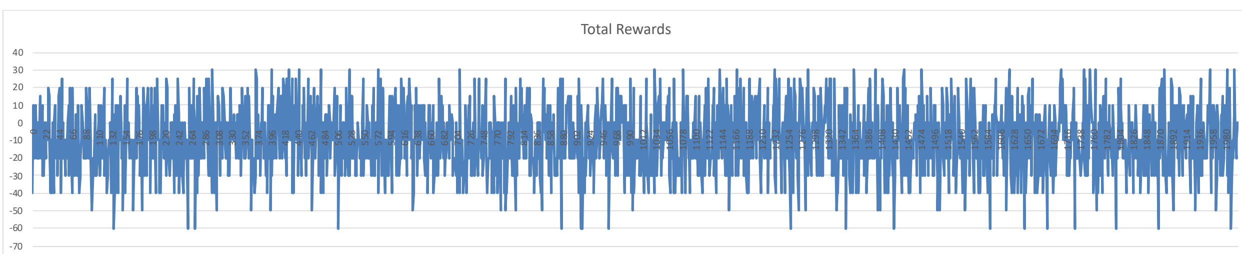
Episode #	otal reward	Penalty	Steps
1	30	0	6
2	30	0	6
3	30	0	6
4	30	0	6
5	30	0	6
6	30	0	6
7	30	0	6
8	30	0	6
9	30	0	6
10	30	0	6

Stochastic Environment:

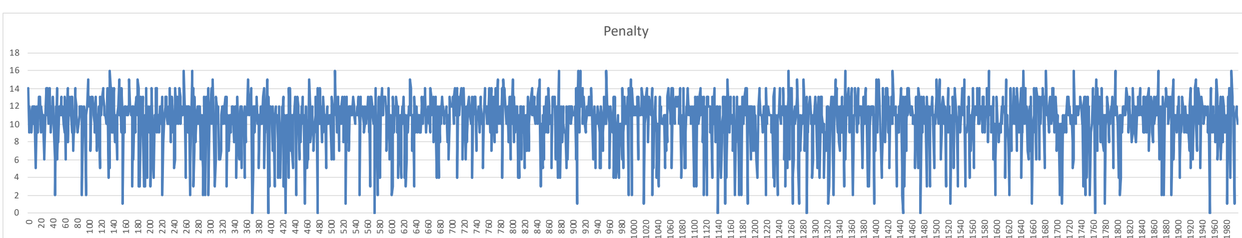
Epsilon Decay rate is 0.9977 per step for 2000 steps



The stochastic environment presents a challenge for the agent. As a result, the agent collects a wide range of rewards during the training cycle like Q Learn in a stochastic environment.



Similarly, the agent receives variable number of penalties throughout training.



Evaluation Results:

Similar to the performance in Q Learning, the agent seldom reaches the goal position.

Episode #	otal reward	Penalty	Steps
1	-60	16	20
2	15	6	15
3	-20	12	20
4	-40	14	20
5	5	9	19
6	0	10	20
7	0	10	20
8	15	6	15
9	-10	11	20
10	-30	13	20

5. Q Learn vs SARSA Comparison

	Training				Evaluation			
	Q Learn Det	Q Learn Stoch	SARSA Det	SARSA Stoch	Q Learn Det	Q Learn Stoch	SARSA Det	SARSA Stoch
Avg steps	7.661	18.672	7.981	18.754	6	18.5	6	18.9
Total Reward	26245	-17000	24525	-19470	300	-35	300	-125
Successful	920	410	915	388	10	2	10	3
Penalties	1206	2129	912	20701	0	96	0	107

From the above result we can clearly say that Q Learn outperform SARSA irrespective of the environment – deterministic or stochastic, on both training and evaluation routines.