INDIAN INSTITUTE OF TECHNOLOGY,
MADRAS

REINFORCEMENT LEARNING PROGRAMMING
ASSIGNMENT 3

CS6700

# HRL and DQN

*Author*
DODIYA KEVAL

*Roll Number*
CS19M023

# 1 Problem 1

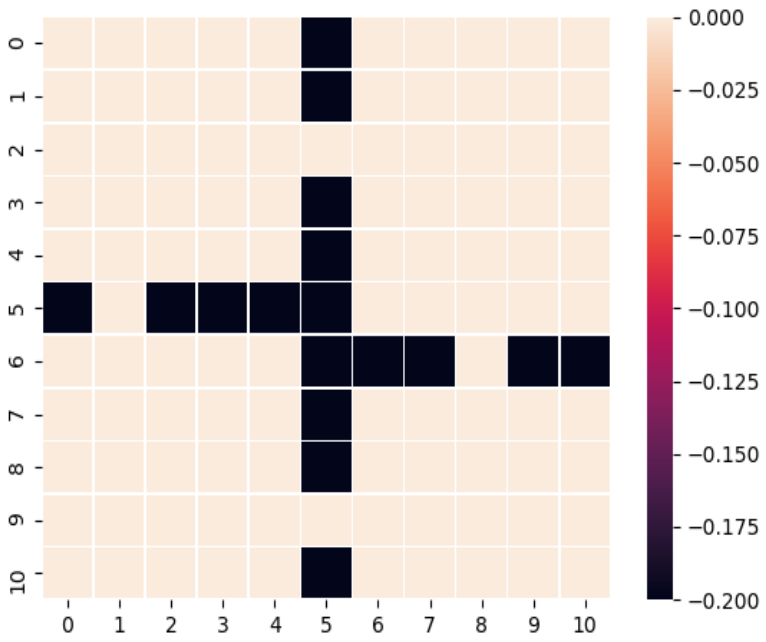



Figure 1: Heat Map of GridWorld without Boarders,Gateways

I made a 11*11 grid world(removed boarders) as above figure.

## 1.1 visualization of Q values

### 1.1.1 Goal 1

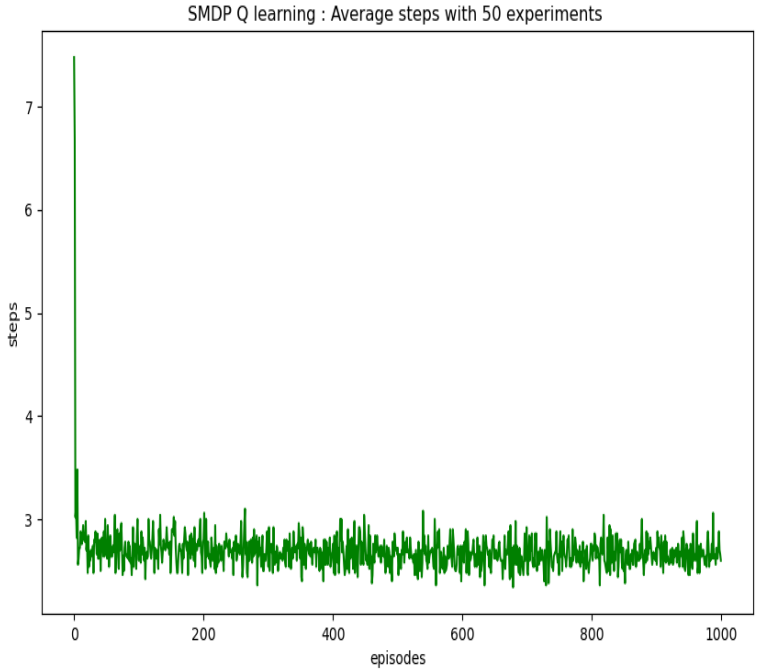



Figure 2: Avg Steps Vs episodes graph over 50 experiments each consists of 1000 episodes for Goal 1 and initial position is random somewhere in room 1.

i took 6 options 4 are primitives and 2 are multi step options. one is clockwise and other is counter clockwise target door selection option.so initially it takes more than 7 steps after learning Q values it'll take avg of 2 or 3 steps(here entire multistep option consider as one step) so multistep option value have more Q value(in Q matrix) obviously hence most of time those options got selected and avg steps are around 2 or 3.

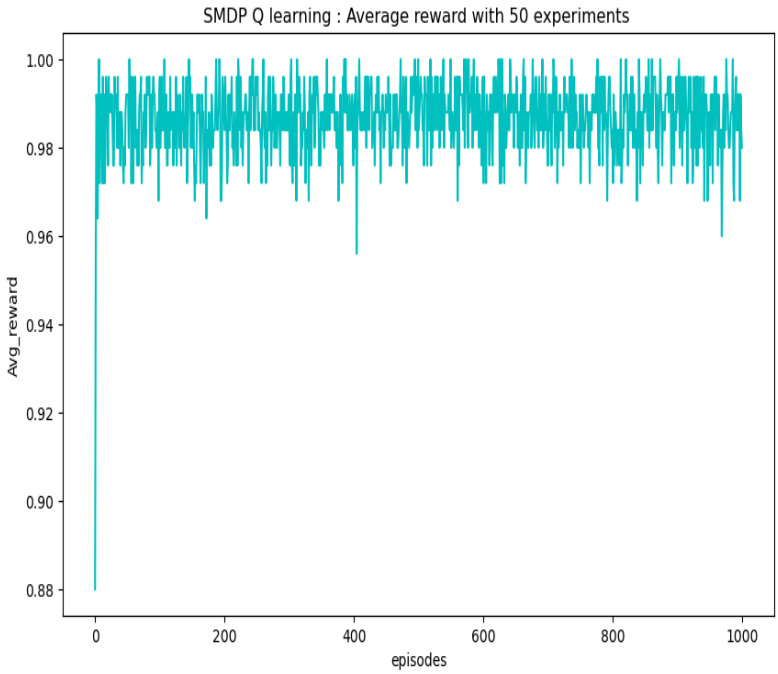



Figure 3: Avg rewards Vs episodes graph over 50 experiments each consists of 1000 episodes for Goal 1 and initial position is random somewhere in room 1

here initially algorithm gives around 0 reward after learning Q values reward will be fluctuate around 0.96 to 1. and heat map is given below.

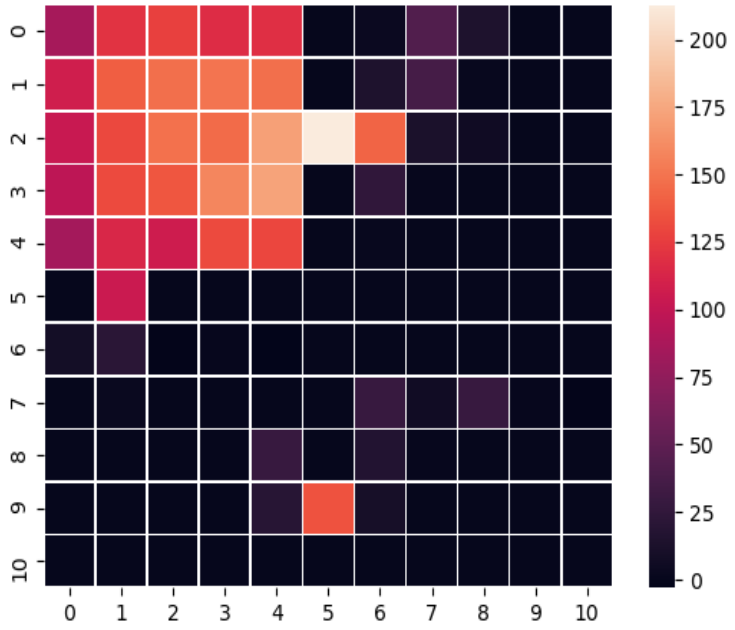
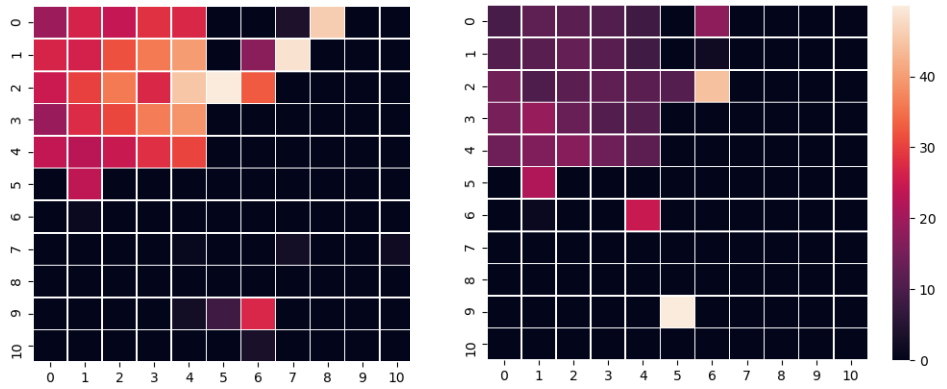



Figure 4: Heatmap over 11*11 grid with sum of all options for each experiments Q values(goal 1).

above heatmap represents avg sum over experiments with sum of all Q values of all options so in room one that values is higher hence that color is light. and again here goal 1 is target goal and initial position starts from anywhere in 1st room.




above figures are visualization of Q values of clockwise(left) and counter-clockwise(right) options. in right image doorway between room 4 and 1 is lighter compare to other doorway indicates Q value of option counter-clockwise explored more. and same explanation we can give for left image also.

### 1.1.2 Goal 2

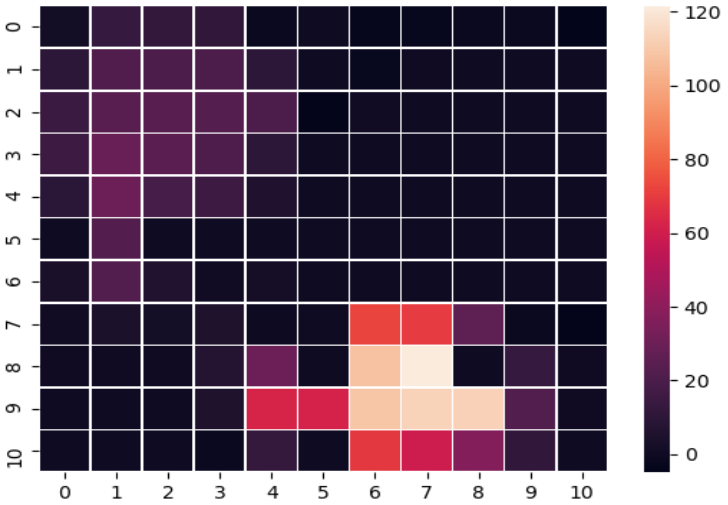
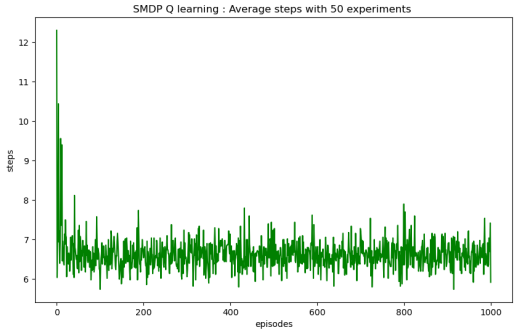


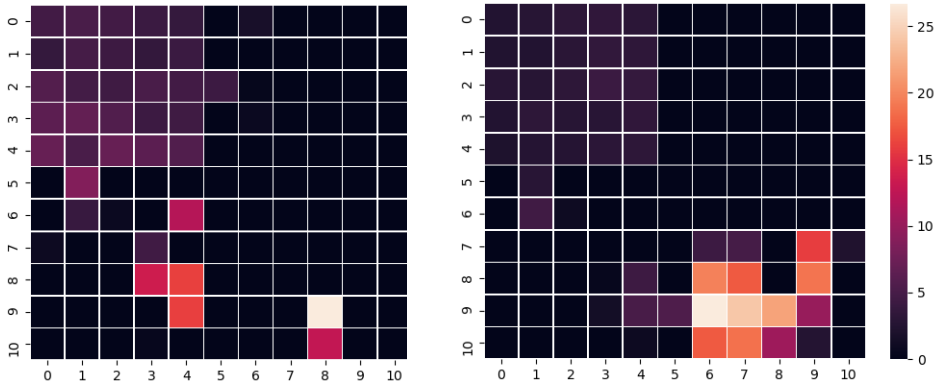

Figure 5: Heatmap over 11*11 grid with sum of all options for each experiments Q values(goal 2).





For goal 2 in heatmap room 4 is lighter and room 1 is little lighter than other rooms. it is because agent stat from room 1 so for room 1's Q values got updated more. and since there are two doorways for goal 2 room 4 vlaues(near goal 2) are very lighter or updated very frequently . and room 2,3 are darker because agent take path from either room 2 or 4 in both cases comparatively Q values less updated. in our case starting position near to doorway $O_1$ more frequently hence doorway $O_2$ were explored more frequently thats why $O_2$ very lighter.

Other 2 graphs are for performance evaluation. as we can see in step Vs episode's graph for initial episodes agent took more steps and then eventually it gets learnt the path and took avg of 5-7 steps.and reward raging around 0.50 to 0.65 thats because sometimes agent took more steps toward boundries and got negative reward so for that reward fluctuation is comparatively more but in the range of 0.50 to 0.65. bellowed figures are Heatmap of clockwise and anti-clockwise Q values.



## 1.2 Problem 2

### 1.2.1 Goal 1 with initial position at centre of room 4

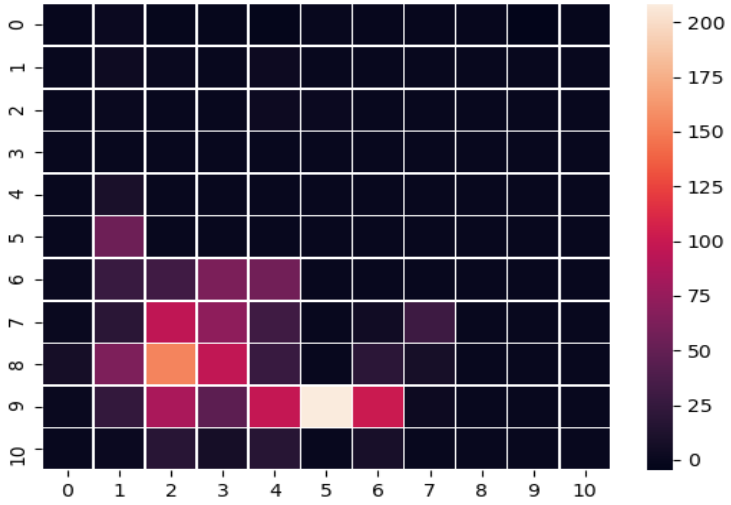
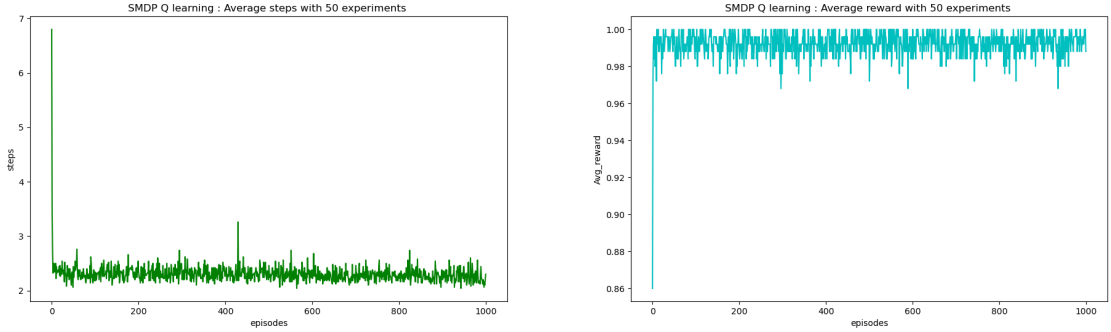



Figure 6: Heatmap over 11*11 grid with sum of all options for each experiments Q values(goal 1 ,initial Position at[8,2]).




As we can see since initial position at [8,2] more likely gateway should be $O_2$ hence that Position in heatmap has very high value. and agent learnt optimal path very quickly(by Step Vs episodes very less steps)and reward also mostly ranging from 0.98 to 1. and almost converge to best path for every experiment. so comapre to initial position at room1 in this situation learning steps are less.

### 1.2.2 Goal 2 with initial position at centre of room 4

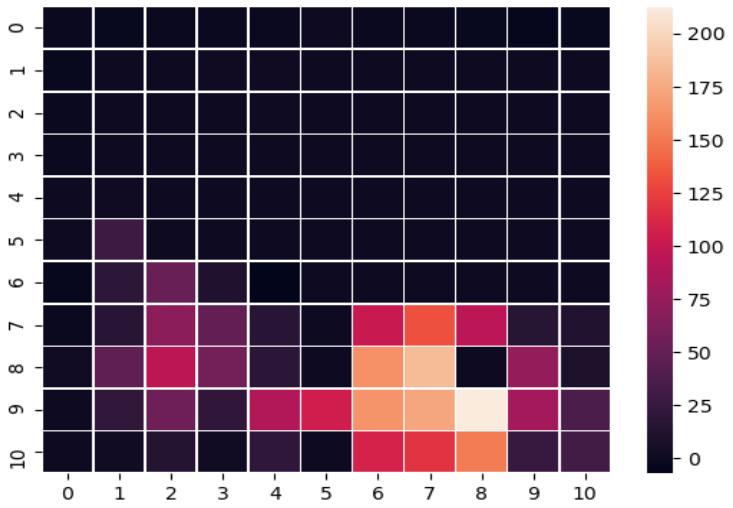
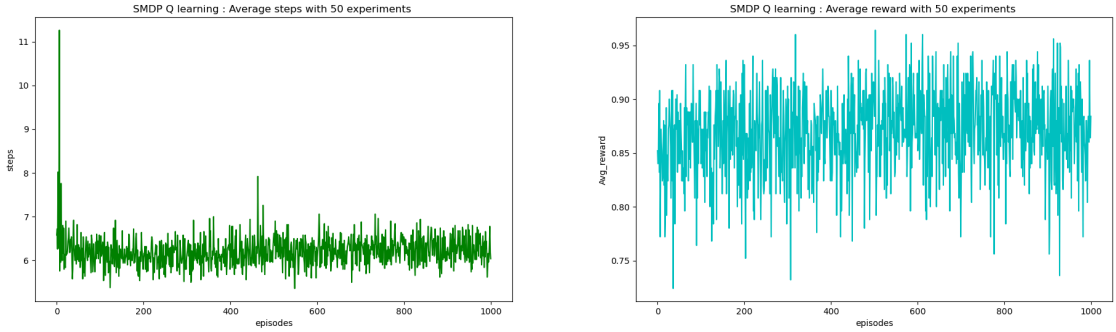
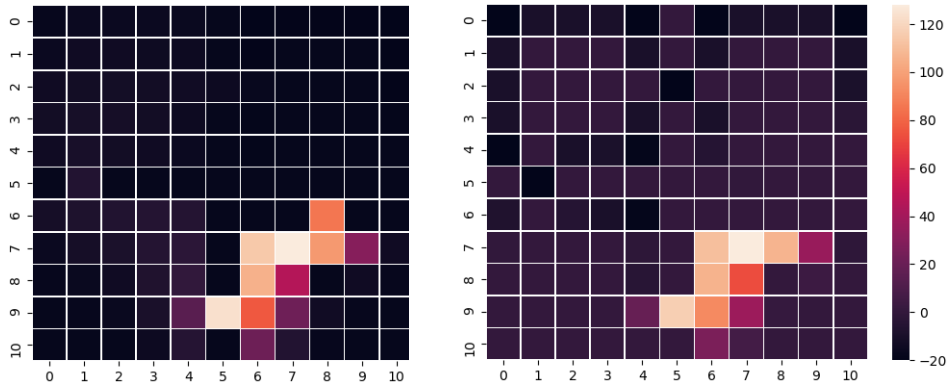



Figure 7: Heatmap over 11*11 grid with sum of all options for each experiments Q values(goal 1 ,initial Position at[8,2]).




As we can see since initial position at [8,2] more likely gateway should be $O_2$ and room 3 gets more explored. and neighbour of Goal 2 are lighter that indicates Q value of each state in room3 perfectly learnt. and avg reward fluctuate 0.75 to 0.95 its because sometimes agent might hit boundaries frequently and for more episodes it can be converge around 0.95.

## 1.3 3 intra-option learning




First figure shows Q value of Goal 1 with intra-option SMDP Q learning and right figure shows Q value of Goal2. as we can see in left figure almost all states of room 1,4,3 has learnt perfectly.by right figure one can says all states has been explored and both multi-options also has been learnt.

So with intra-option SMDP time consumption is more because of options learning but it's eventually find out optimal path for multi-step option. and for room2,3 are little light indicates Q value for every states has computed and hence every state has best policy to go next state unlike normal SMDP in that agent wondering in multi-step options.

# 2 Problem 2

## 2.1

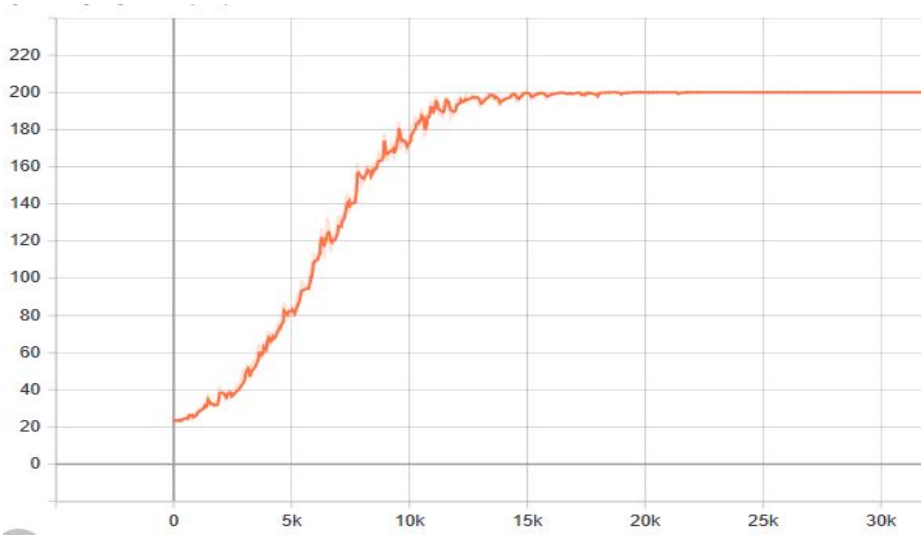



Figure 8: Avg reward over 100 episodes with slide Vs episodes graph over 1 experiments each consists of 30100 episodes.

Above plot shows avg reward over 30100 episodes with sliding window of 100 episodes. as we can see between 15k to 20k episodes start getting reward around 200 and stays same for further episodes. that means agent perfectly learnt the game.

## 2.2 Hyper-Parameters

I choose following hyper parameters.

- Layers : 2 sequential Dense layers(hidden) + 1 input + 1 output layer. each hidden layer having 160 neurons and 'relu' as an activation function.
- Adam Optimizer with Learning Rate(alpha) : 0.01
- Epsilon with Decay : starting with 1(eps) and its value decay with every iteration. decay value is 0.9997 and minimum epsilon is 0.01.
- Batch Size : 32
- $\gamma$(gamma) : 0.95.
- Reply Buffer Size : Maximum = 15000, minimum = 100 experiences.

## 2.3 Affects Of Hyper-Parameters

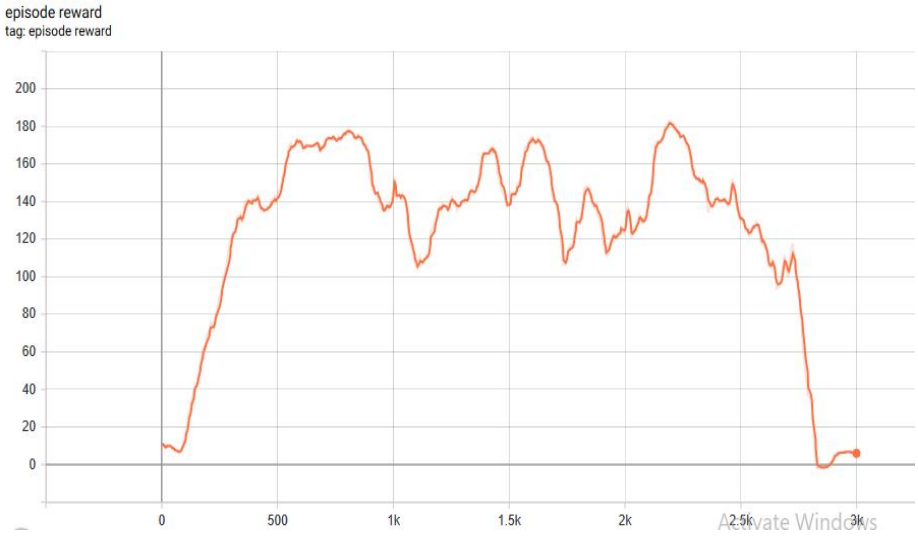



Figure 9: Avg reward over 100 episodes with slide Vs episodes graph over 1 experiments each consists of 3000 episodes with 2 Layers.

Above plots shows Avg reward Vs episode with 2 layer of neural network each with 60 , 60 units and relu is the activation function. so here after some episodes relu neurons gets undershoot(parameters values become negative) or dead hence avg rewards comes down to 15-20 because relu neurons gives output zero with negative value of weights. hence in conclusion i kept 3 layers with 160 neurons each. (here output layer has linear activation function).

I have also tried tanh as activation function but network was not learning after 5k episodes rewards still around 25-30 that is not appropriate.

if we increase decay to 0.9999 then epsilon decrease rate is very slow and took more time to learn hence i kept 0.9997 that is appropriate for exploration and exploitation. and if we don't have enough exploration then reply buffer doesn't have appropriate experience and result in bitter action for given state.

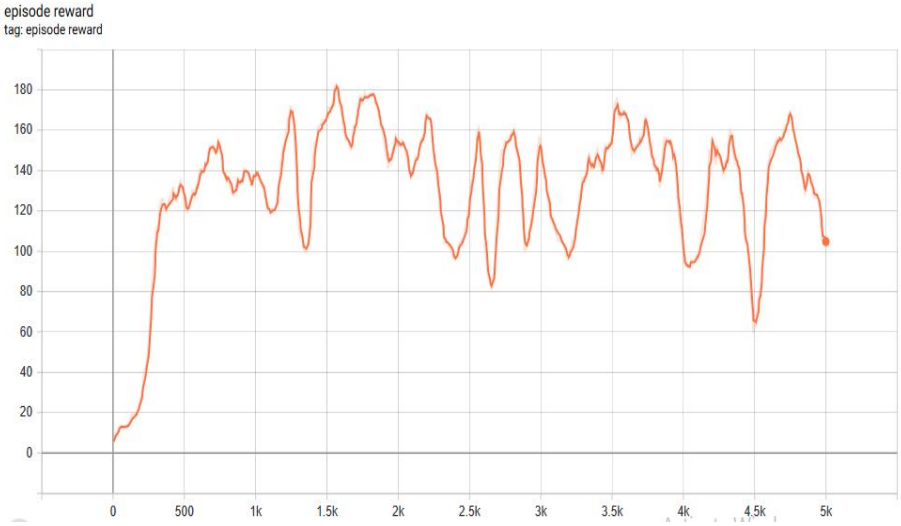



Figure 10: Avg reward over 100 episodes with slide Vs episodes graph over 1 experiments each consists of 5k episodes with batch size 64.

Here by above plot we can see with batch size 64 and neurons 250 ,250(high) model is unstable and learning rate is high compare to our final model with has batch size 32 and neurons 160,160,160 in each layer. if we wait(50k episodes), it might eventually gives perfect output but since batch size is 64 it gets more time to train.
(note: Everytime relu is the activation function).