Assignment 2

All the plots/gif are in plot folder and videos in video folder.

Also uploaded here

All Videos

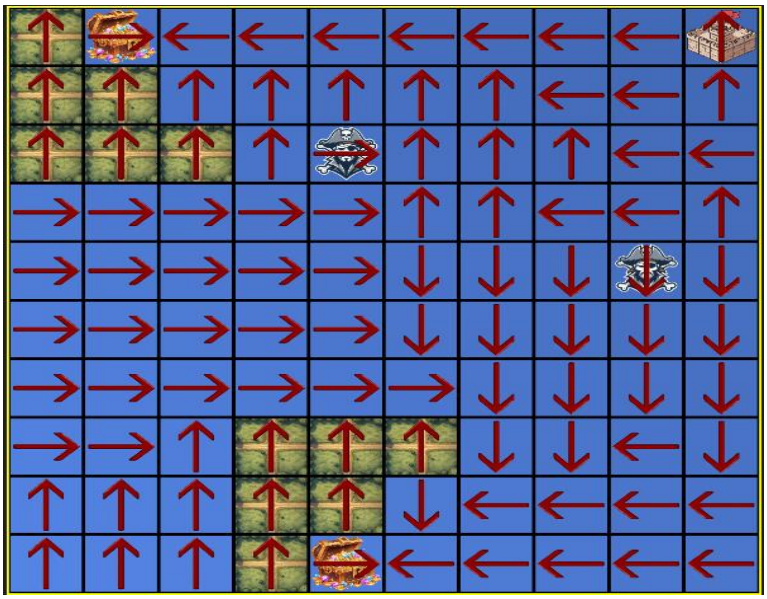
https://csciitd-my.sharepoint.com/:f:/g/personal/aib232073_iitd_ac_in/EuMH3cns2B9HqJUDr-zlXF0B5N14ENWk_91c3I_VWkyLCA?e=JYWEpc

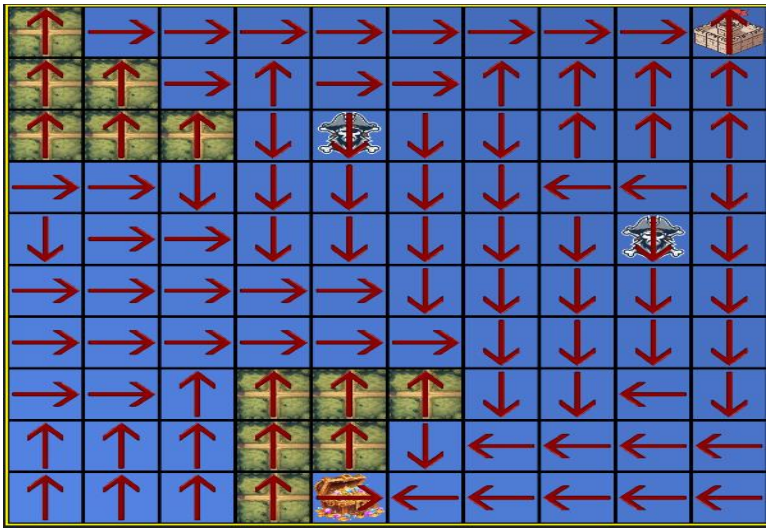
All Plots

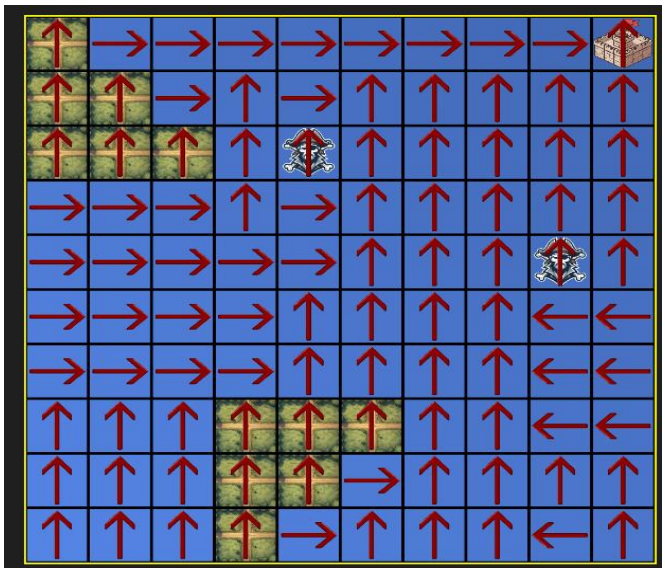
plots.zip

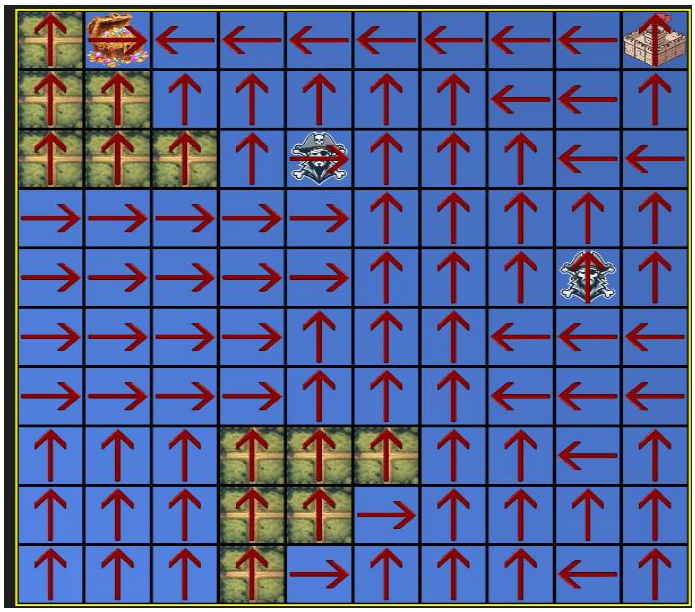
https://csciitd-my.sharepoint.com/:u:/g/personal/aib232073_iitd_ac_in/EZFxBpExUClInDeNYOnPjXYBQwlD2DBuVS3kQMwOelDtAA?e=T31mcO

1.1 Policy Iteration

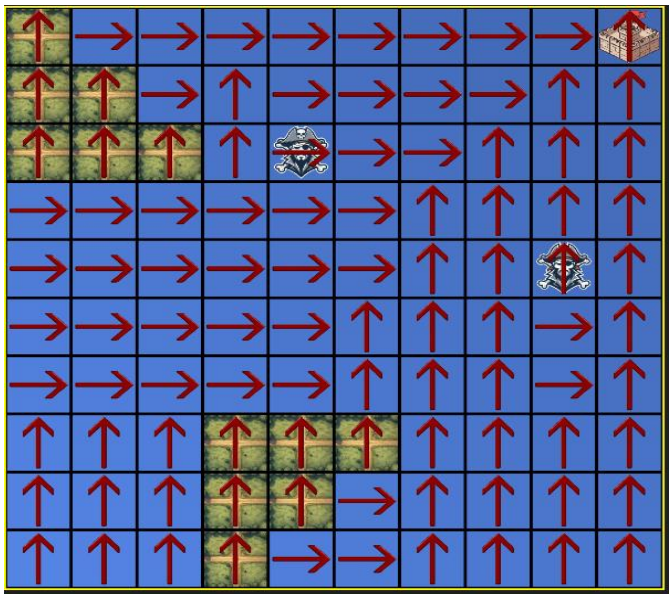


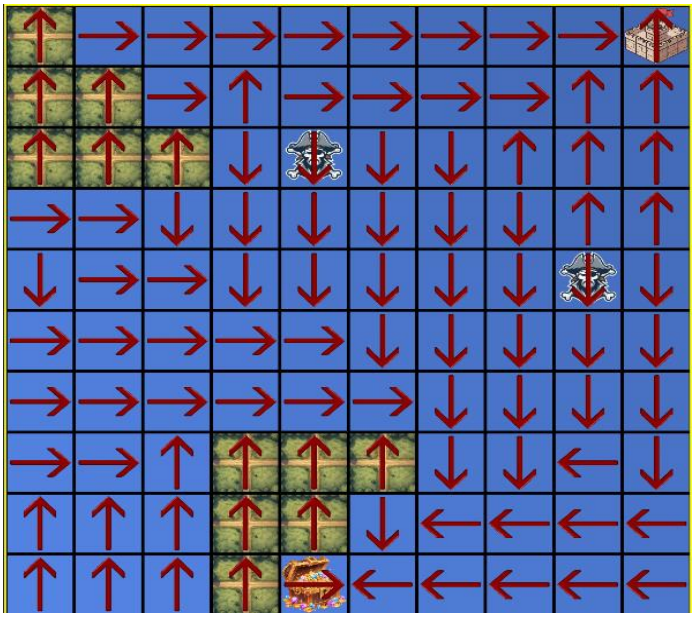


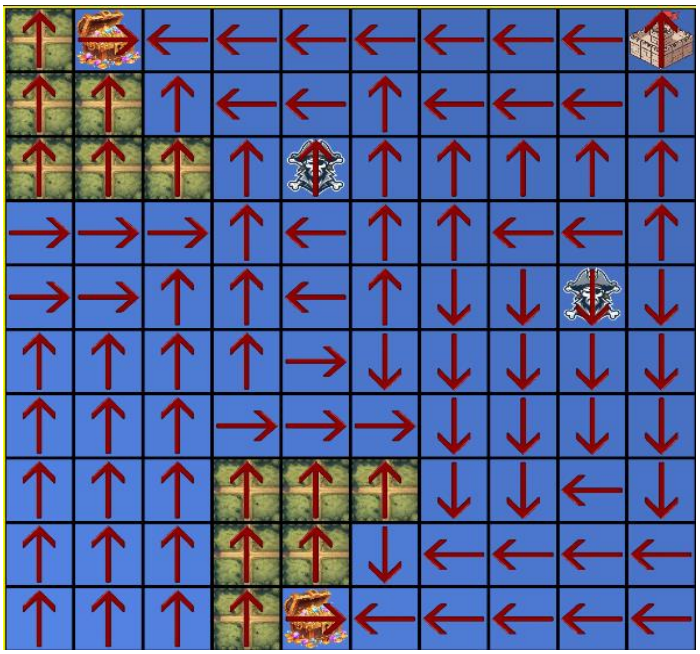




1.2 Value Iteration

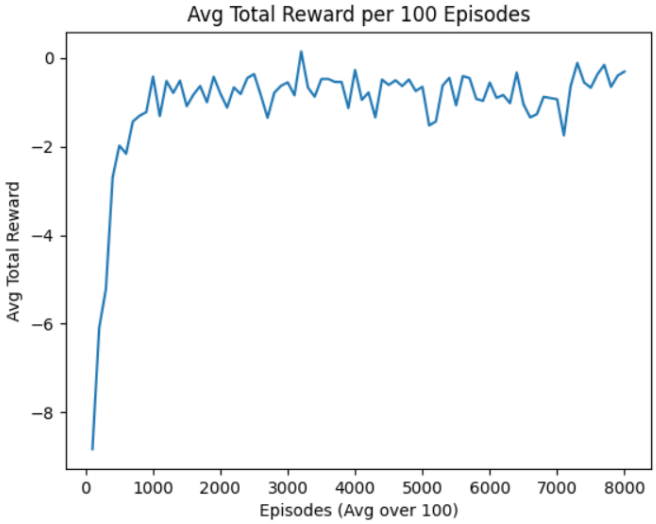






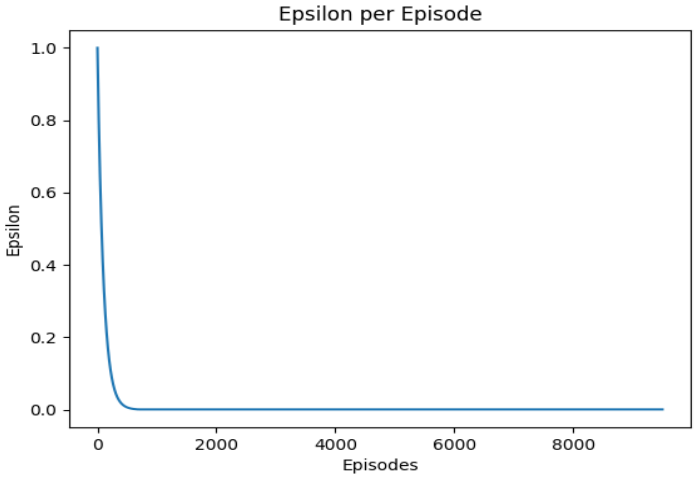
2.1 SARSA Treasure Hunt v1

I have plotted rewards for over 100 episodes.
Convergence criteria was reward over 100 episodes > -0.4



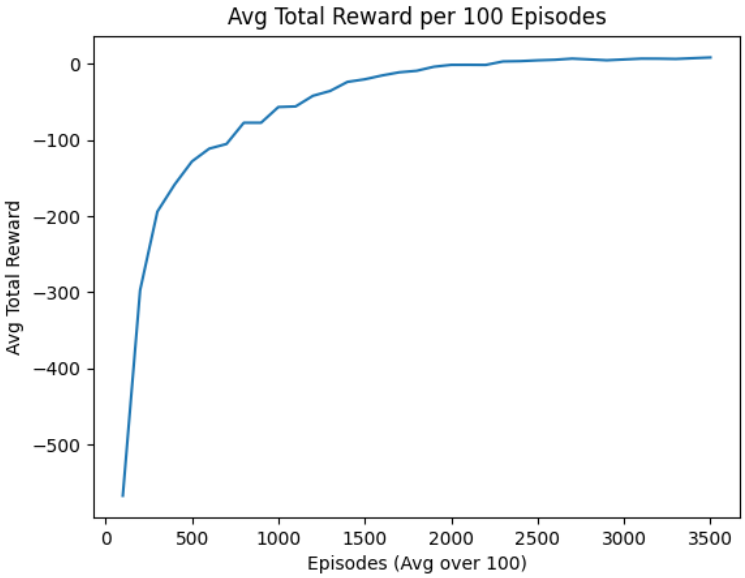
Check logs and plots file for results. Time taken for getting optimal policy was completed in 7.57 seconds. The criteria were rewards of last 100 episodes greater than –0.4 when trajectory size is 100. Reward for ideal trajectory is –0.201 which collects rewards and reaches fort in shortest trajectory. In this environment the agent receives a constant reward of 0.01 on reaching the fort.

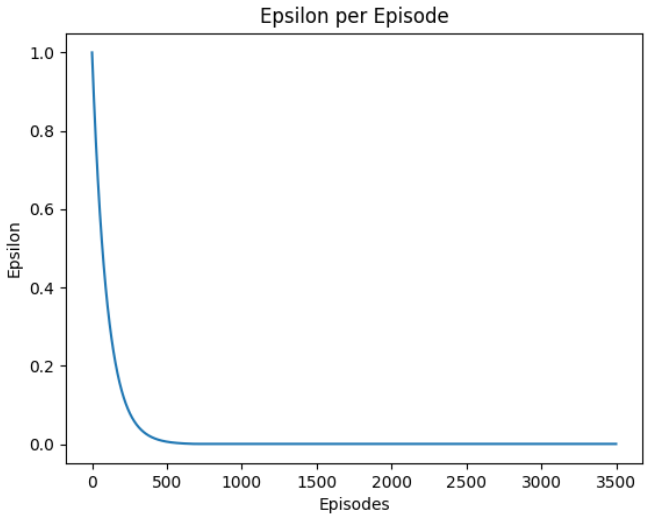
2.2 Q Learn Treasure Hunt v1



Time taken to converge is 28 seconds.

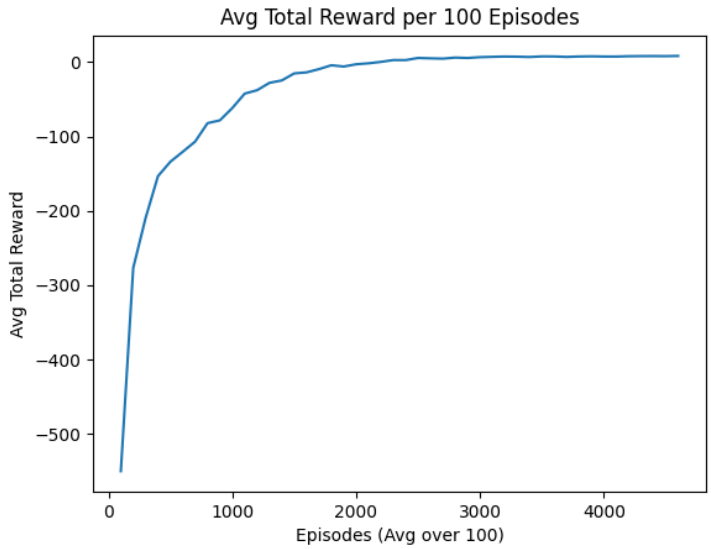
2.2 Sarsa TaxiV3

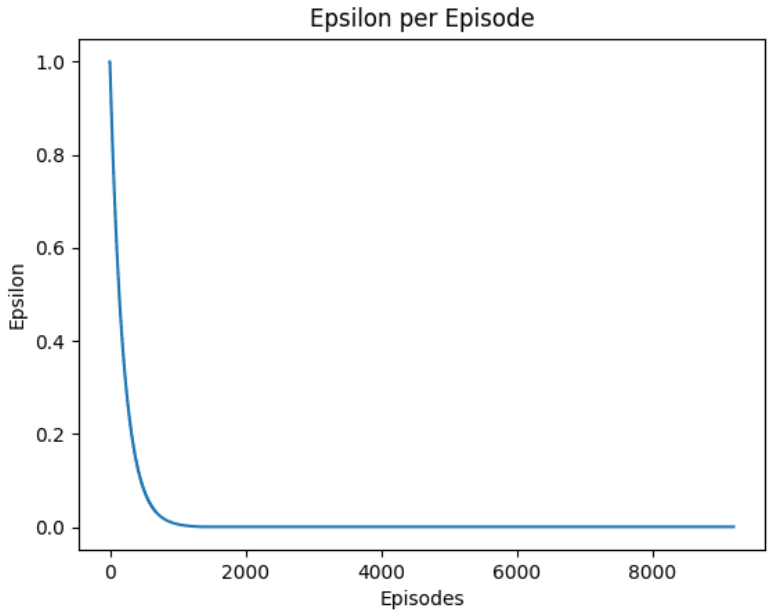




Convergence completed in 5.97 seconds. Criteria avg rewards over 100 episodes more than 8.

2.2 Qlearn Taxi V3



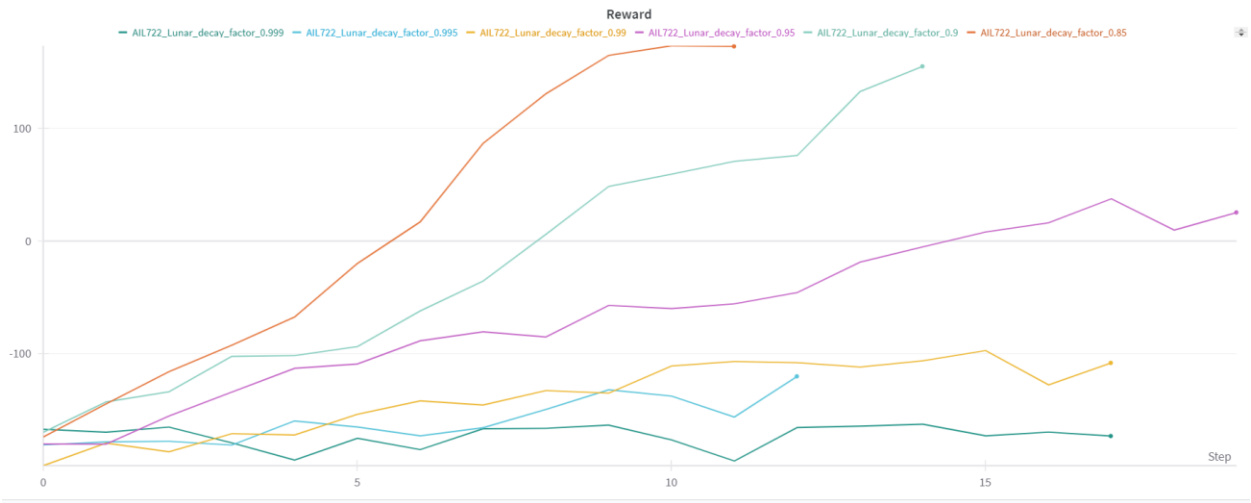


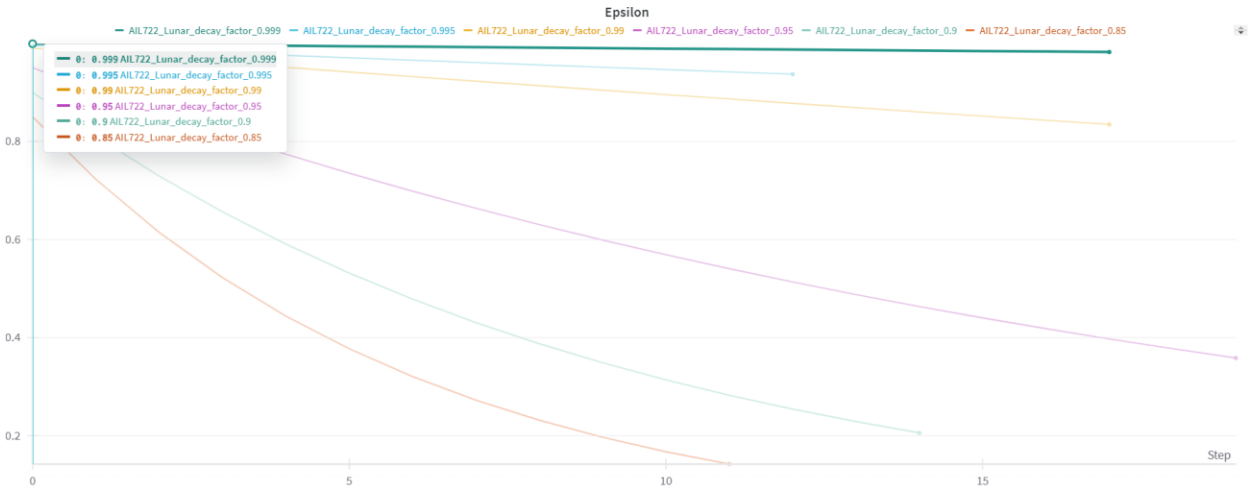
Convergence completed in 7.89 seconds. Convergence criteria avg reward over 100 episodes is 8.

3.1 Lunar Lander

Model for both wind and without wind are saved in saved model folder.

Comparison over different decay factor.





Inference: Model converges faster when decay factor is low. It achieves lower epsilon early , so exploration decreases due to this and it converges early.

Effect of Wind : With wind the average reward is like 128 . Videos are in the video folder.

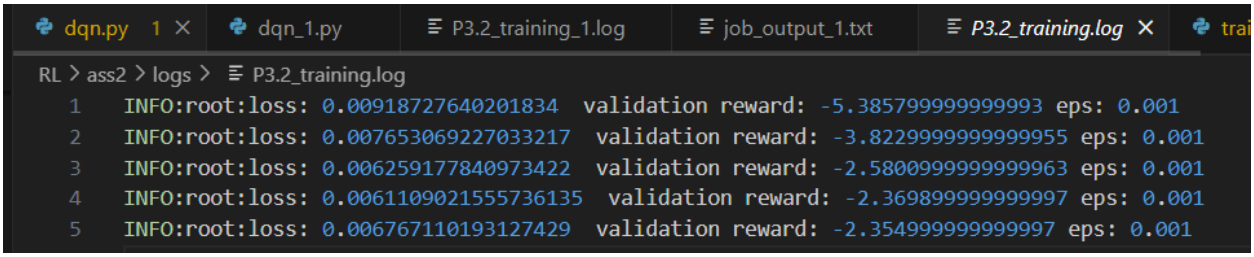
Comparison with random agent: A random agent is provided on webpage

https://gymnasium.farama.org/environments/box2d/lunar_lander/

Compared with the trained model, a random model has very less probability of landing on right spot compared to trained with avg rewards more than 200.

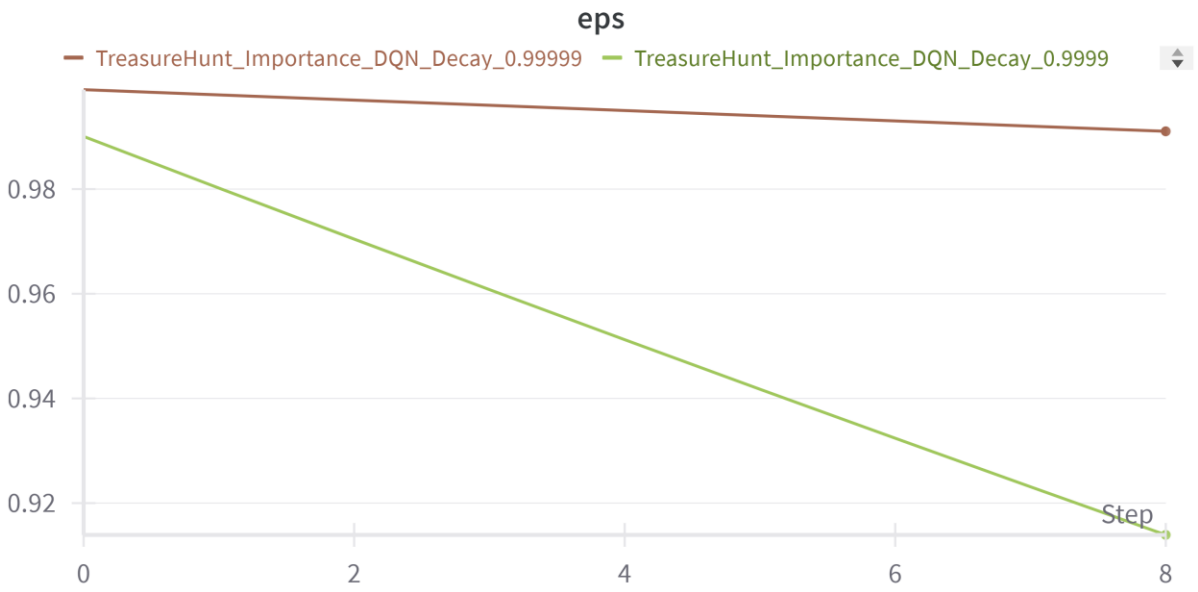
P3.2 Treasure Hunt V2

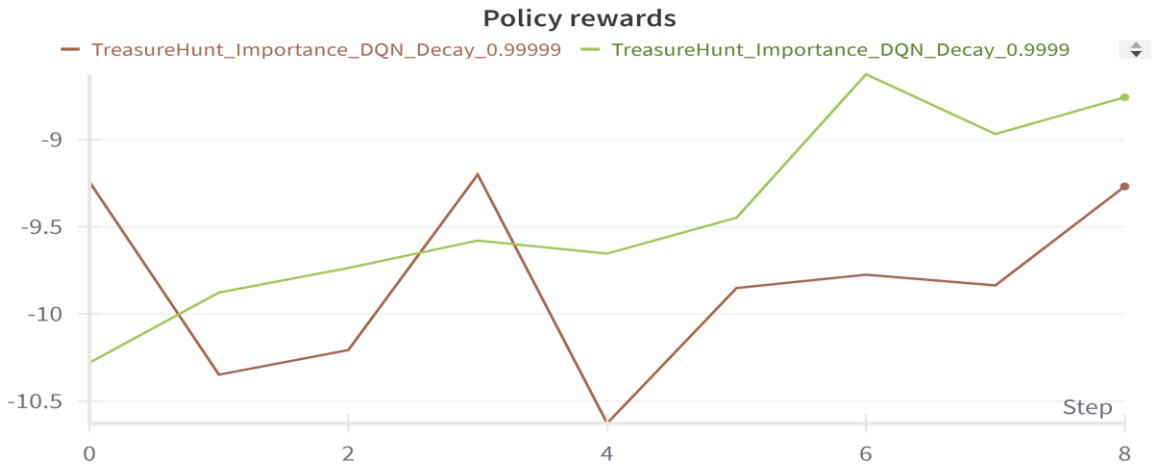
The videos of the implementation are in the videos folder. Most iterations taken were 4000.

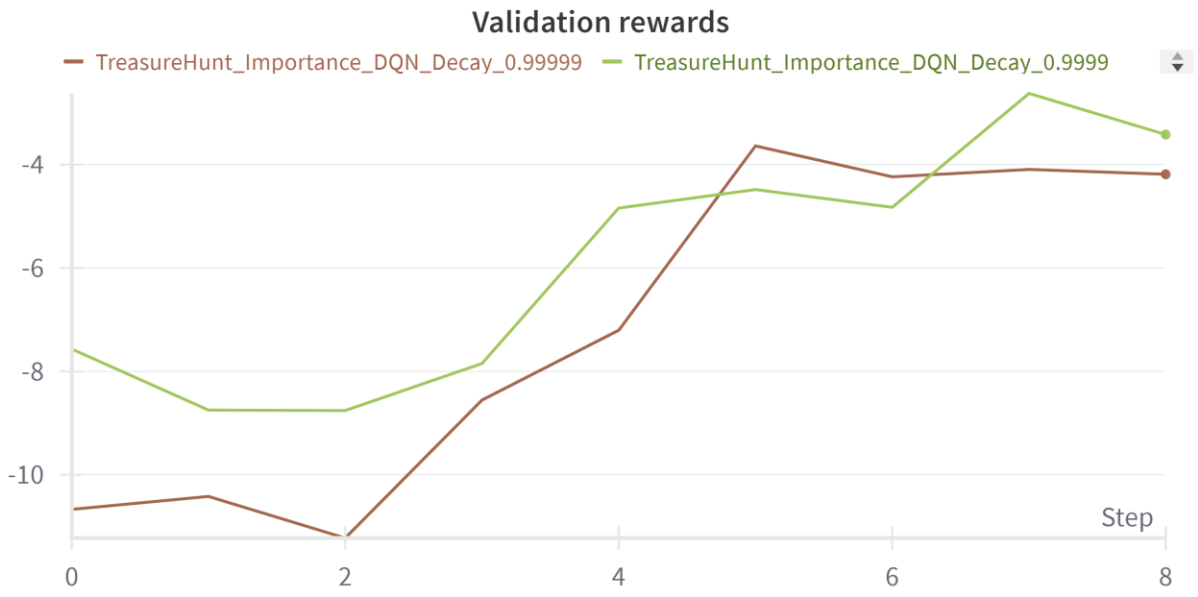


Rewards could reach around –2 with 4000 episodes of training .

Comparison Between decay 0.9999 and 0.99999







Validation rewards are averaged over Demo Environments

Policy Rewards are rewards over last 100 episodes in running DQN.

We can observe from these graphs that 0.9999 decay policy is better than 0.99999 .

Visualization of running policy is in plots folder.