CSE 574: Introduction to Machine Learning, Sections C&D
Spring 2022

Instructor: Alina Vereshchaka

Assignment 4

Defining and Solving Reinforcement Learning Task

"We certify that the code and data in this assignment were generated independently, using only the tools and resources defined in the course and that we did not receive any external help, coaching or contributions during the production of this work"

PART 1

1. **Describe the environment that you defined. Provide a set of actions, states, rewards, main objective, etc.**

ANS:

a) In our defined environment, we have following set of actions:

1) U(up)
2) D(down)
3) L(left)
4) R(right)

All the address or co-ordinates of the 16 states are mentioned below

S1 ==	(0, 0)	Origin point of Agent
S2 ==	(0, 1)	
S3 ==	(0, 2)	Reward -1
S4 ==	(0, 3)	Penalty - 2
S5 ==	(1, 0)	
S6 ==	(1, 1)	
S7 ==	(1, 2)	
S8 ==	(1, 3)	
S9 ==	(2, 0)	
S10 ==	(2, 1)	
S11 ==	(2, 2)	Reward - 2
S12 ==	(2, 3)	
S13 ==	(3, 0)	
S14 ==	(3, 1)	Penalty - 1
S15 ==	(3, 2)	
S16 ==	(3, 3)	Goal position

Rewards: -

In our environment, following are the rewards allocated: -

- + 0.5: For each of the two reward grids individually
- -0.3: For each of the two penalty grids individually
- + 1: For the final goal grid position.

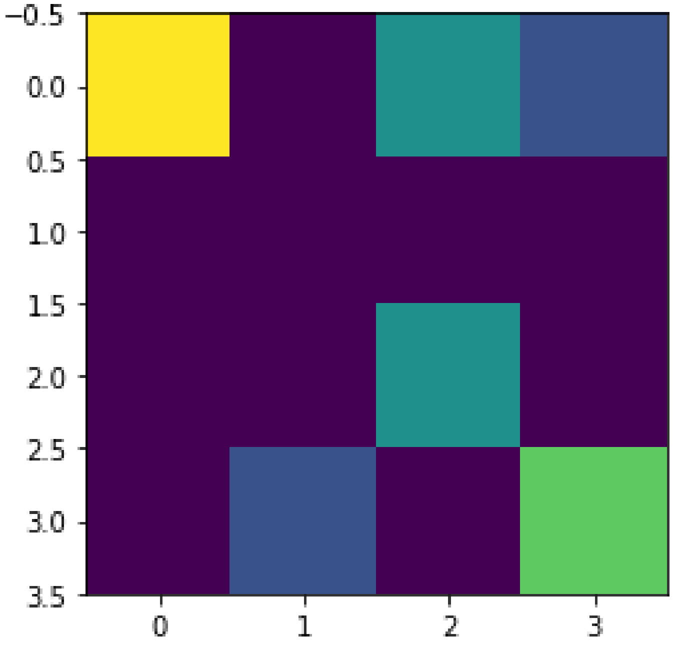
Objective: -

The objective of this designed environment is to start from agent's origin position and reach to the goal position by collecting the maximum rewards and minimizing the penalties.

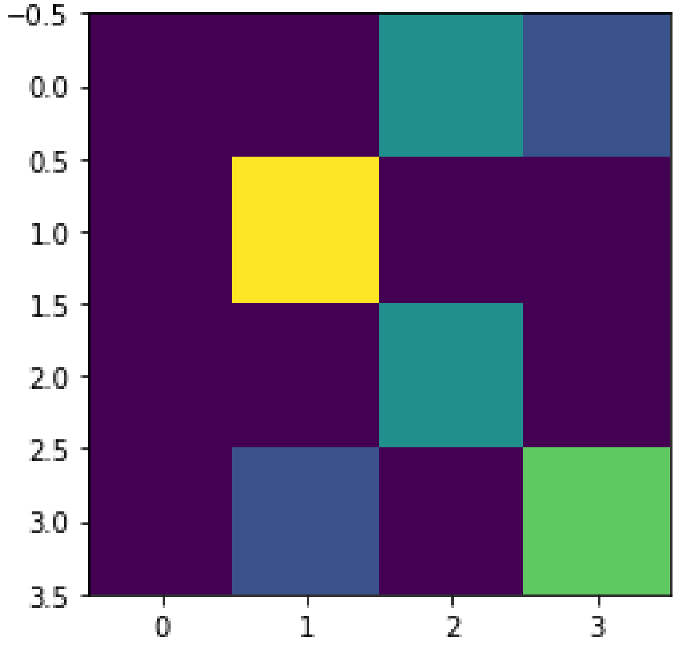
2. Provide visualization of your environment.

ANS:

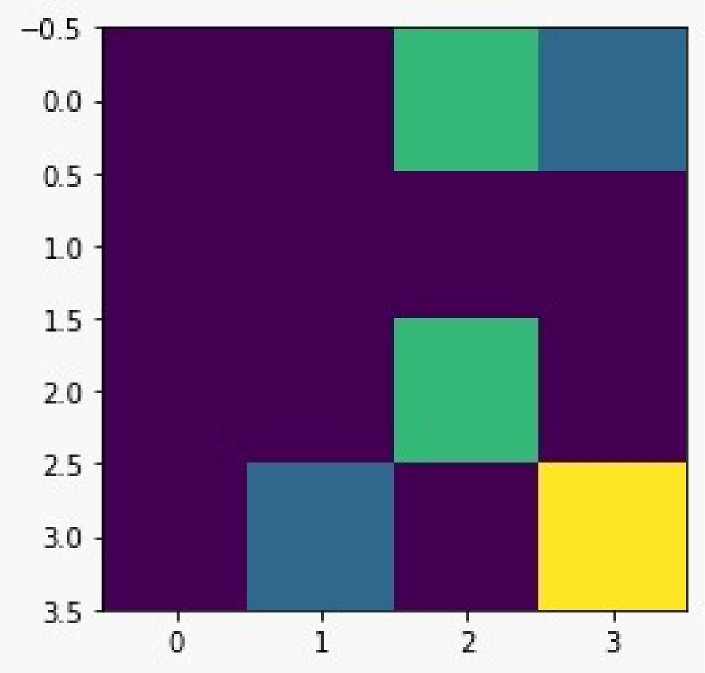
1. In this fig., agent is at (0,0) position which is origin position.



2. In this fig., agent is at (1,1) after a few time-steps from start where it is computing rewards and penalties.



3. In this fig., the agent has reached the final goal, collecting maximum rewards and minimum penalties.



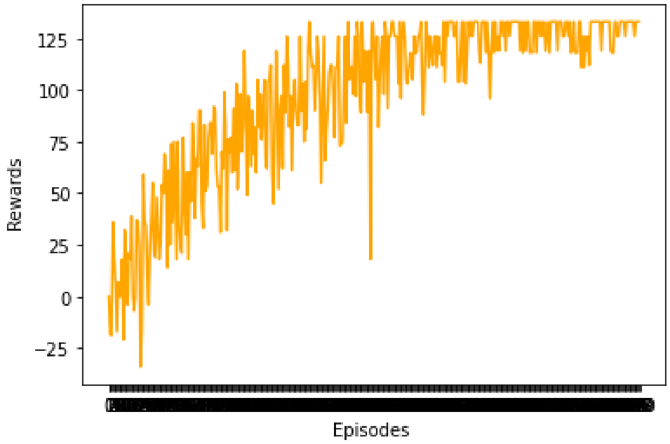
The model will begin from the goal position and travel lot of paths ultimately finding the best path by also utilizing rewards and penalties.

3. **Safety in AI: Write a brief review explaining how you ensure the safety of your environment.**

ANS: In reinforcement learning there is no predefined model. The model takes actions and learn in course of time as it takes actions. Here we have 4 actions up, down, and left, right. While taking these actions it tries to find the goal state, but it will have to collect maximum rewards and minimum penalties. In short, we are optimizing the model. We have given the model some boundaries which keep the check on the actions taken by the model as it will not go outside the grid etc.

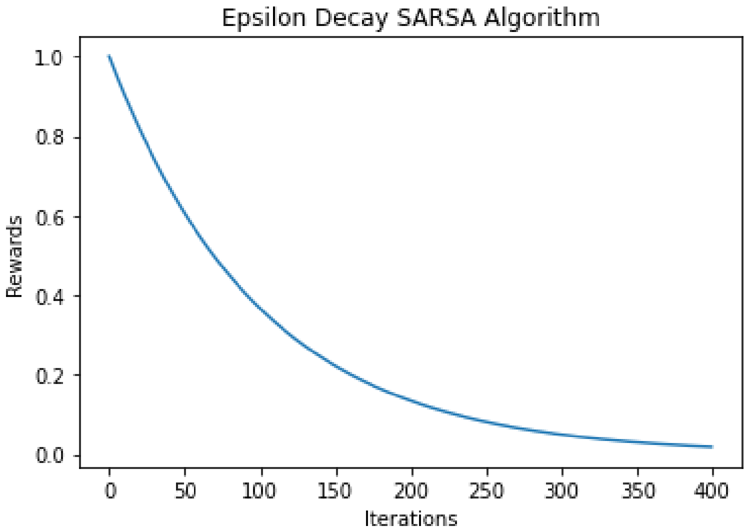
Part 2

1) Show and discuss the results after applying SARSA to solve the environment defined in Part I.

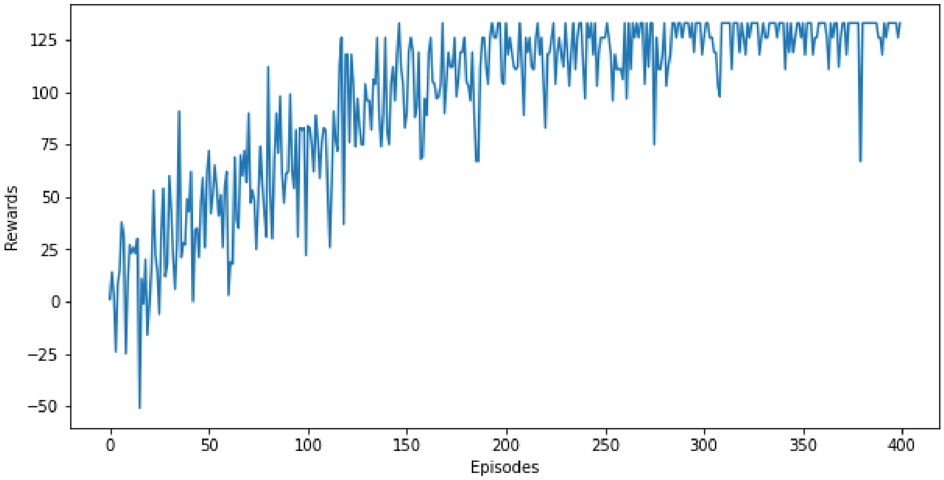


After applying SARSA and using the hyper parameters as alpha = 0.2, gamma = 0.8, episodes = 400 and decay factor as 0.990 we observed that the final goal epochs to reach final point collecting rewards decreased.

2) Provide a plot for epsilon decay

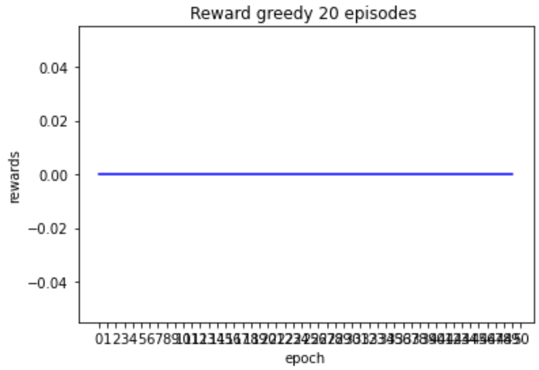


3) Provide a plot for the total reward per episode.



For the total path achieved the algorithm was learning on its own and we can see as the episodes were incrementing the algorithm trained itself to collect max rewards and reach goal position soon.

4) Provide the evaluation results. Run your environment for at least 10 episodes, where the agent chooses only greedy actions from the learnt policy. Plot should include the total reward per episode.



Here although we set for being greedy it did not collect any reward as well. May be for later stage for higher episode it will collect the rewards.

5) Give your interpretation of the results.

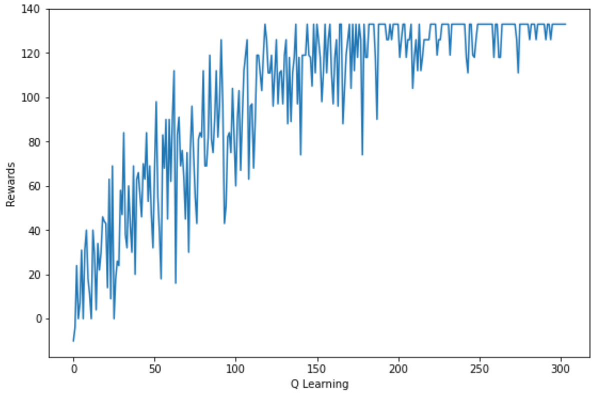
The agent while going from start position to end position follows path of up, down, left, right. While collecting the rewards it will not reset but while taking the penalties it does reset. It hence after going from start to end goal it learns by itself the maximum rewards points and tries to figure out the position to collect maximum rewards and reach the end goal. The value of rewards and penalties must be proportionate but does not matter if low or high.

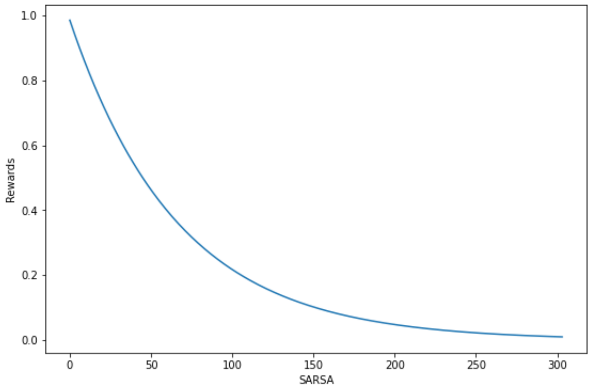
6) Briefly explain these tabular methods: SARSA and Q-learning. Provide their update functions and key features.

Q in Q learning quality and it is also known as off policy reinforcement learning algorithm. It does not consider the current steps while performing the next step. That is it is independent of current state but takes the action from the state where value of Q is maximum. SARSA is also called on policy RL algorithm. While performing the action or next step it considers the S value from the current step to next step (s-).

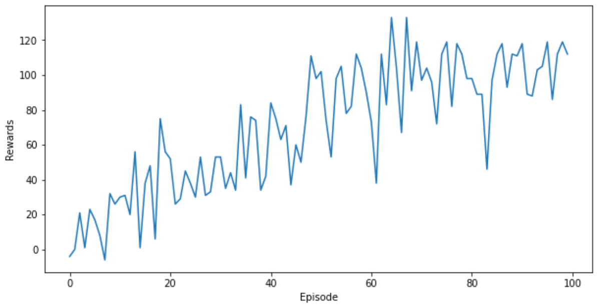
Q learning will not pay attention at which policy is under consideration instead of that it will use the maximum Q value. On the other hand, SARSA updates Q value strictly on the basis of the previously executed policy.

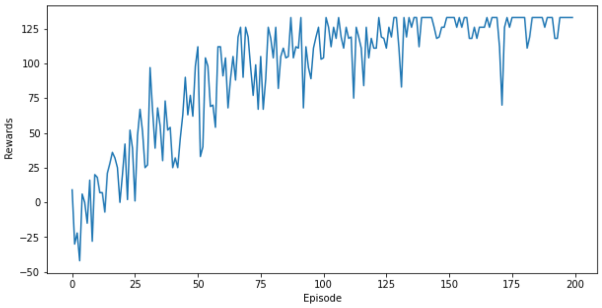
To estimate the future reward Q learning uses the max operation while SARA does not.





7) Provide the analysis after tuning at least two hyperparameters from the list above.



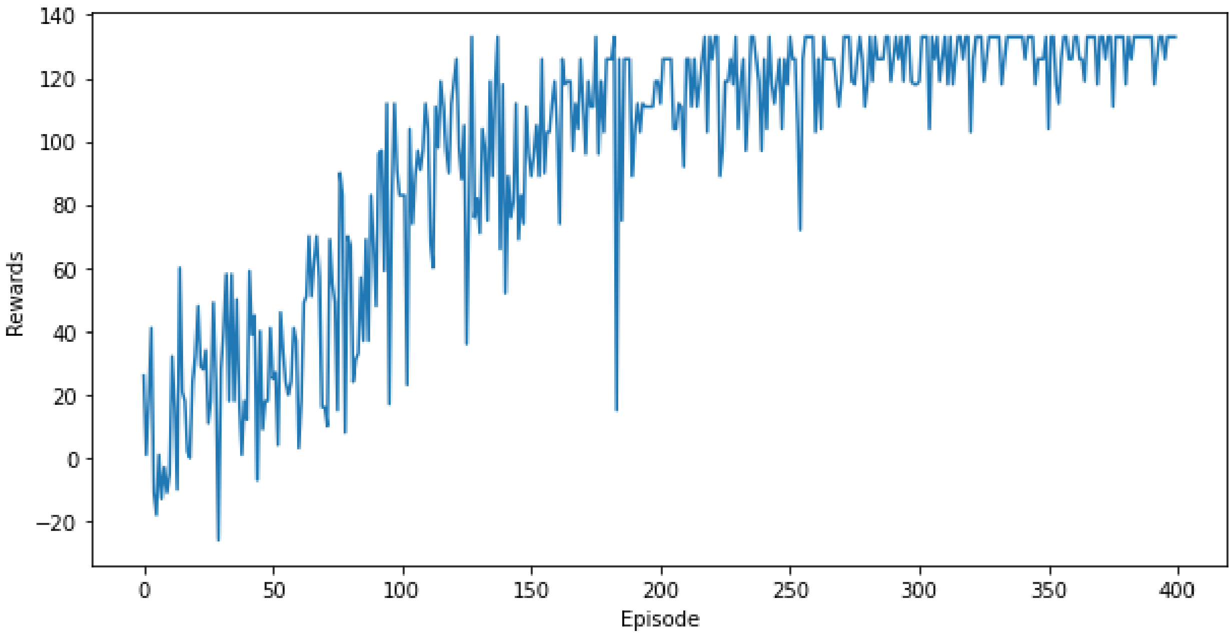


Ep = 100 Ep = 200

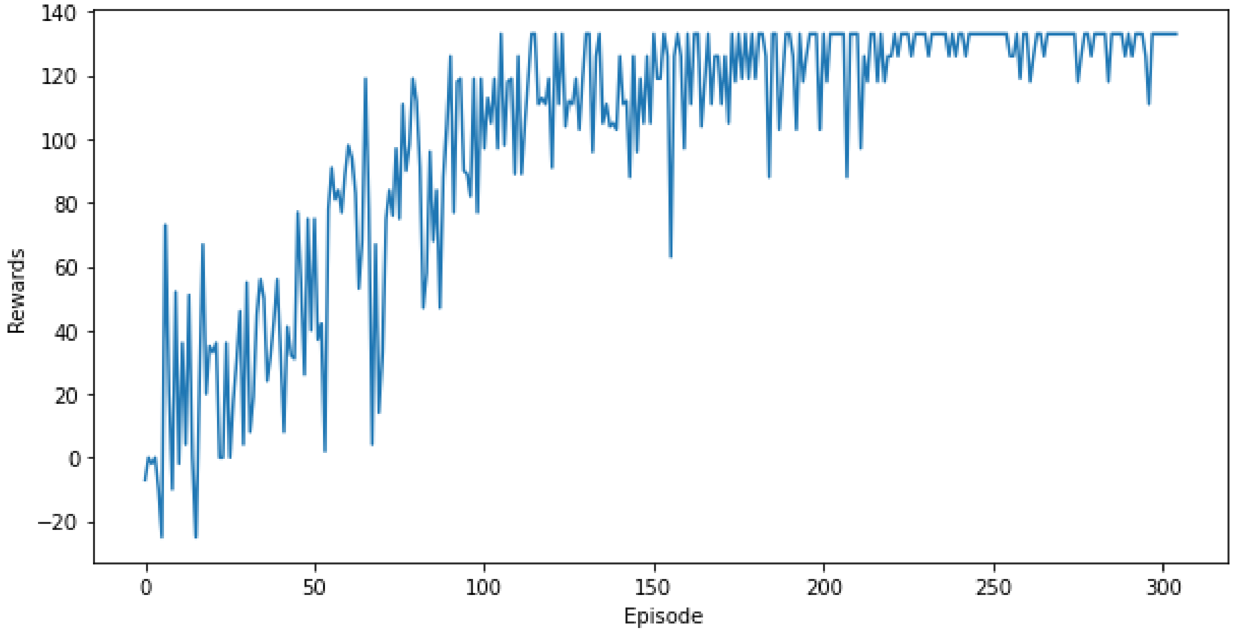
We can see after changing the episodes there comes one episode from where it learns with max rewards and min penalties and is almost constant in collecting rewards and penalties once it is learned. Hence after around 140 episodes we notice almost flat or linear output.

8) **Try at least 3 different values for each of the parameters that you choose. Provide the reward graphs and your explanation for each of the results. In total you should have at least 6 graphs and your explanations. Make your suggestion on the most efficient hyperparameters values for your problem setup.**

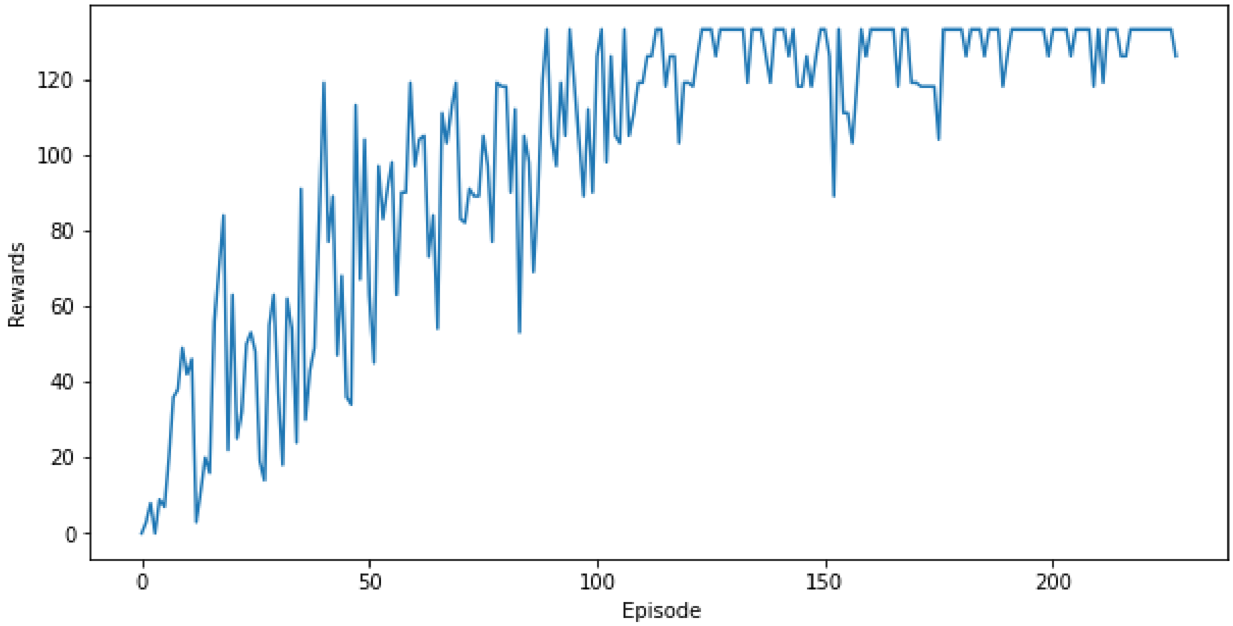
Delay Factor = 0.990



Delay Factor = 0.985

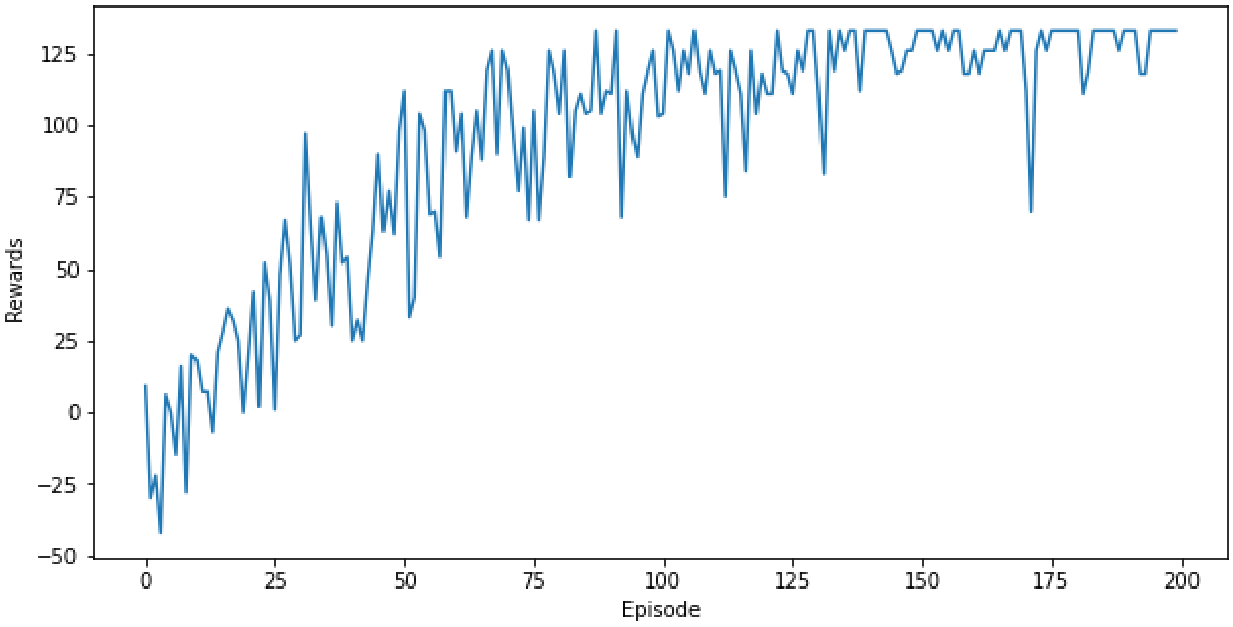


Delay Factor = 0.980

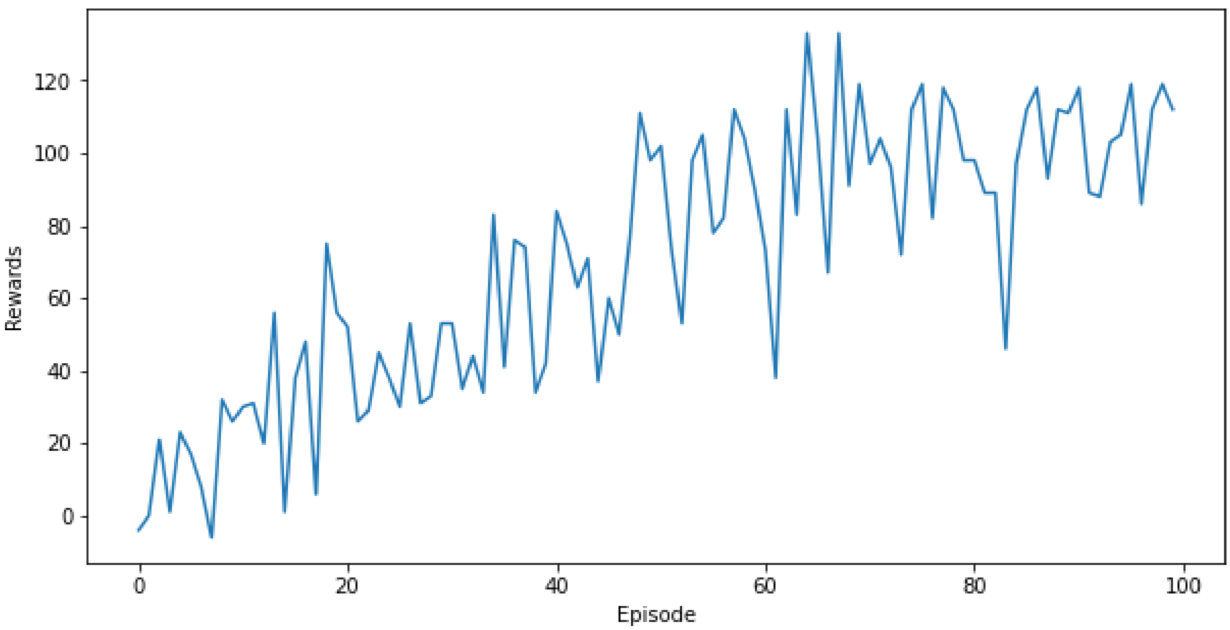


Episodes = 300

Episodes = 200



Episodes = 100


Rewards
120
100
80
60
40
20
0
0
20
40
60
80
100
Episode

Team Member	Assignment Part	Contribution
Harshvardhan Tanpure	1, 2	50,50(%)
Chaitanya Desai	1, 2	50,50(%)

REFERENCES:

1.Geeks for geeks for graphs

2. https://www.geeksforgeeks.org/sarsa-reinforcement-learning/

3. https://en.wikipedia.org/wiki/State%E2%80%93action%E2%80%93reward%E2%80%93state%E2%80%93actio

4. https://towardsdatascience.com/simple-reinforcement-learning-q-learning-fcddc4b6fe56

5. https://www.simplilearn.com/tutorials/machine-learning-tutorial/what-is-q-learning

6. Instructor’s notes/slides/references codes