REPORT ASSIGNMENT 3

COL 333

PART A

QUES 1->A

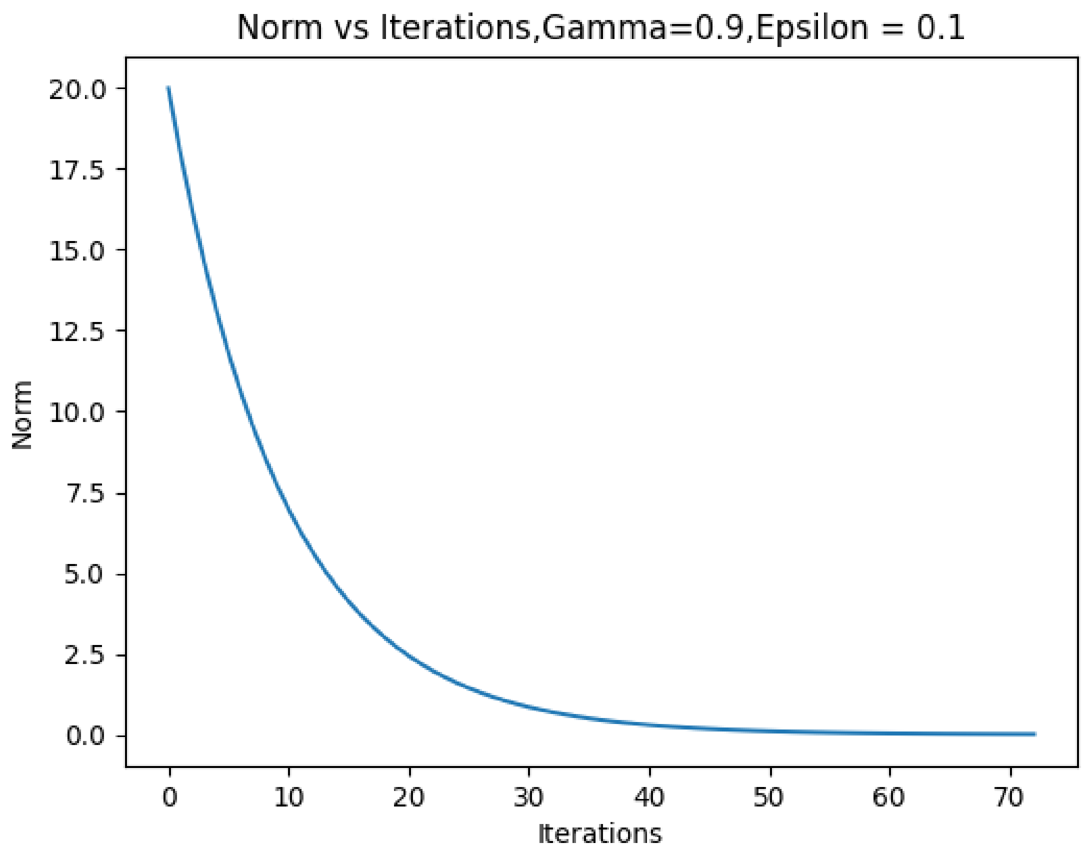
A. Action space : {North ,South ,East ,West ,Pickup ,Putdown}
B. State space : {Location of passenger, Location of Taxi, Location of destination}

⇨ 25 location for Taxi , 5 location for passenger, 4 Location for destination => total states = 25*5*4 = 500 state space

C. Reward space : {-1 for every action,+20 for destination,-10 wrong pickup and putdown}
D. Transition state : the probabilities of the states i.e P(s'|s,a)

QUES 2->A

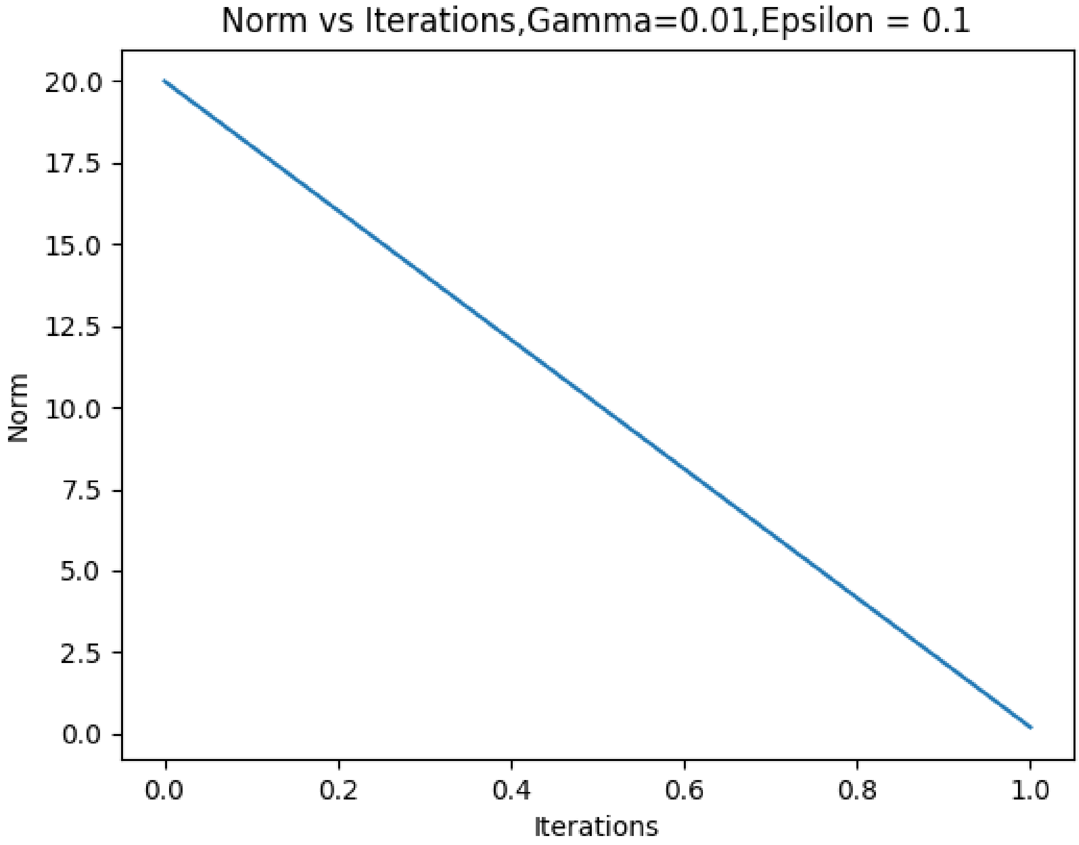
➢ **Iterations = 73 gamma =0.9 epsilon = 0.1**



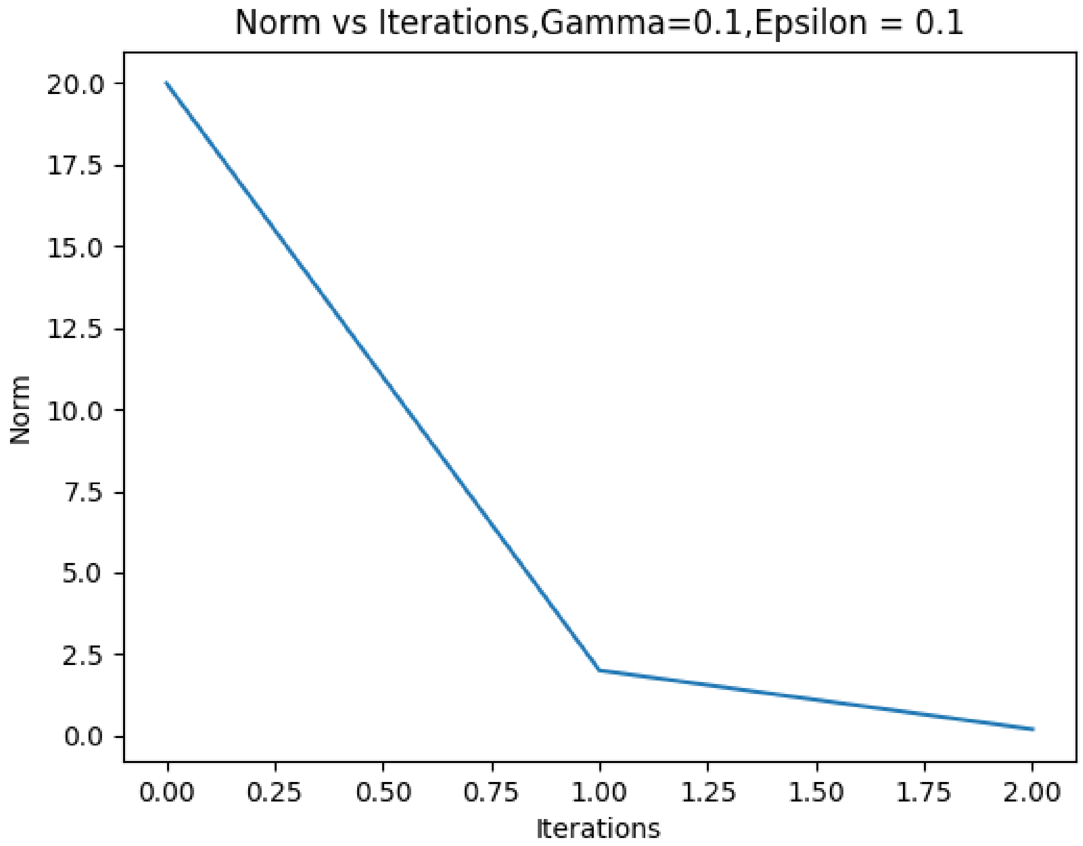
In this part we can able to see that the curve for discount fact 0.9 and number of iterations as iterations is increase the norm is decrease and finally our graph is converged.

QUES 2 B->

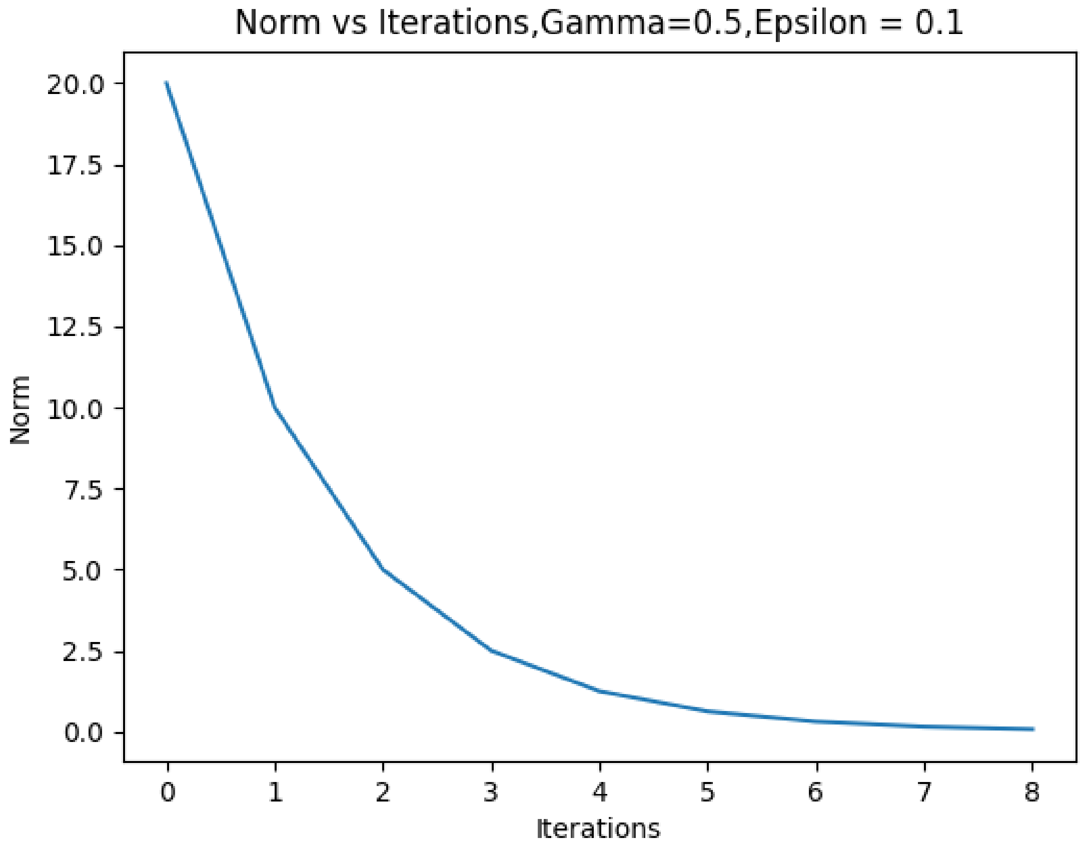
- **Iterations = 2 gamma =0.01 epsilon = 0.1**



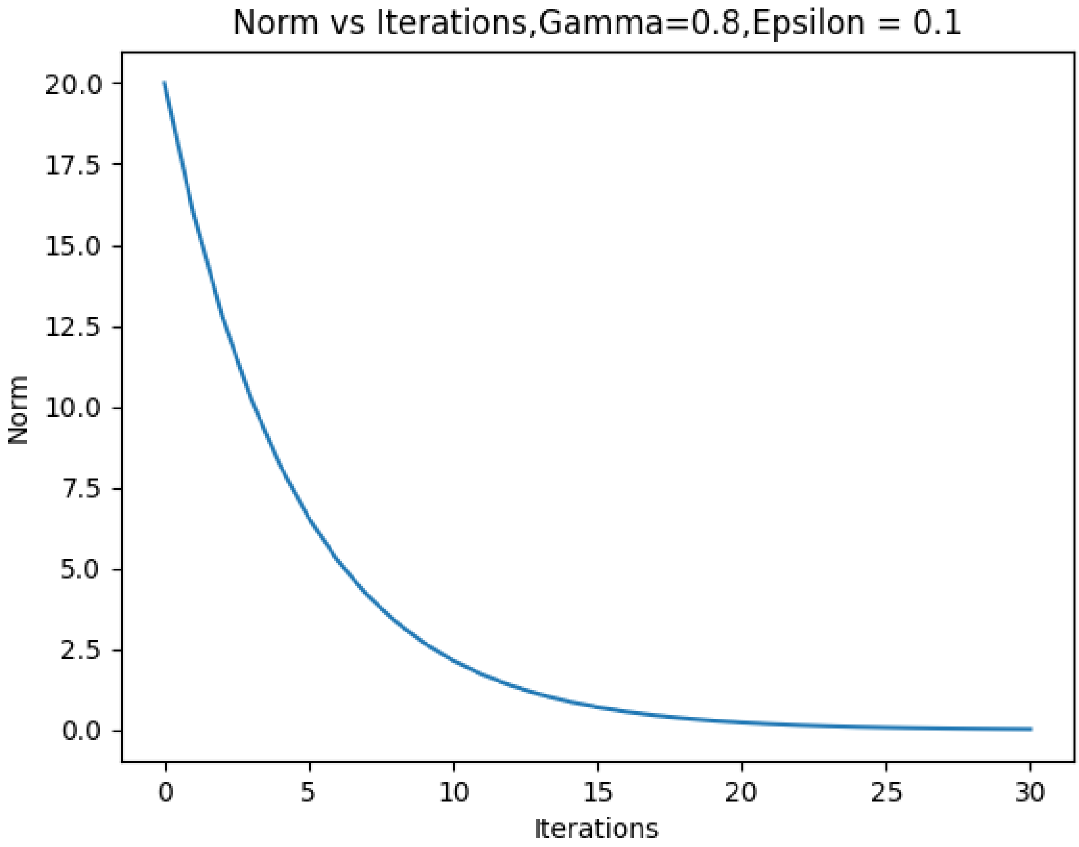
- **Iterations = 3 gamma =0.1 epsilon = 0.1**



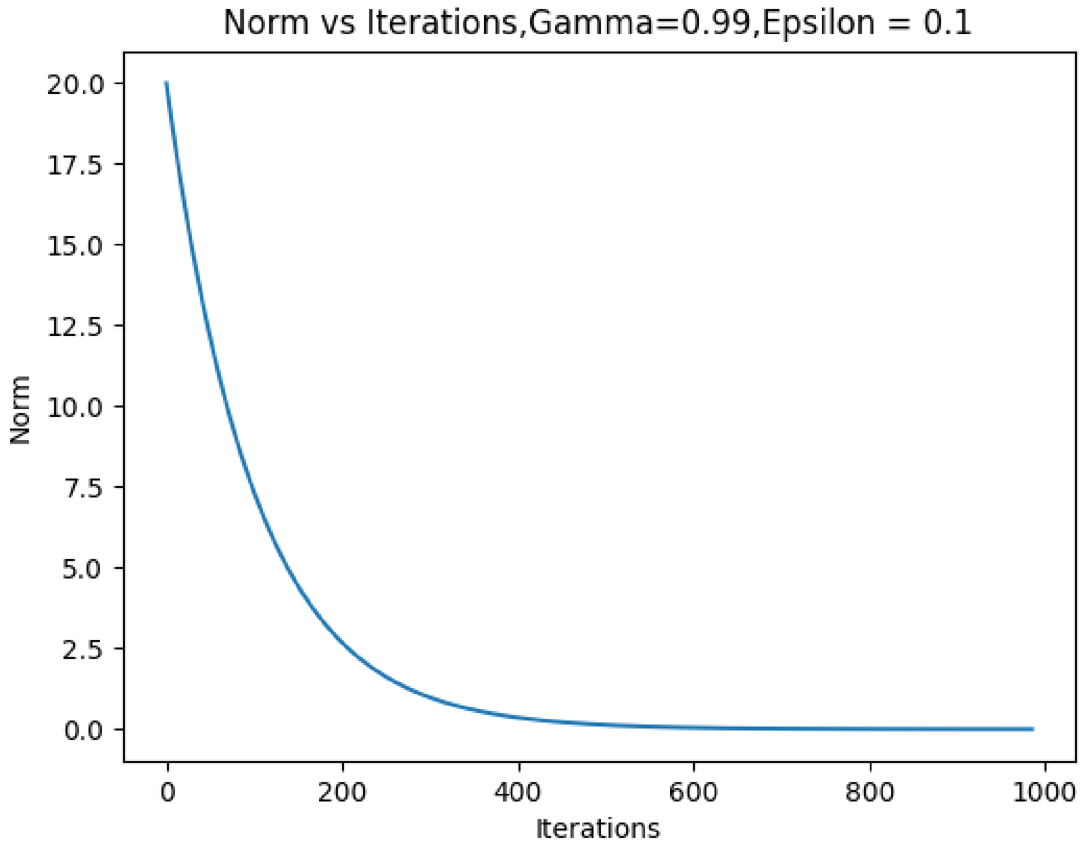
- **Iterations = 9 gamma =0.5 epsilon = 0.1**



- **Iterations = 31 gamma =0.8 epsilon = 0.1**



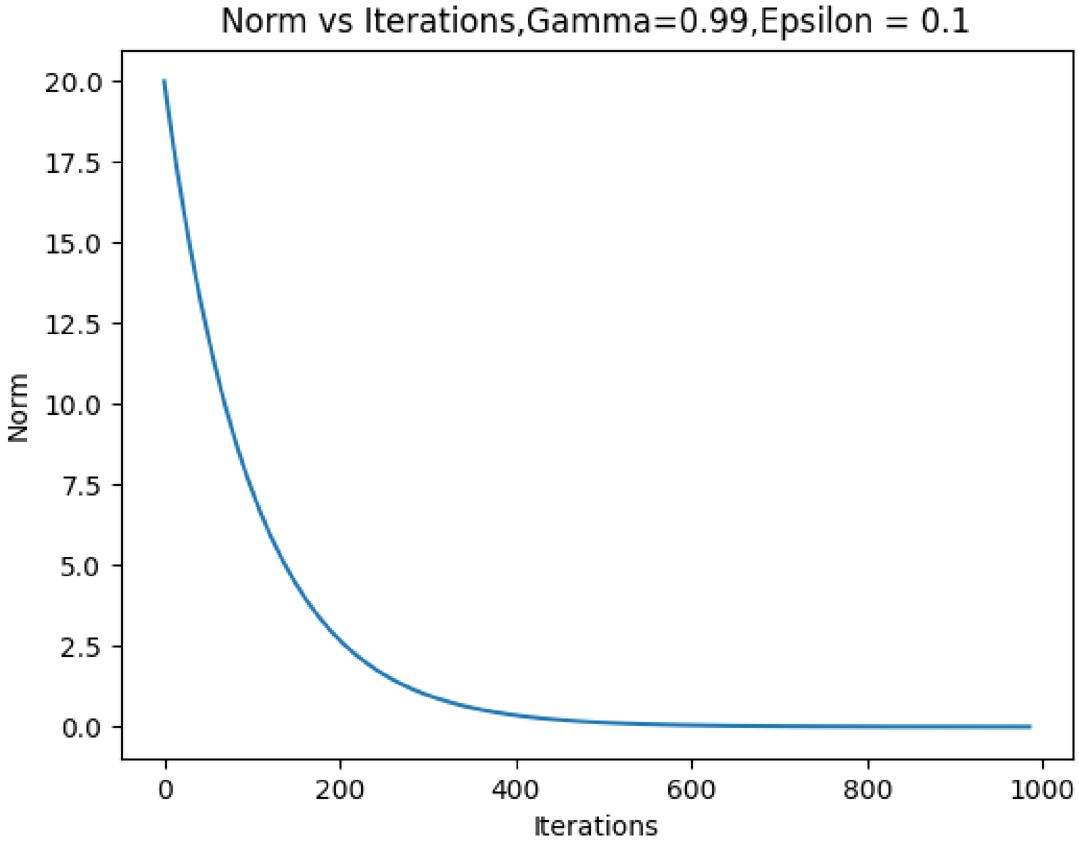
- **Iterations = 986 gamma =0.99 epsilon = 0.1**



In this part we can see that we analyze our data for the different values of the gamma(discount factors) so form here we can easily able to see that as we increase gamma then number of iteratioins is also increase for the convergence. This is because when gamma is low then we are propagating less =>our delta (error bound) was reached early same opposite in the reverse case when gamma is high.

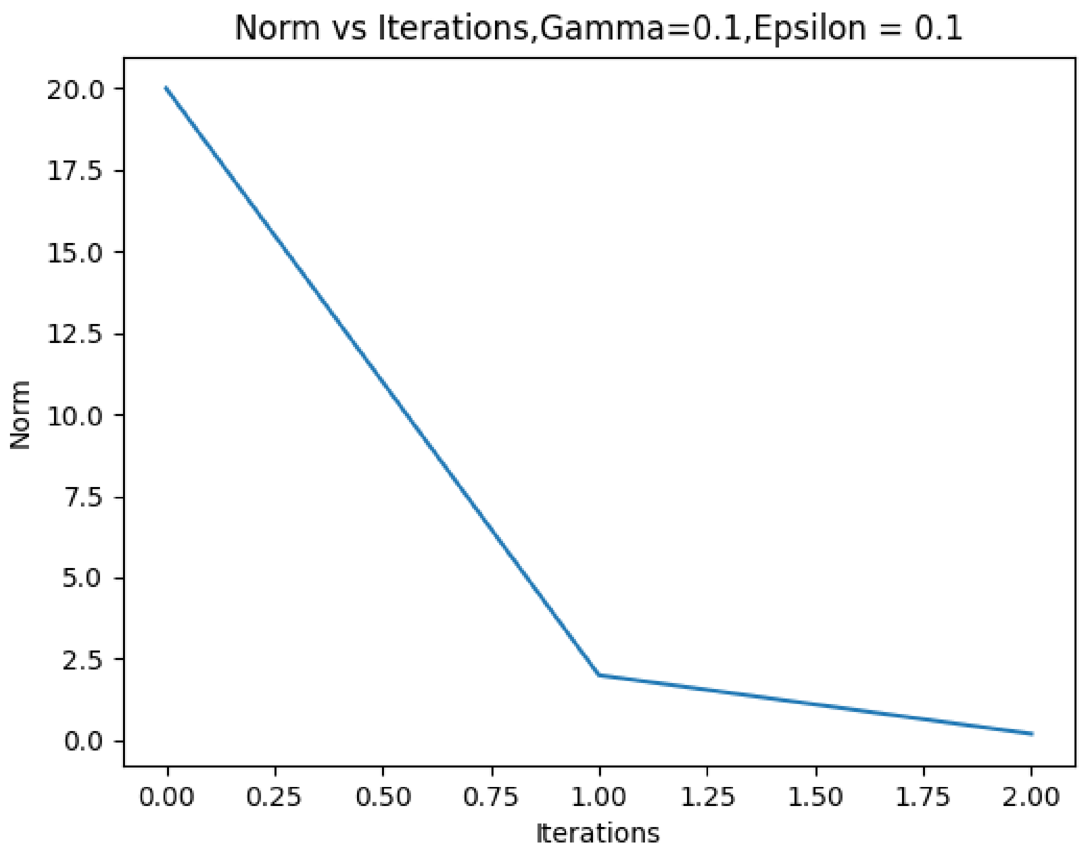
QUES 2->C

- **Iterations = 986 gamma =0.99 epsilon = 0.1**



First 20 states of policy = ['Putdown', 'North', 'North', 'North', 'South', 'North', 'North', 'North', 'South', 'North', 'East', 'East', 'South', 'North', 'East', 'East', 'South', 'Putdown', 'South', 'South']

- **Iterations = 3 gamma =0.1 epsilon = 0.1**



First 20 states of policy = ['Putdown', 'Putdown', 'Putdown', 'Putdown', 'South', 'Putdown', 'Putdown', 'Putdown', 'South', 'North', 'Putdown', 'Putdown', 'Putdown', 'North', 'Putdown', 'Putdown', 'Putdown', 'Putdown', 'Putdown', 'Putdown']

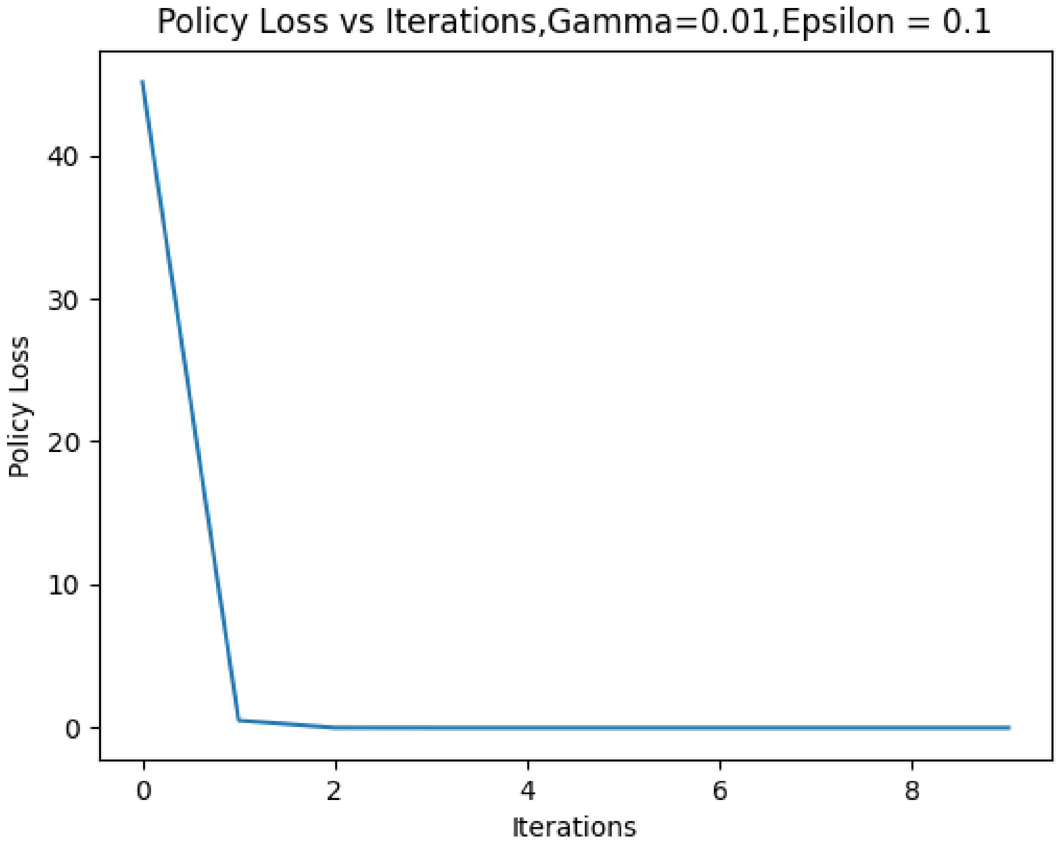
In this part we have to see the optimal policy along with the corresponding states and for that which will be taking two value of discount factors. We can conclude that if the discount factor is very low than the taxi will try to take rewards as early as possible and if it very large than the taxi will try to take rewards anytime and for that we will be able to see the corresponding states vs rewards.

QUES 3->A

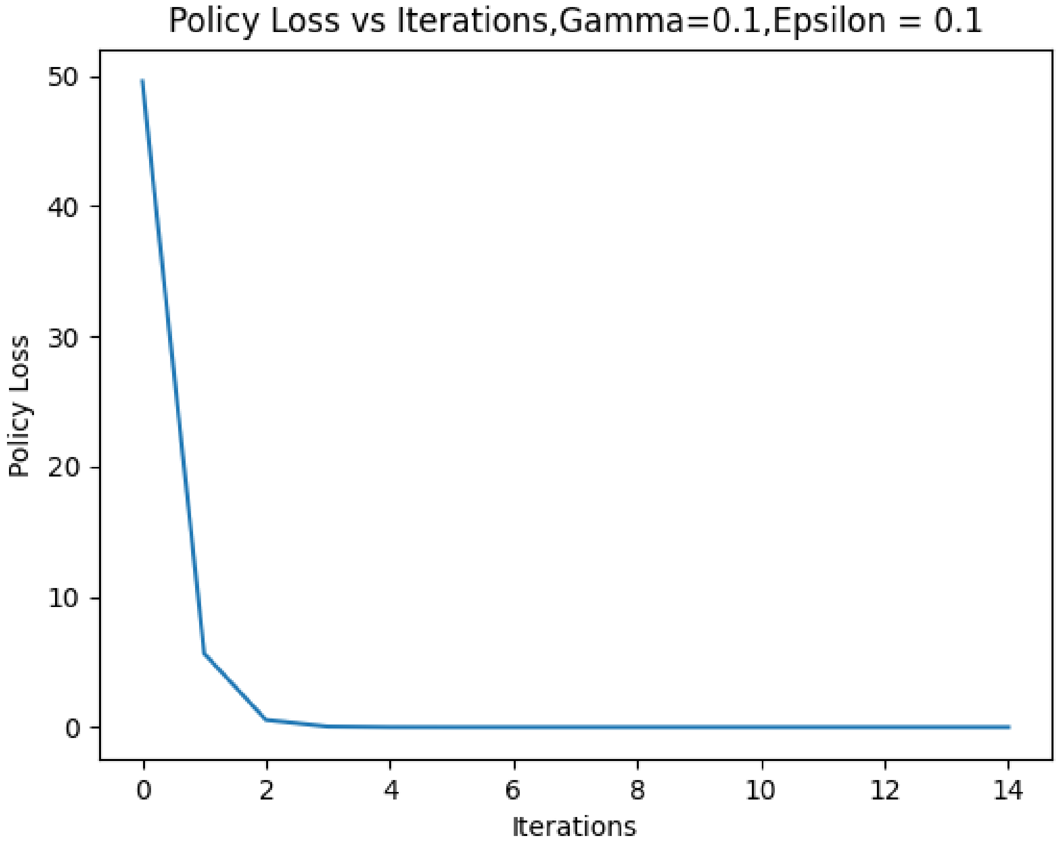
In policy iteration we will be evaluating any policy and then keep it updating and finally it converges to optimal policy. We will start with a random policy and then will keep on updating it using the argmax (Adjacent state utilities / among 6 actions) and then will change the policy according to it.

QUES 3->B

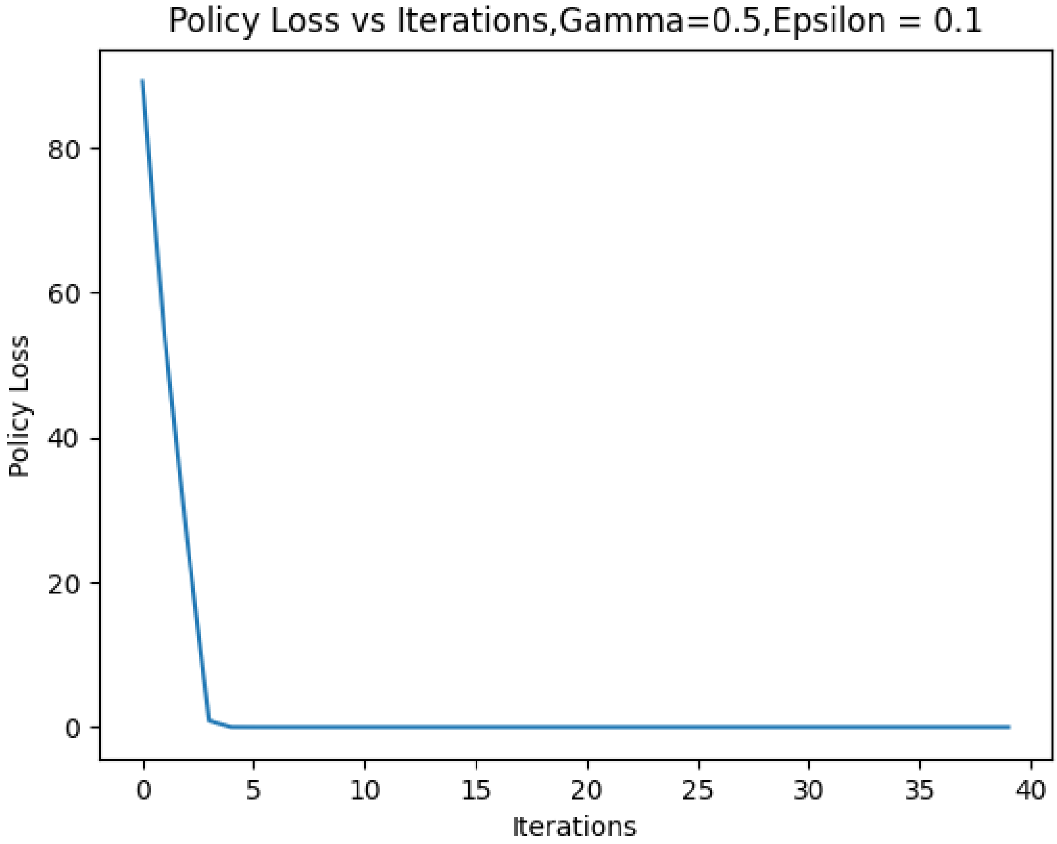
- **Iterations = 10 gamma =0.01 epsilon = 0.1**



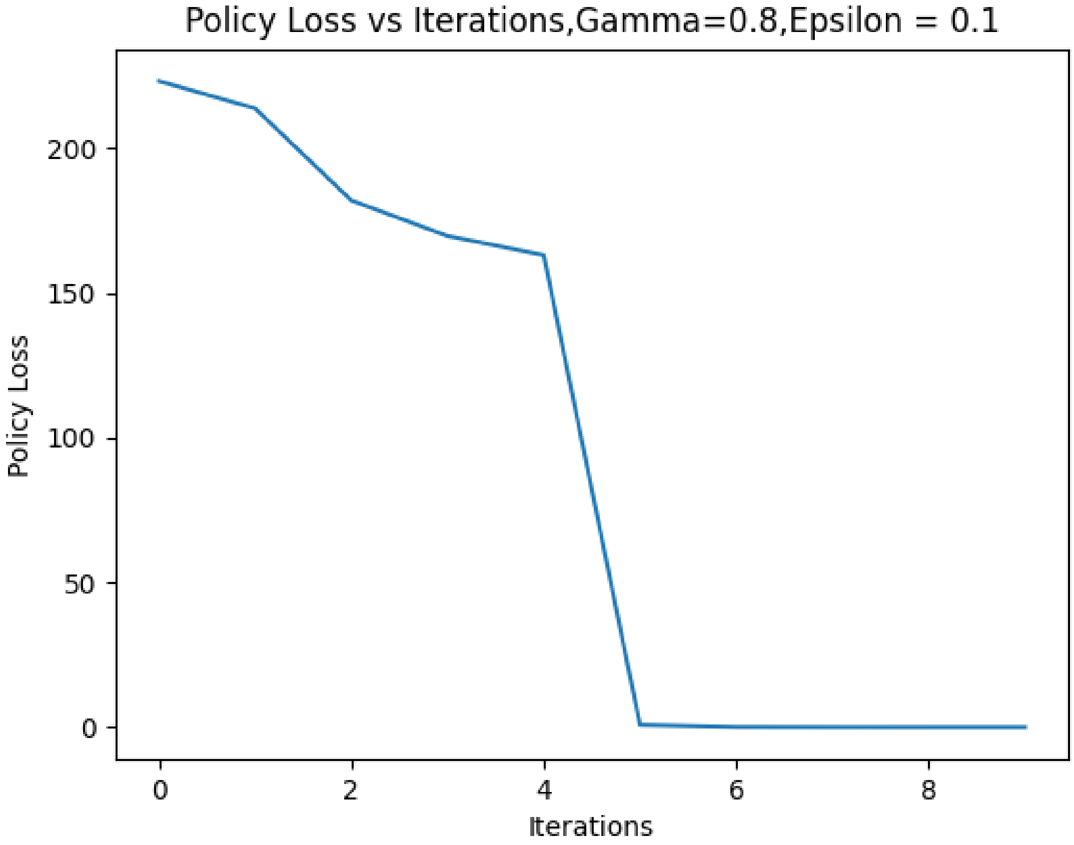
- **Iterations = 15 gamma =0.1 epsilon = 0.1**



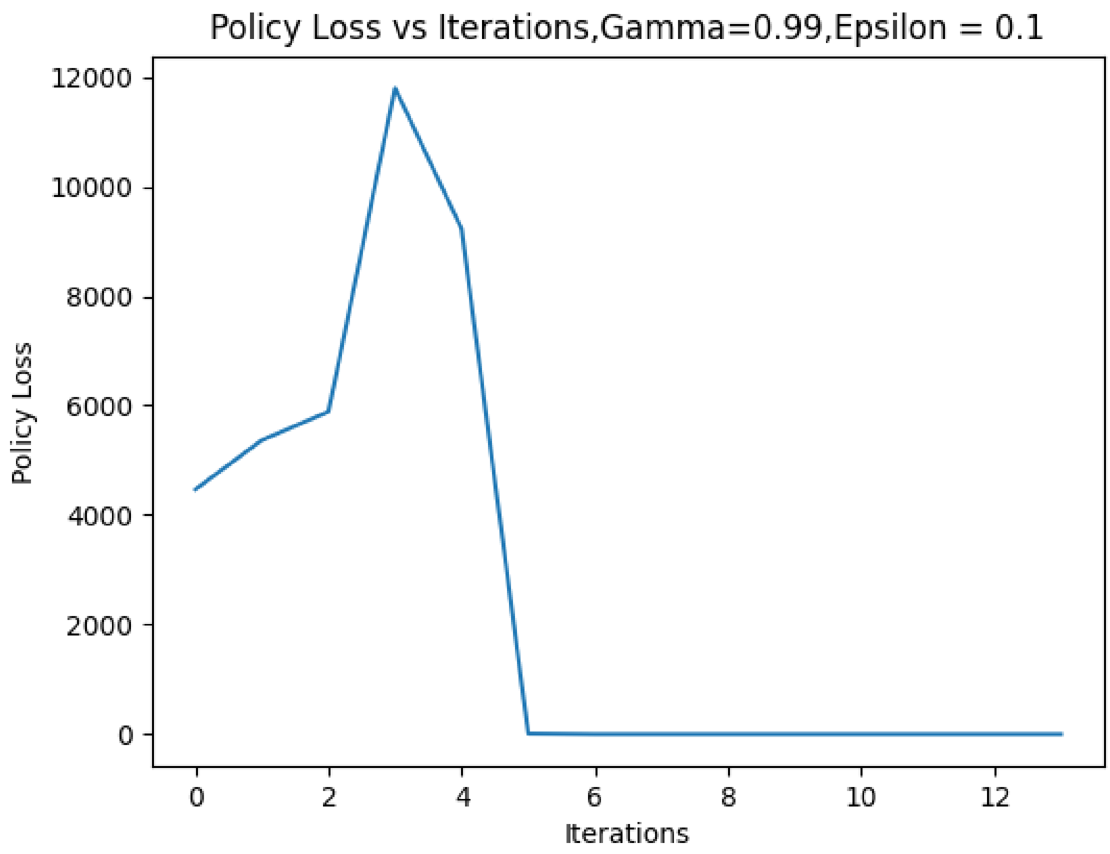
- **Iterations = 40 gamma =0.5 epsilon = 0.1**



- **Iterations = 10 gamma =0.8 epsilon = 0.1**



- **Iterations = 14 gamma =0.99 epsilon = 0.1**

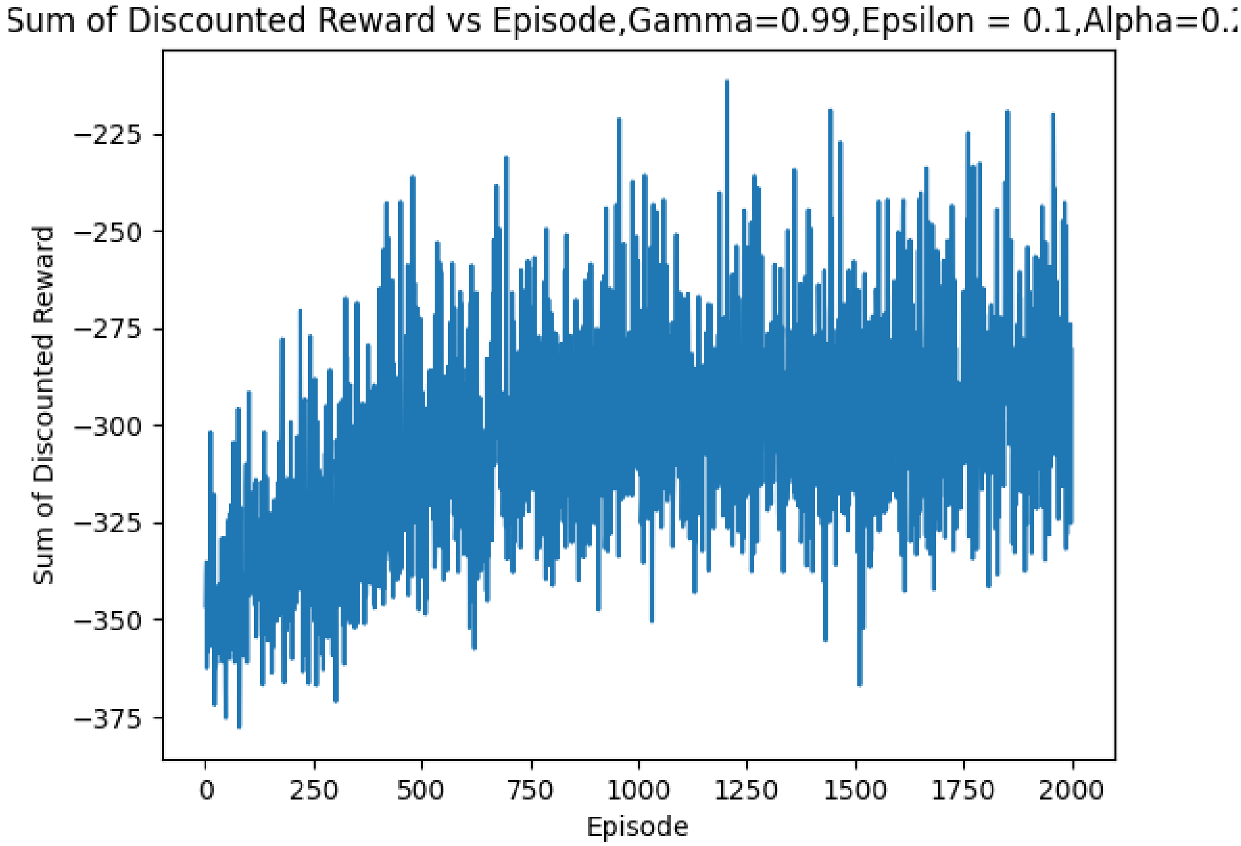


Now we are studying the effect of change in the value of gamma on our MDP which is the taxi domain problem for us. We can clearly see that numbers of iterations are increasing as we are increasing our gamma and that is because gamma is the discount factor which tells us about the preference of present rewards as compared to future rewards. Less is the value of gamma, more we will tend to receive rewards earlier.

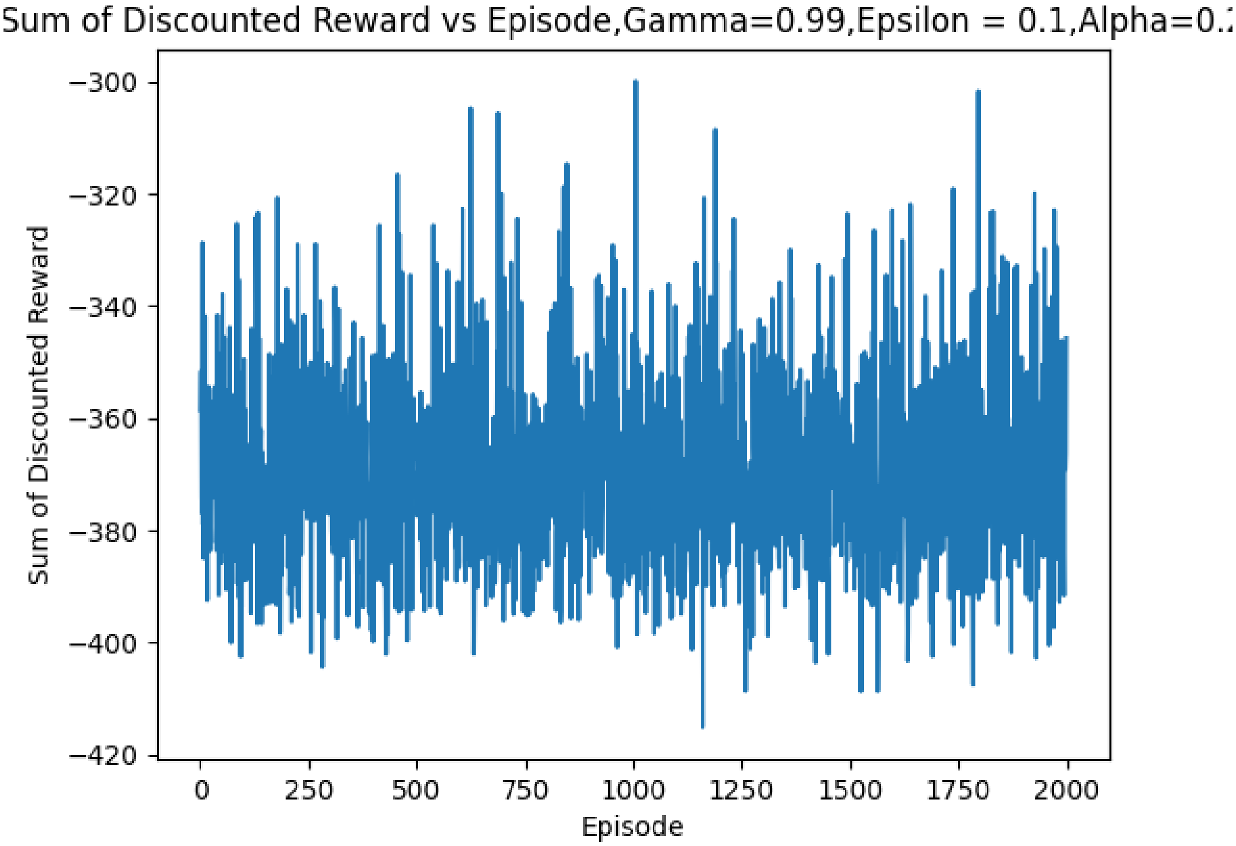
PART B

QUES 2->

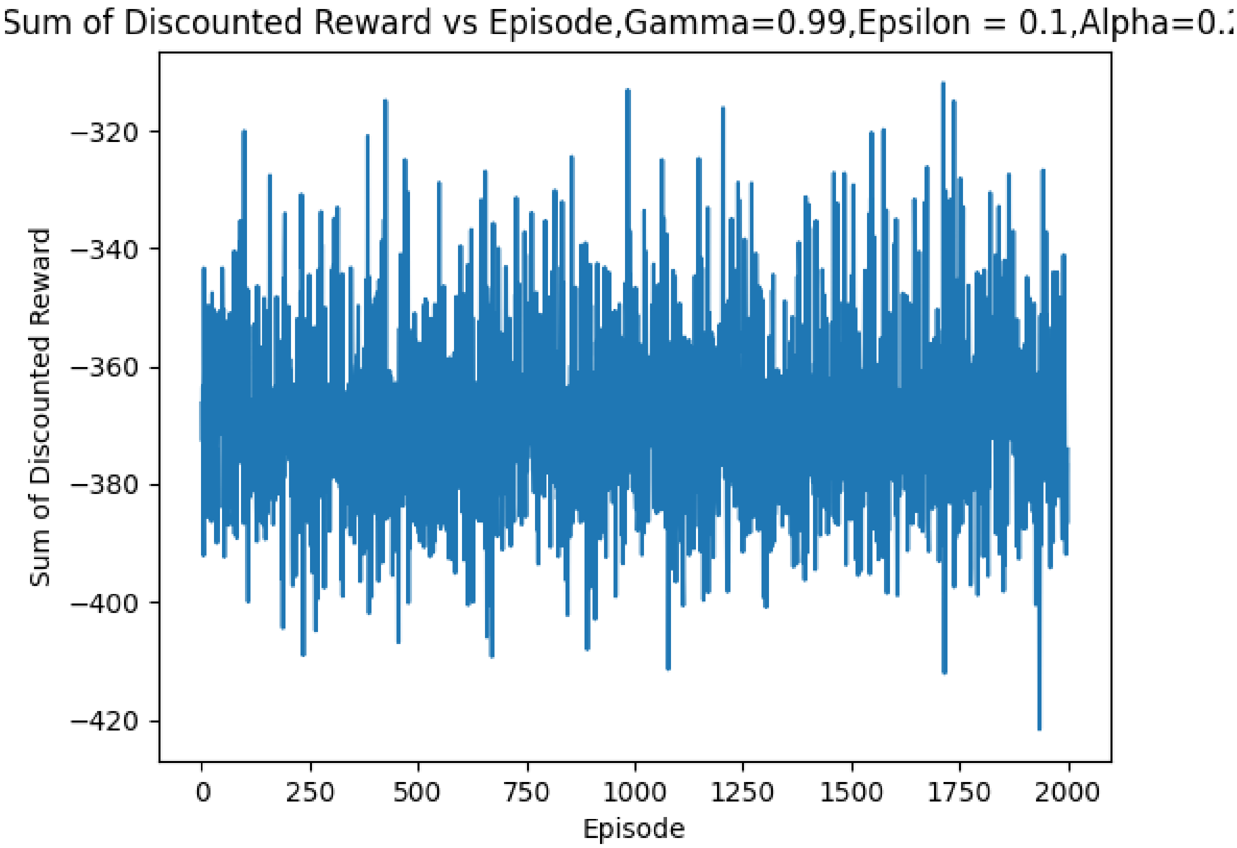
- **Q-learning a 1-Step look-Ahead using constant Epsilon**



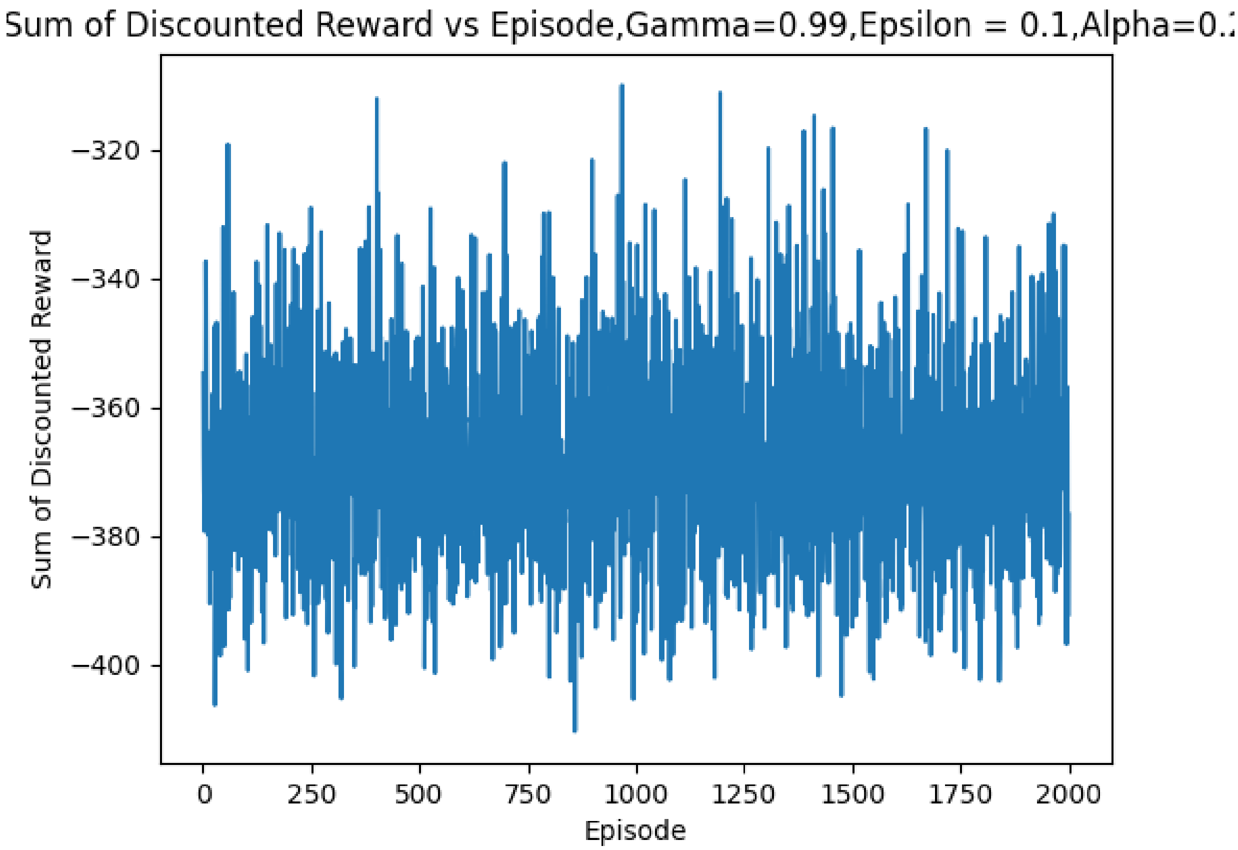
- **Q-learning a 1-Step look-Ahead using decaying Epsilon**



- **Q-learning SARSA using constant Epsilon**



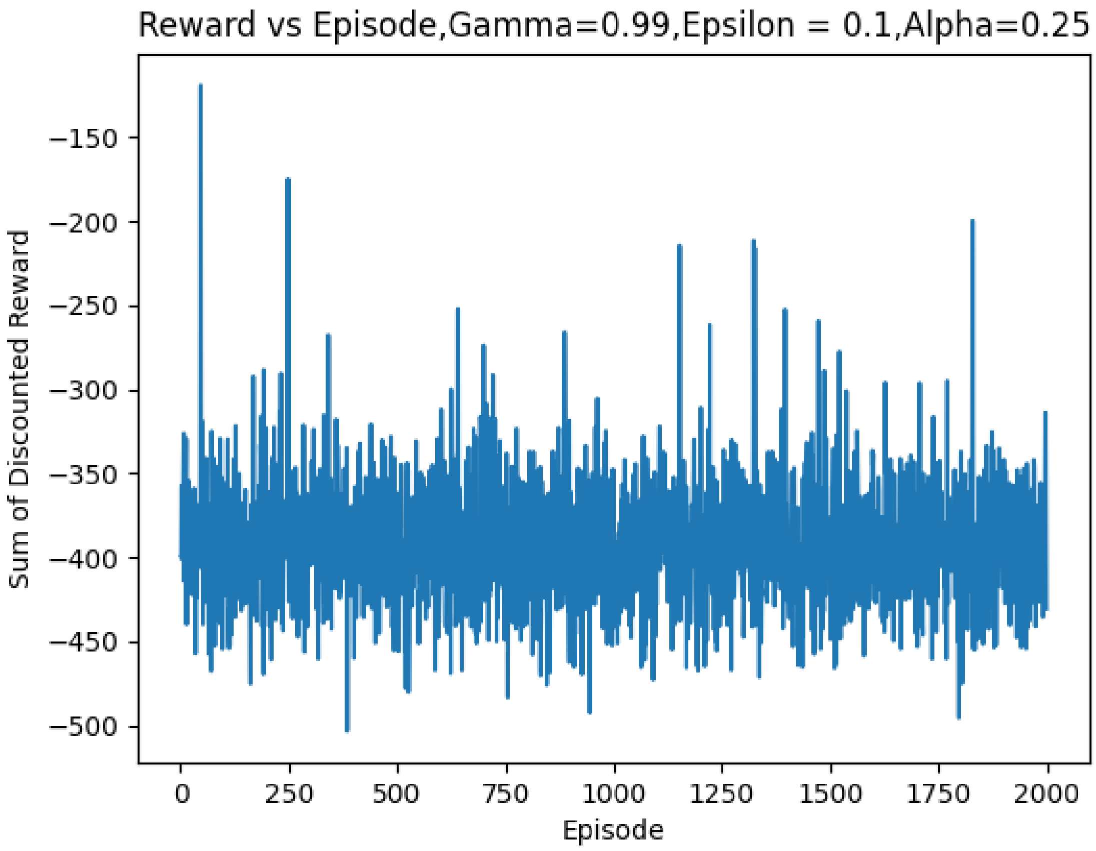
- **Q-learning SARSA using decaying Epsilon**



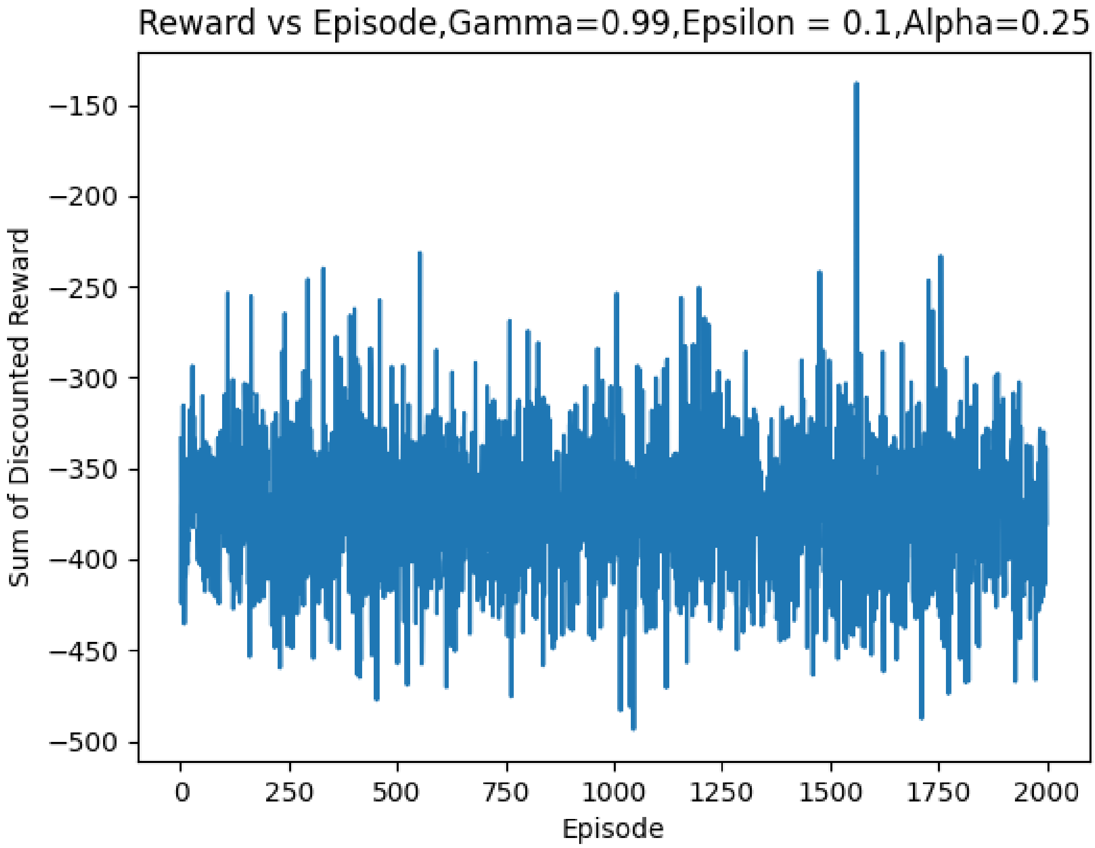
Now we have drawn the sum of discounted rewards with the episode numbers and we can see that in these cases the sum of discounted rewards is converging and we can see that it is converging to a large negative values in all these cases i.e. nearly -350 and that is the case in all of them. The large negative reward can be because of the fact that the taxi is taking a long time to reach the destination and in the process the taxi driver is getting a large negative reward.

QUES 3->

- **Y (Initial passenger depot) = (0,0) and R (Initial taxi location) = (3,0), G (passenger destination location) = (0,4)**

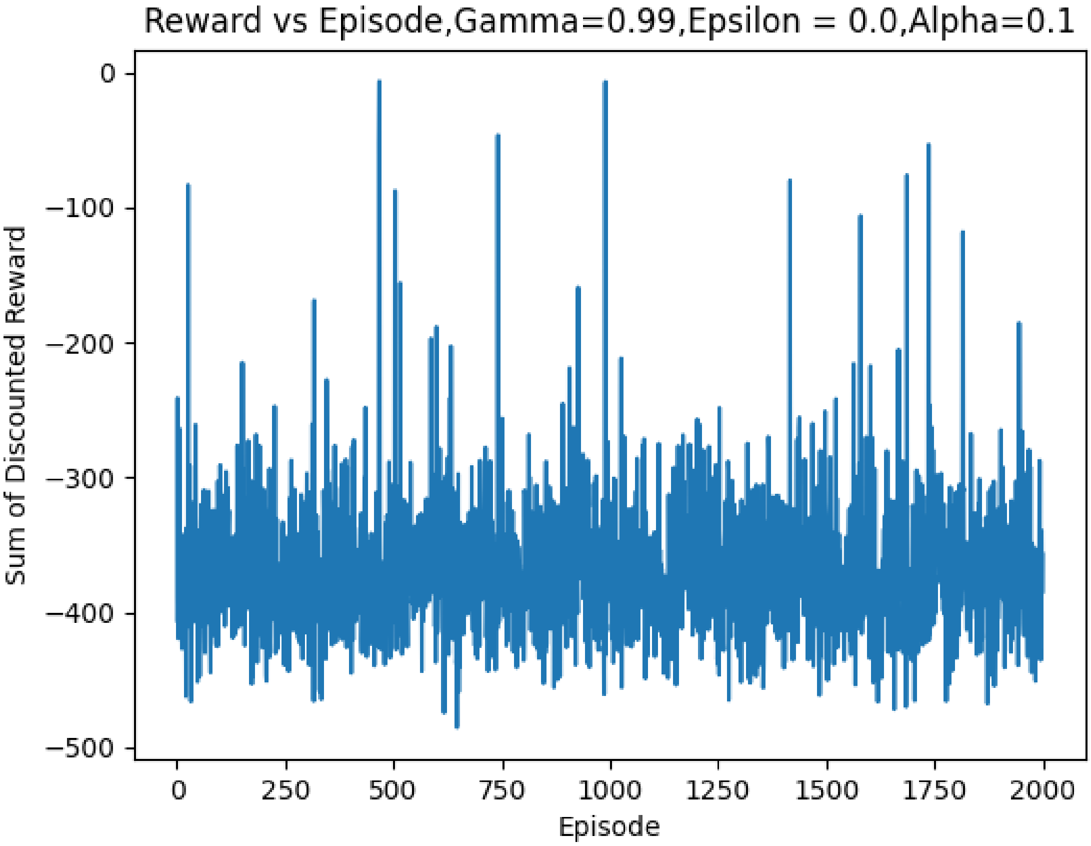


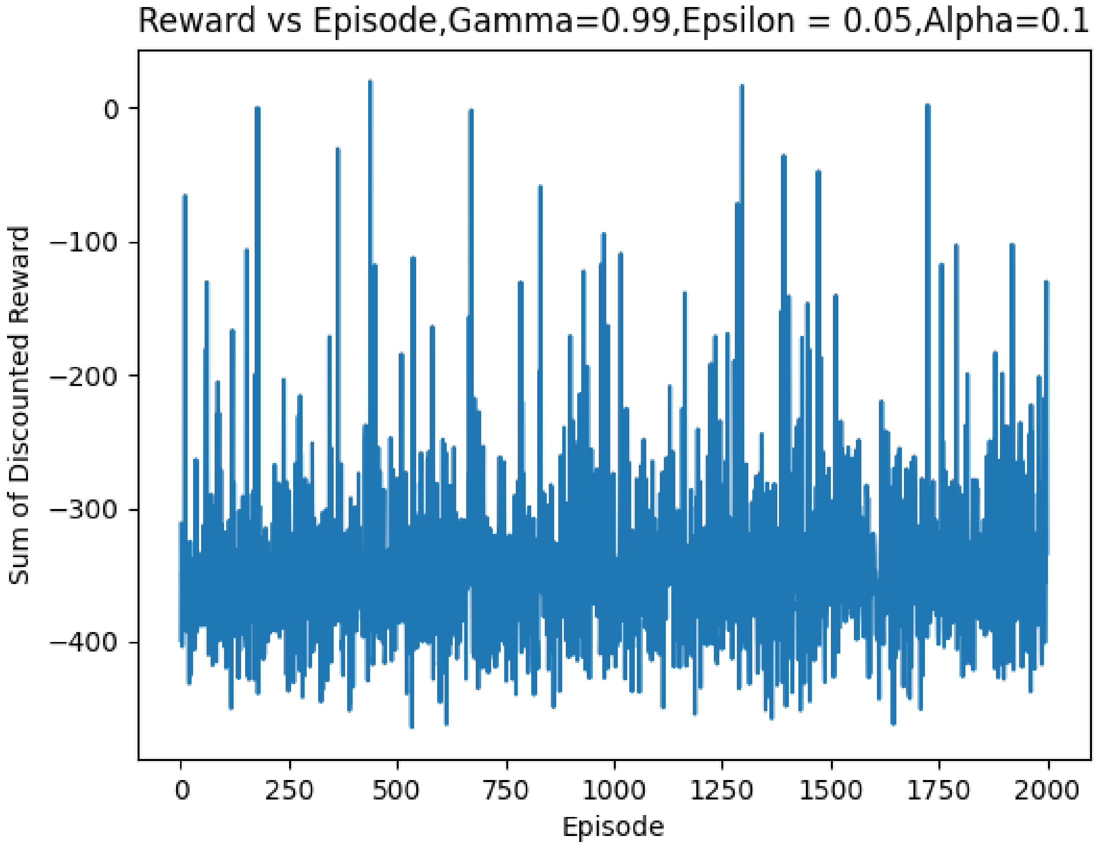
- **Y (Initial passenger depot) = (4,4) and R (Initial taxi location) = (3,0), G (passenger destination location) = (0,4)**

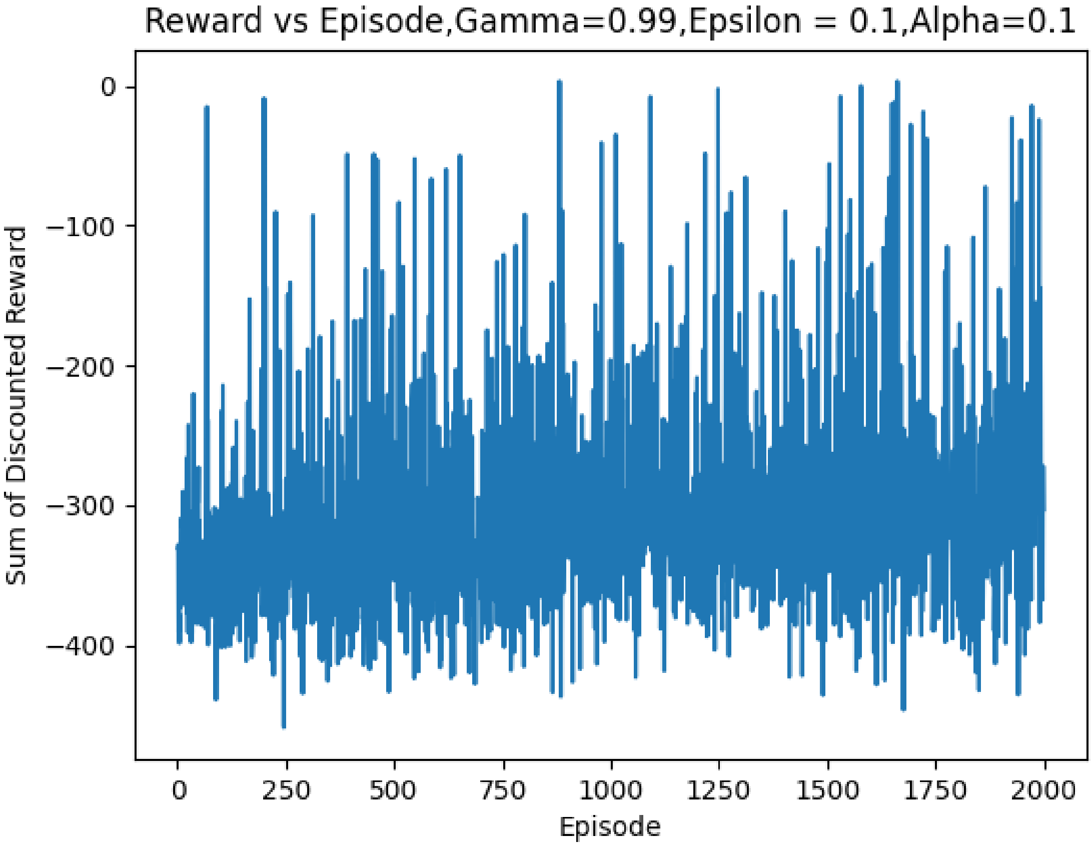


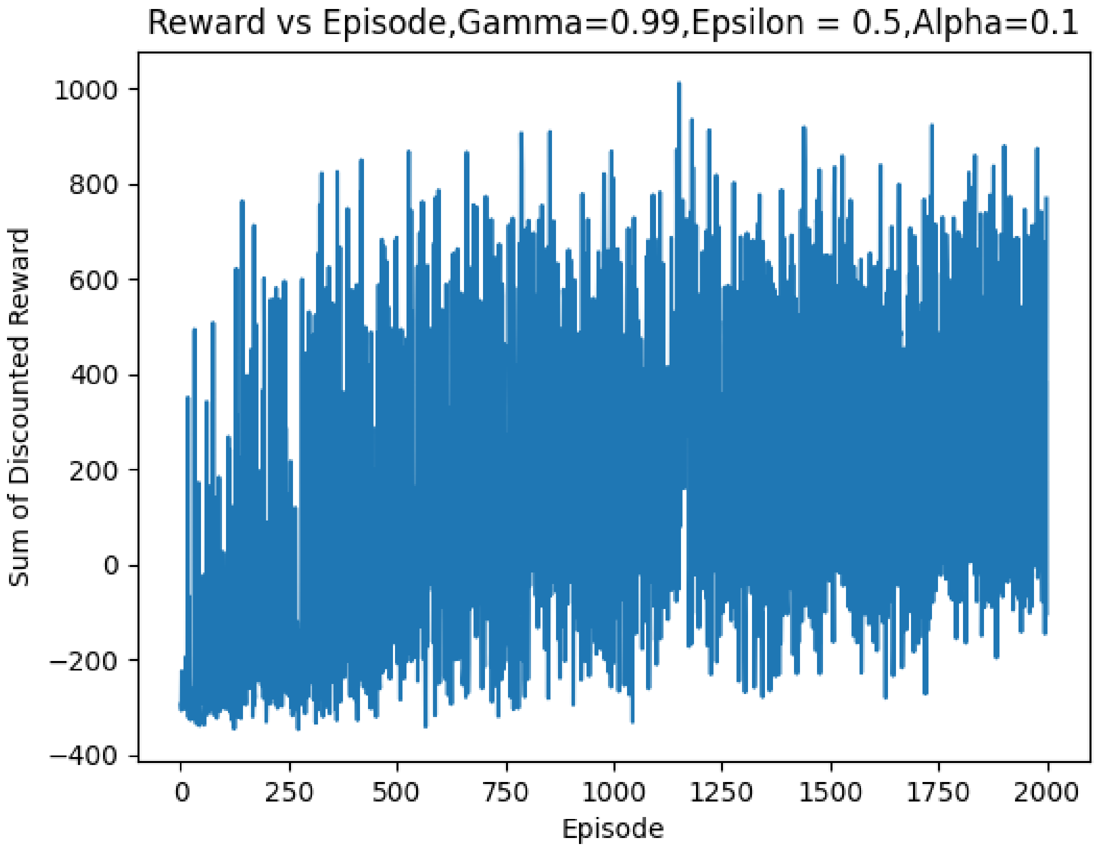
Among all the above methods the one which is getting highest cumulative rewards is Q-Learning with constant epsilon. In it the rewards accumulated are more than any of the other three. Now, we will be changing the instance of the problems chosen and then we deduced that on changing the instance of our taxi domain the rewards are also changing a bit.

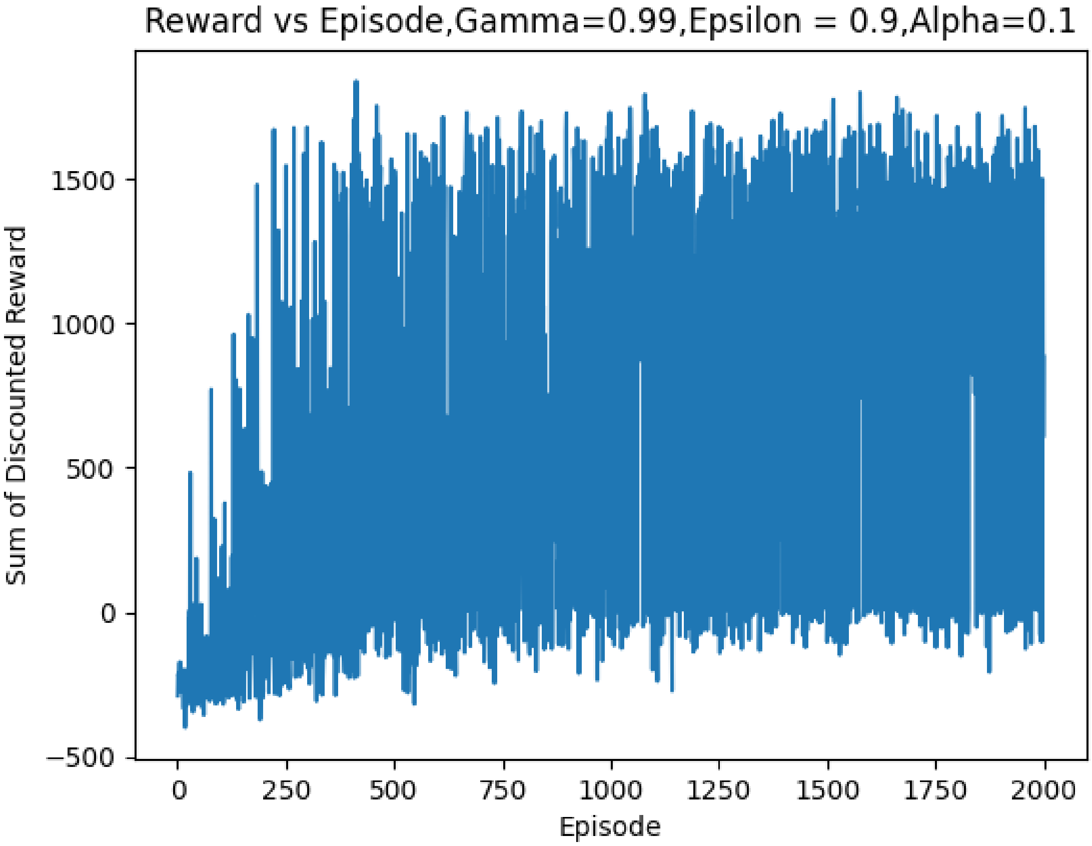
QUES 4->


Reward vs Episode,Gamma=0.99,Epsilon = 0.0,Alpha=0.1
Sum of Discounted Reward
0
−100
−200
−300
−400
−500
0
250
500
750
1000
1250
1500
1750
2000
Episode

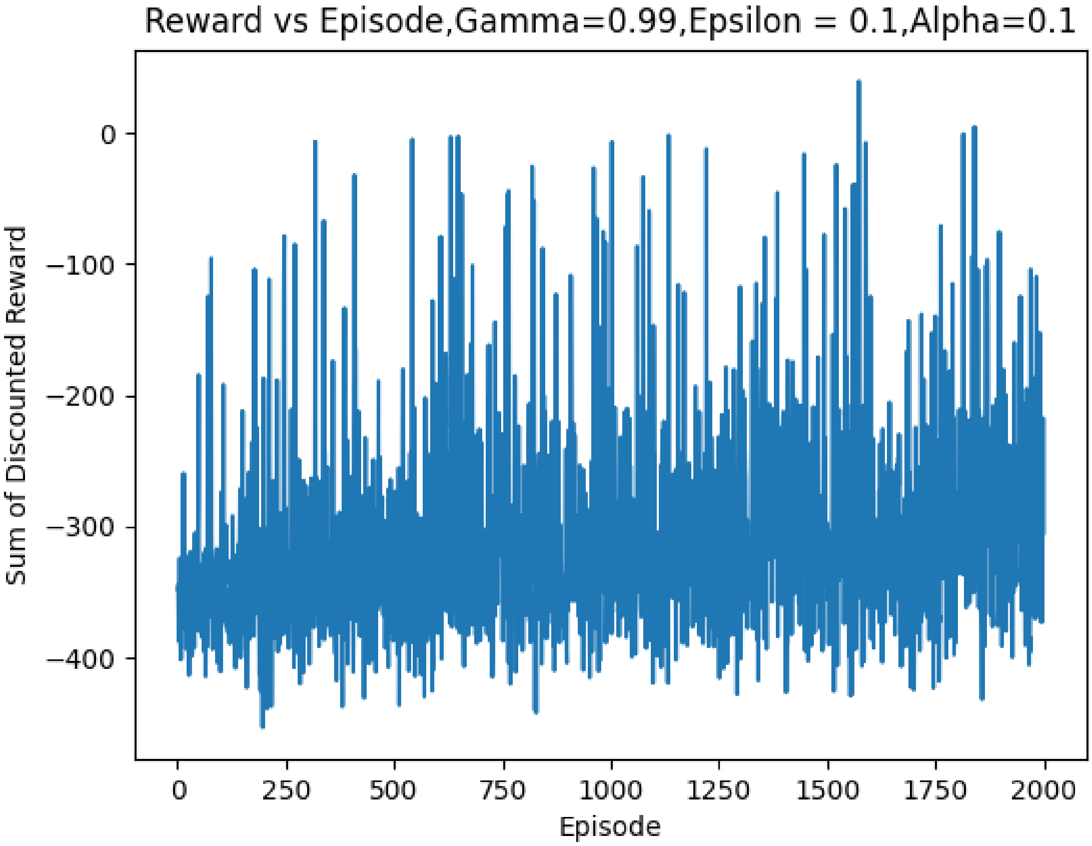

Reward vs Episode,Gamma=0.99,Epsilon = 0.05,Alpha=0.1
Sum of Discounted Reward
0
−100
−200
−300
−400
0
250
500
750
1000
1250
1500
1750
2000
Episode

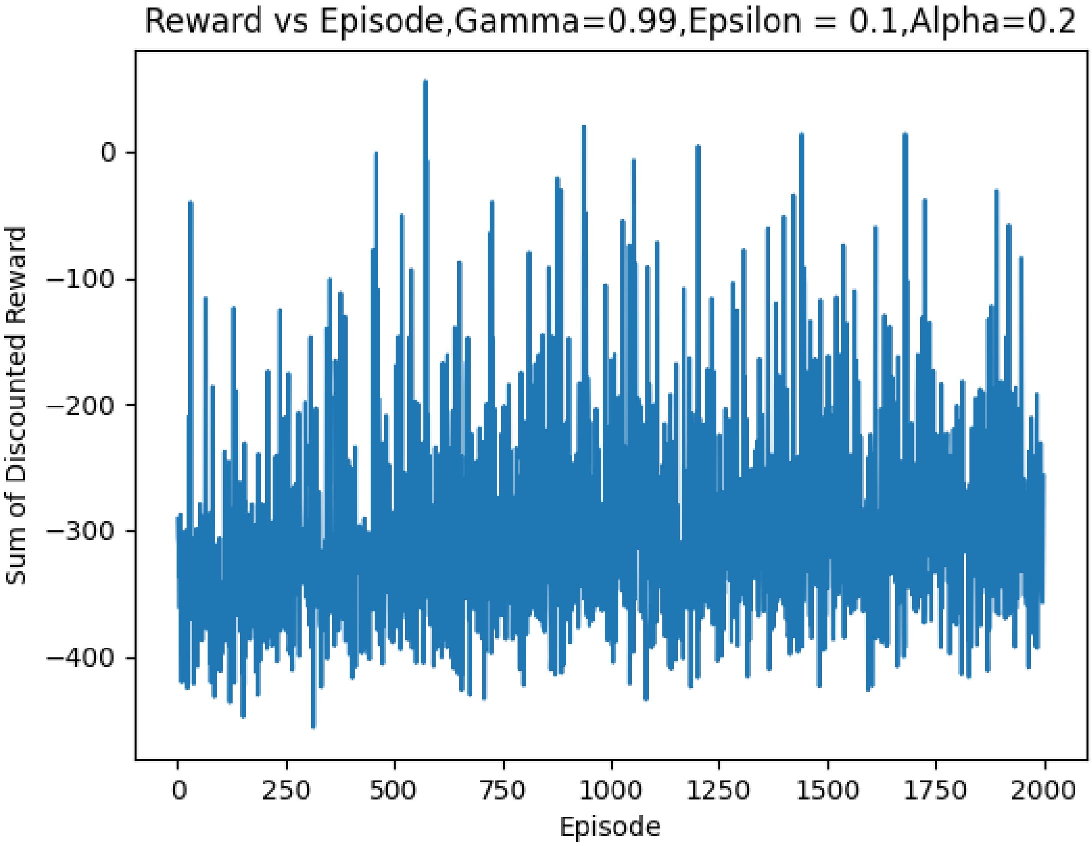

Reward vs Episode,Gamma=0.99,Epsilon = 0.1,Alpha=0.1
Sum of Discounted Reward
0
−100
−200
−300
−400
0
250
500
750
1000
1250
1500
1750
2000
Episode

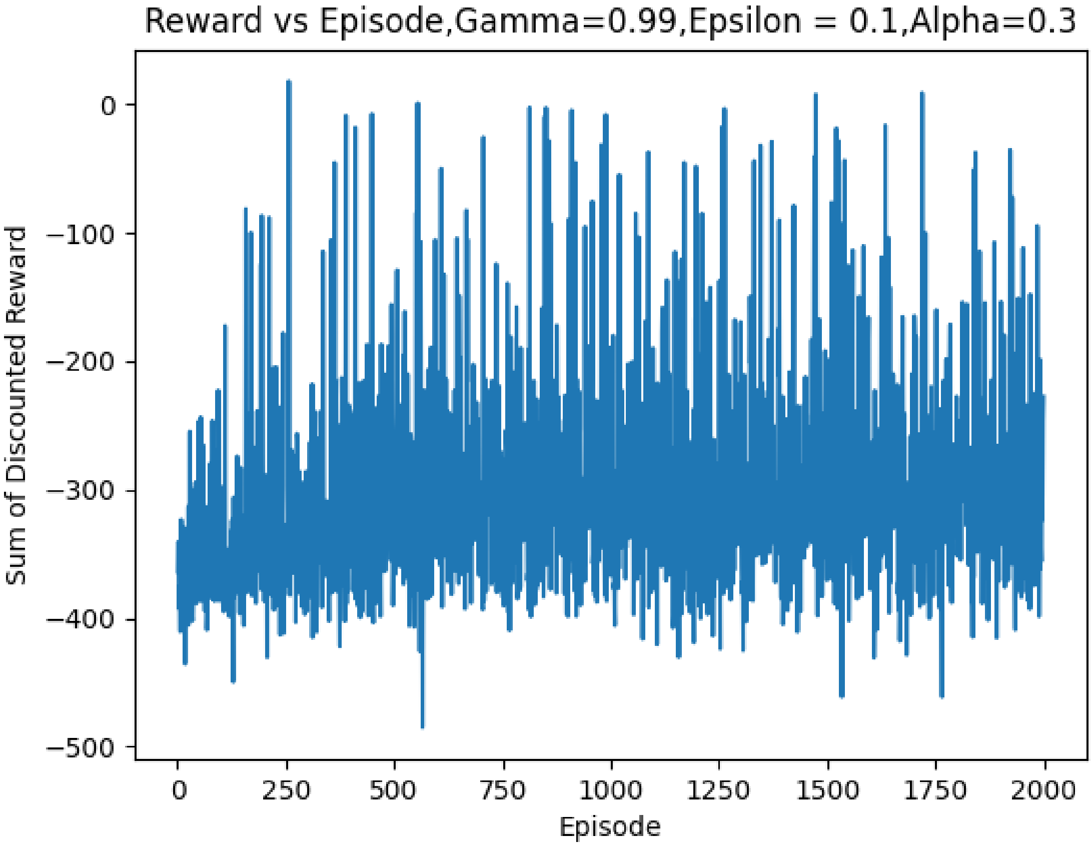

Reward vs Episode,Gamma=0.99,Epsilon = 0.5,Alpha=0.1
Sum of Discounted Reward
1000
800
600
400
200
0
−200
−400
0
250
500
750
1000
1250
1500
1750
2000
Episode

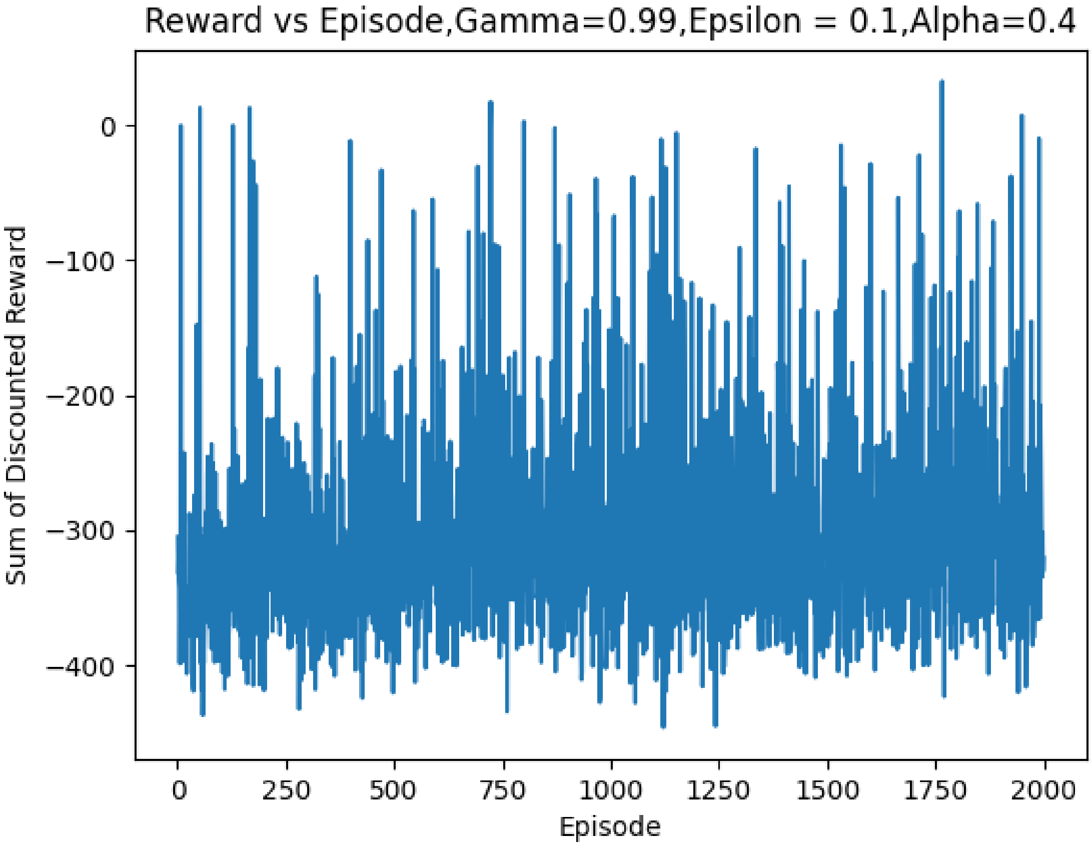


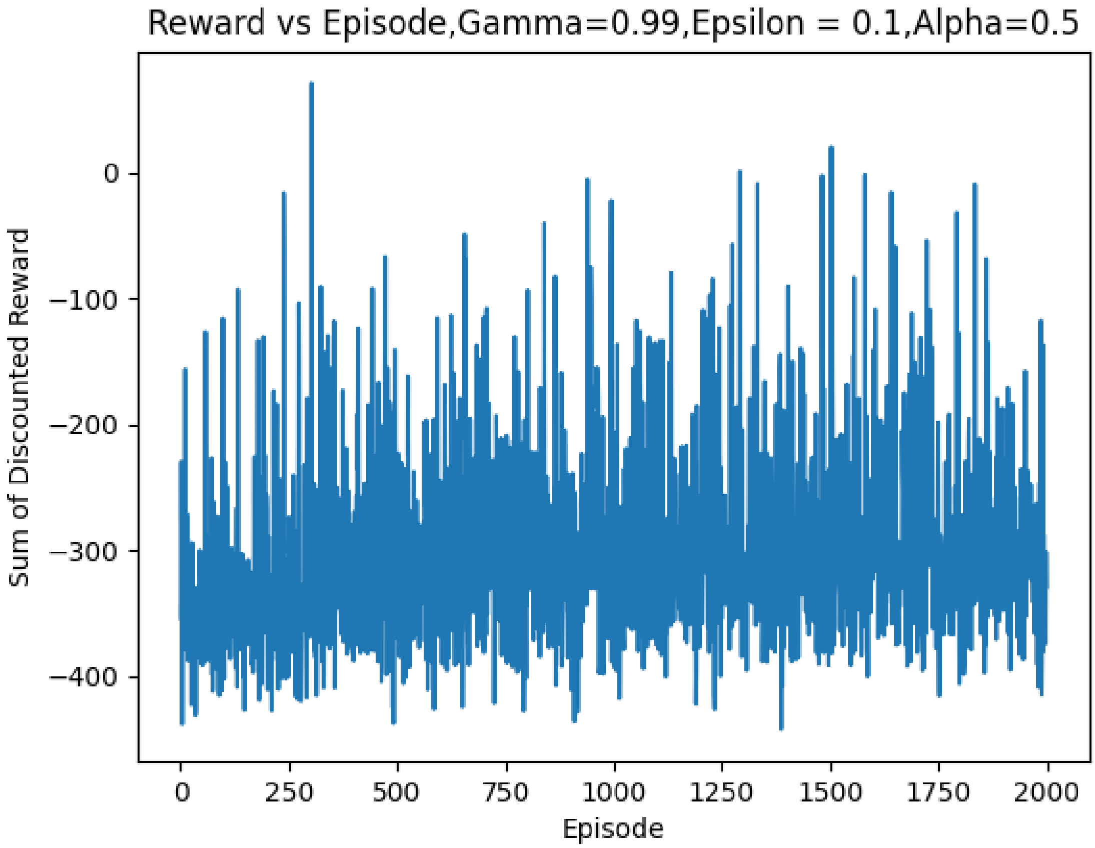
Now, from the above graphs we can conclude that on changing the value of epsilon (exploration rate) as we increase it the cumulative rewards we are getting is also increasing. This is because we will get more commulative rewards as the learning rate is increased.


Reward vs Episode,Gamma=0.99,Epsilon = 0.1,Alpha=0.1
Sum of Discounted Reward
0
−100
−200
−300
−400
Episode
0
250
500
750
1000
1250
1500
1750
2000


Reward vs Episode,Gamma=0.99,Epsilon = 0.1,Alpha=0.2
Sum of Discounted Reward
0
−100
−200
−300
−400
0
250
500
750
1000
1250
1500
1750
2000
Episode

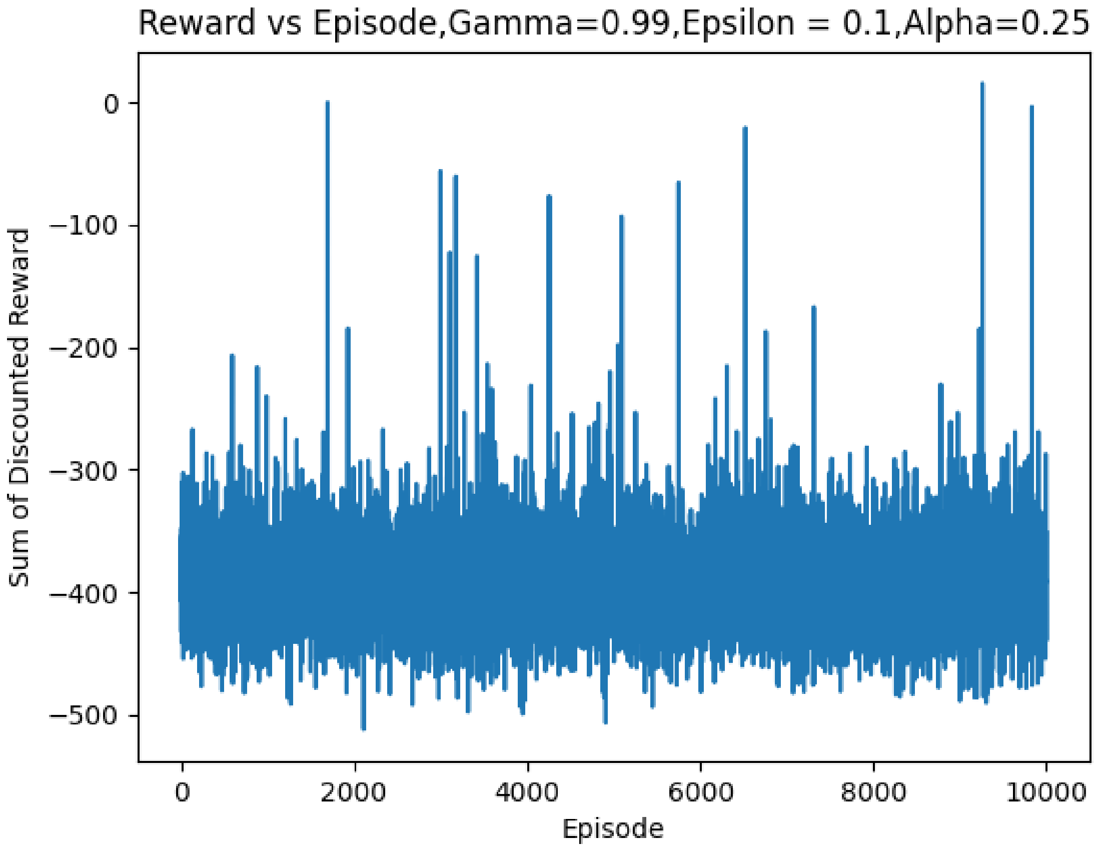

Reward vs Episode,Gamma=0.99,Epsilon = 0.1,Alpha=0.3
Sum of Discounted Reward
0
−100
−200
−300
−400
−500
Episode
0
250
500
750
1000
1250
1500
1750
2000


Reward vs Episode,Gamma=0.99,Epsilon = 0.1,Alpha=0.4
Sum of Discounted Reward
0
−100
−200
−300
−400
0
250
500
750
1000
1250
1500
1750
2000
Episode



On changing the value of alpha we can see that there is almost no change in the convergence of the rewards. This is because of the fact that alpha is the learning rate and the model is learning slightly but not as much to converge.

QUES 5->



On changing the domain to 10*10 problem we see that there is a large increase in the state space of the problem and because of that the rewards accumulated will be highly negative as compared to previous one. Then we will take the average rewards on chosen 5 instances of this same problem.