Assignment-3
Reinforcement Learning
RanjithKumar Hiremath

1. Goal:

The aim of this project is to implement Q-learning environment given the data and the environment. The Algorithm should be able to make a profitable trade over a period of time, learning from the environment.

2. Data Set:

As data set we have historical stock price for NVIDIA over 5years, which make up to 1258 entries. The features are open price, high, low, closing price and adjusted closing price. Q learning algorithm is trained on looking at the data of 10 days.

80% of data is used for training and the rest is used for testing.

3. Environment:

In any kind of Reinforcement learning algorithm the environment plays a very important role. In the defined environment we have 3 actions and 4 states resulting as the consequence of these actions.

The Three actions are:

- BUY - encoded as 0.
- SELL- encoded as 1.
- HOLD - encoded as 2.

The states resulting from these actions are defined as group of variables measure as part of actions.

- Increase in price(binary).
- Decrease in price(binary).
- Stock held(binary).
- Stock not held(binary).

Combination of these 4 variables form 4 states.

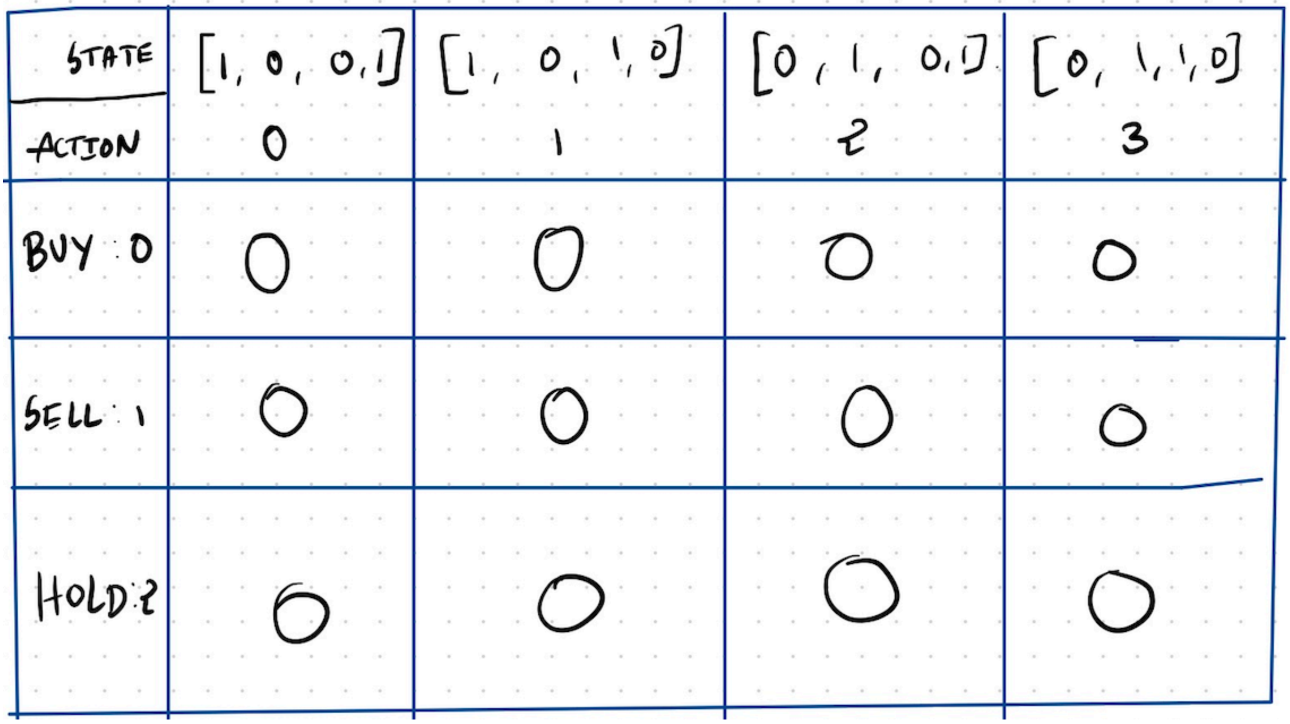
- [1, 0, 0, 1] - encoded as 0.
- [1, 0, 1,0] - encoded as 1.
- [0, 1, 0, 1] - encoded as 2.

- [0, 1, 1, 0] - encoded as 3.

The initial investment is 100000 dollars.

4. Q-learning implementation and training:

The combination of the above mentioned states and actions form a 3*4 Q-table, whose values are initialised to zero as shown below.



STATE	[1, 0, 0, 1]	[1, 0, 1, 0]	[0, 1, 0, 1]	[0, 1, 1, 0]
ACTION	0	1	2	3
BUY : 0	0	0	0	0
SELL : 1	0	0	0	0
HOLD : 2	0	0	0	0

Q learning, tries to find the best next action on the basis of current state, using a metric called q value. It uses the current reward and the maximum q value returned after the action taken to update the q value of previous state. .This updation is handled by the below equation. Gamma here is the discount factor and alpha is the learning rate.

$$Q(s,a) \leftarrow (1-\alpha)\cdot Q(s,a) + \alpha\cdot(r+\gamma\cdot \max_{a'} Q(s',a'))$$

Model parameter initialised are

Learning rate = 0.01.
Discount factor = 0.90.

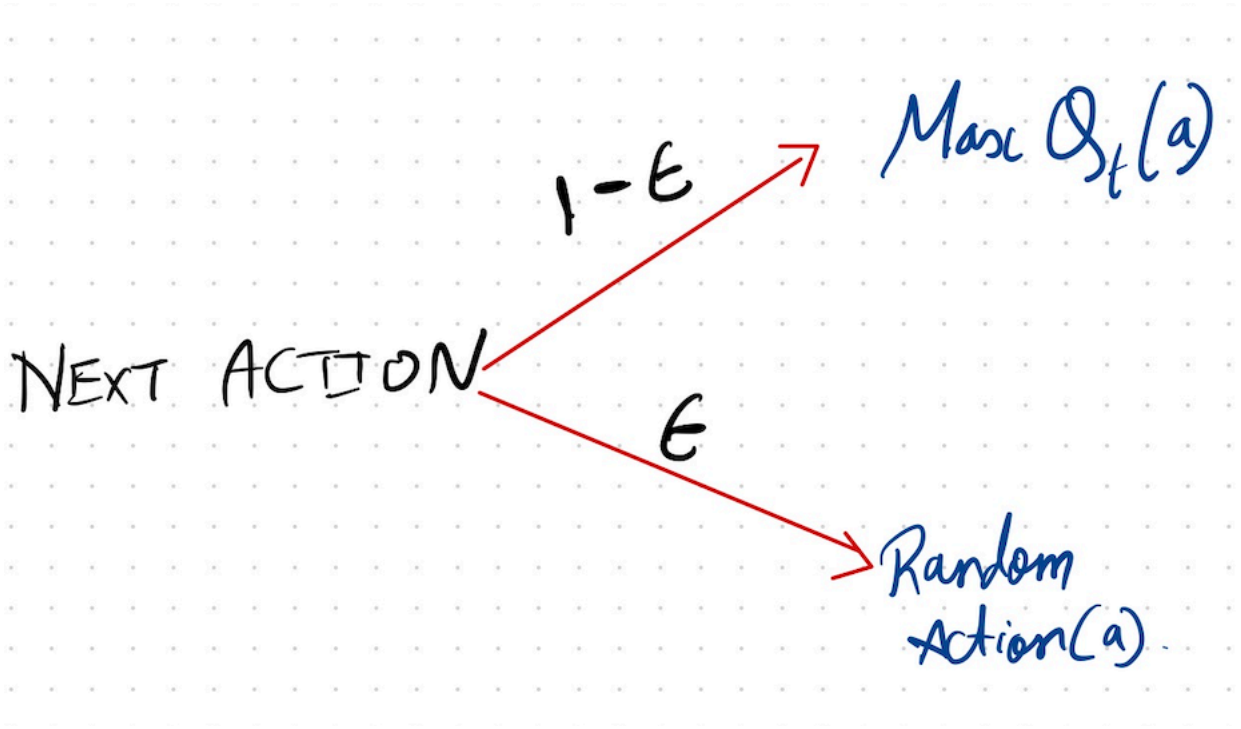
Note: The equation used for training was posted by professor in piazza.

Apart from using the max value of Q to take next action we also use epsilon greedy algorithm where we randomly choose different action rather than the best action according to the Q value.

The process of choosing a random action results in new states, this is called exploration.

The process of choosing the best action is called the exploitation. Since it always yields the maximum reward.

If E(epsilon) is the probability of exploration then the algorithm chooses optimal action for 1-E times and take a random action E times.

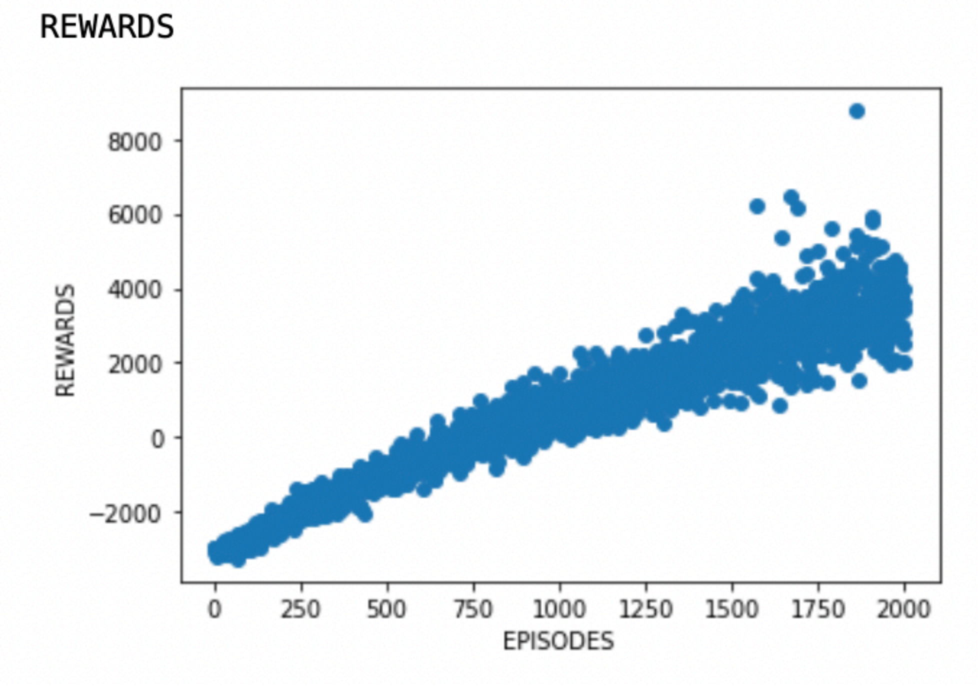


In current implementation epsilon is set to 1 initially then it is decayed at rate of 0.001 exponentially. Delta, the discount factor is set to 0.9. The minimum decayed value of epsilon is 0.01. The learning rate used is 0.01.

The training is run for 2000 episodes and check for the rewards.

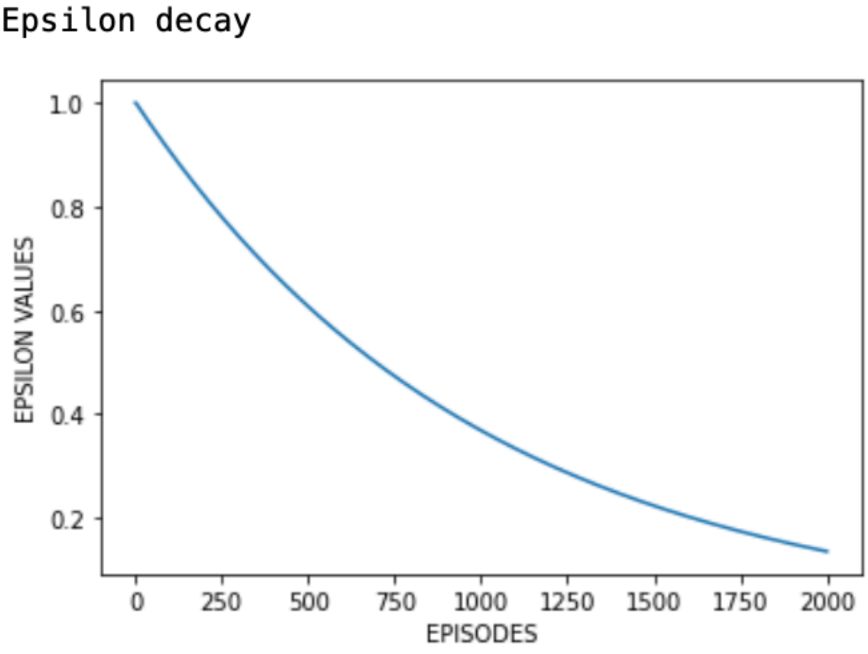
The environment stops every episode if 75% of the cells in q table are effected.

Total rewards and Epsilon decay while training are as shown below.



We notice that the algorithm initially started with very low rewards and as it got trained, the rewards follow an increasing trend.

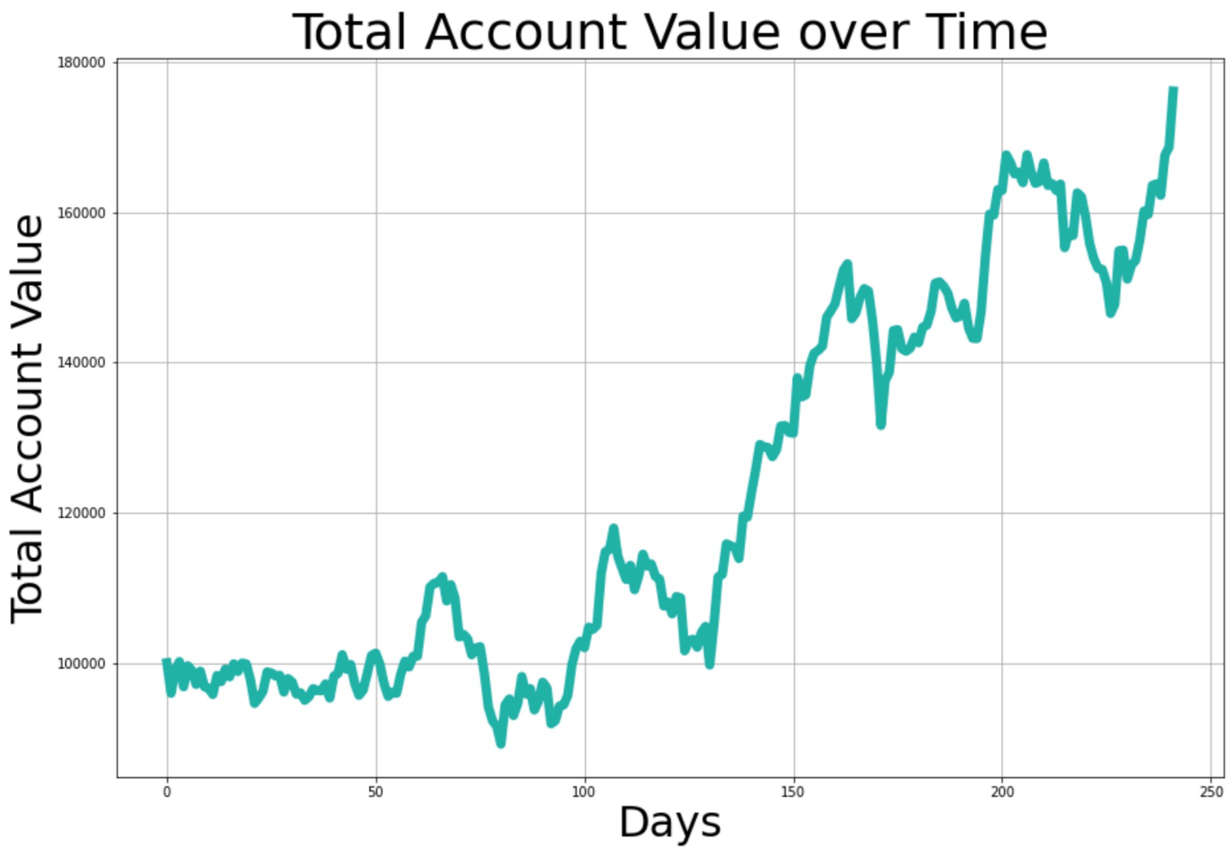
The epsilon decay graph shows the exponential decay of the epsilon, which is the probability of exploration. As the algorithm learns more, the optimal action is chosen rather than randomly selecting the action.



5. Testing phase:

After training the model for 2000 episodes the q-table is frozen and the table is tested on the remaining data, to check whether the algorithm has learnt to make good amount of profit.

To do this we use the testing step defined in the environment and extract the total balance over time (indirectly after sequence of actions performed using the Q table) Below are the results of the same.



```
176207.54779799999
```

It is observed that over a period of 250 days the account value has increased from 100000 to 176207, thus making the q table estimate a valid one.