TABLE OF CONTENTS

INTRODUCTION

Reinforcement Learning are a class of problems and its solutions are classed as Reinforcement Learning algorithms. These problems are primarily concerned with a machine or software agent learning how to map situations to actions to maximise a certain numerical reward. This reward signal is known as the reinforcement signal. The aim is for the agent to decide on the best action based on its current state. When this step is repeated this type of problems are called Markov Decision Process (MDP). One of these solutions is the Q Learning algorithm[1][2]. The Q Learning algorithm is one such method of solving reinforcement learning problems. It is our aim to evaluate the effect of parameterisation on the Q Learning algorithm.

PROBLEM DEFINITION

We formulate our problem as that of finding the shortest path in a grid by our agent. To do this we initially define a (6, 6) grid and then increase the scope of the problem to be more challenging by making the grid size larger to (10, 10) and then (100,100). In all cases the task is for the agent to find the shortest route or optimal policy through this. We consider two formulations of the Q Learning algorithm in this study: deterministic and stochastic. Some of the analysis is performed with the deterministic version and some are conducted with the stochastic version. Hence, going forward we will present these two cases separately and assess the effect of different parameterisation on their respective convergence and final results.

TOOLS

We use IPython [3] for this analysis in addition to the NUMPY[4], MATPLOTLIB[5], SEABORN[6] and PANDAS[7] packages. In addition to these the stochastic version of the Q Learning algorithm utilises the MDPTOOLBOX[8] package. It is worth noting that Python uses 0 based indexing.

ANALYSIS ROADMAP

Since we are considering two cases of the Q Learning algorithm we present an overview of the analysis conducted with each implementation.

Q Learning: Deterministic	Q Learning: Stochastic
We consider a small (6, 6) grid from now on referred to as small grid and we conduct the following analysis: **Basic Case:** • For the small grid define a learning rate as a function of the number of episodes • Define a State Transition function and the Reward matrix, R • Define an epsilon greedy policy, epsilon = 0.8 • Define discount factor, gamma = 0.2 • Initialise the Q matrix as zeros and show updates after 1200 iterations • Represent performance by looking at the discrepancy value at each episode, the final V values and comparing number of steps at each episode **Advanced Case 3: Different Learning Rates** • We define a fixed learning rate and compare the results for epsilon greedy policy **Advanced Case 4: Different policy** • Define the learning rate as a function of the number of episodes • Consider the results of an epsilon-greedy and softmax policy with a constant gamma value of 0.2 • Consider the results of an epsilon-greedy and softmax policy with a constant gamma value of 0.8	We consider the larger grid sizes of (10, 10) and (100,100) for this implementation in addition to the (6, 6) case. **Advanced Case 3: Different Learning Rates** • We define learning rate and epsilon as a function of number of episodes. **Advanced Case 2: Different Gamma Values** For the small grid: • We rerun the initial analysis with different discount factor values of 0,0.5 and1 **Advanced Case 5: Different State and Reward functions** A Medium grid world (10, 10) • We change the transition probability matrix and reward matrix by changing the dimensions of the problem from a (6, 6) to a (10, 10) grid. • Here we use the stable parameters from the previous experiments which we deem to be discount factor of 0.5 and vary the maximum number of iterations from 10k to 20k and compare the results.
Extra 1: Expanding the scope of the problem A Big grid world (100,100) • We scale our problem to a 100 by 100 grid and define the P and R matrix to run our experiment. Here we use an agent with discount factors 0 and 0.5 with 50K iterations **Extra 2: Comparison of all the above cases** • We compare the optimal policies, V values, performance values and runtimes for all our stochastic experiments. • We compare performance values, V values and number of steps for all our deterministic experiments.	

THE Q LEARNING ALGORITHM

Q Learning is an off policy temporal difference learning algorithm which in its simplest form can be defined as follows:

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha\left[R_{t+1} + \gamma \max_a Q(S_{t+1}, a) - Q(S_t, A_t)\right].$$

Equation 1: Q learning algorithm update rule[1]

$$\mathrm{dQ} = \alpha[R_{t+1} + \gamma max_a \mathrm{Q}(S_{t+1}, \mathrm{a}) - \mathrm{Q}(S_t, A_t)]$$

Equation 2: The update term for the Q learning algorithm

$$\text{discrepancy} = \text{abs}(dQ)$$

Equation 3: This is the performance measure that we use for the Q Learning algorithm.

We note that the discount factor and learning rate are incorporated in the delta and they both act as a damping factor on the values that the Q matrix is updated with. Therefore the relevant inputs to the Q Learning algorithm then become:

1. The Reward Matrix, R
2. The discount parameter, γ
3. The learning rate, α
4. The Q matrix, initialised as zeros

Q LEARNING: THE LEARNING RATE, α

The learning rate α determines how much the new information that the agent is acquiring will change what he already knows. Instead of using a fixed value for the learning rate, we defined α as a function of the number of training episodes. The alpha parameter is defined as follows:

$$\alpha = \frac{1}{(\sqrt{n+2})}; \quad n = \textit{episode number}.$$

Equation 4: The learning rate, alpha is defined as a function of the number of episodes. It is defined as the inverse square root of the number of episodes with a constant added to prevent division by zeros. .

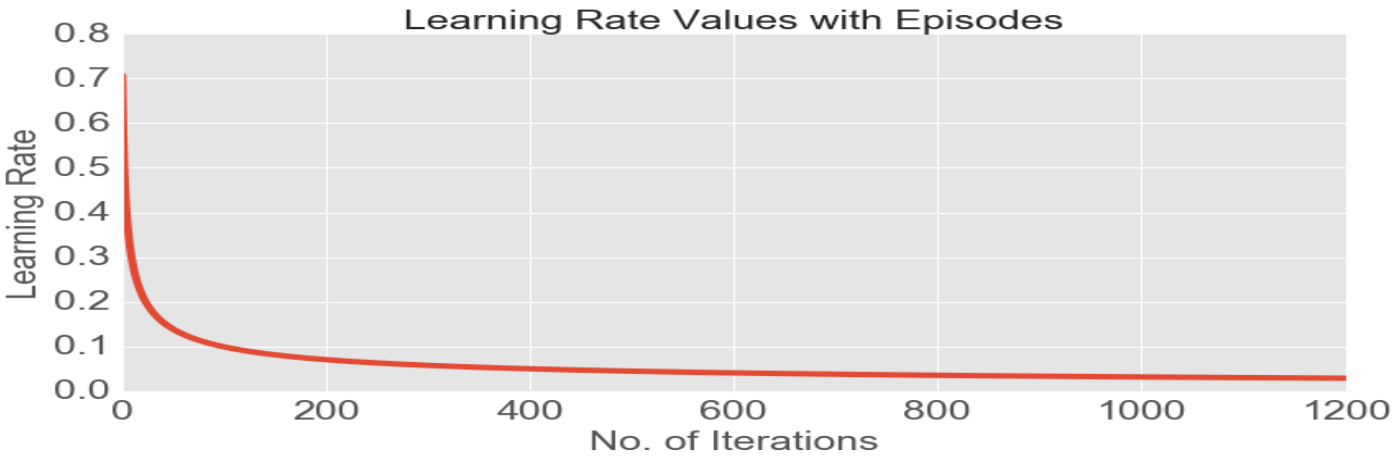

Figure 1: Plot of learning rate, alpha with increasing number of episodes. This is a more dynamic approach to having an adaptive learning rate than a static fixed value.

The impulse response of the learning rate is shown in Fig 1 as defined by Equation 4.This has the effect of applying an exponential smoothing to the Q learning update values. Also the magnitude of its impact is much smaller over time than the gamma value which we keep constant. Since our alpha values change with each episode we say this represents our **Advanced Case 3** and this underlies all our experiments that we present from this point unless explicitly stated otherwise.

Q LEARNING: DETERMINISTIC CASE

In this section we present our analysis with the deterministic implementation of the Q Learning algorithm. Here we define our problem space as a (6, 6) grid. Which gives our agents 6 states and actions to consider. The agent starts at a random position and our goal for him is to reach the state zero. The agent can move only one state at a time. Since we implement an epsilon greedy policy, it compares the epsilon value 0.8 to a number generated from uniform random distribution. When this number is lower than the epsilon the agent is motivated to exploit and in the converse case the agent is motivated to explore. For faster convergence we preferred to bias our agent more towards exploitation than exploration.

Q LEARNING: BASIC CASE

We represent the small grid case and its associated state and reward functions in Fig 1 and Table 1. Fig 1 shows the areas of the grid with high numerical reward with the yellow squares representing the terminal areas. These are the areas that we want the agent to find its way to and the negative reward areas are the ones which we want the agent to avoid. These represent obstacles and in the real world would slow the agent down. But it is more realistic as for example when finding the shortest route one may encounter road works or traffic jams which these could represent. The rewards are given in terms of the states that the agent passed through. The only high positive rewards are given for the terminal state 0, since we want him to stay there as well as for state 2 which is the only point from which the agent is permitted to pass without a penalty.

To do so we initialised a (6, 6) matrix with zeros and assigned to it the values in Table 1. The agent is encouraged to move to the terminal state which has a reward of 100 and avoid the areas marked by negative rewards indicated in yellow.

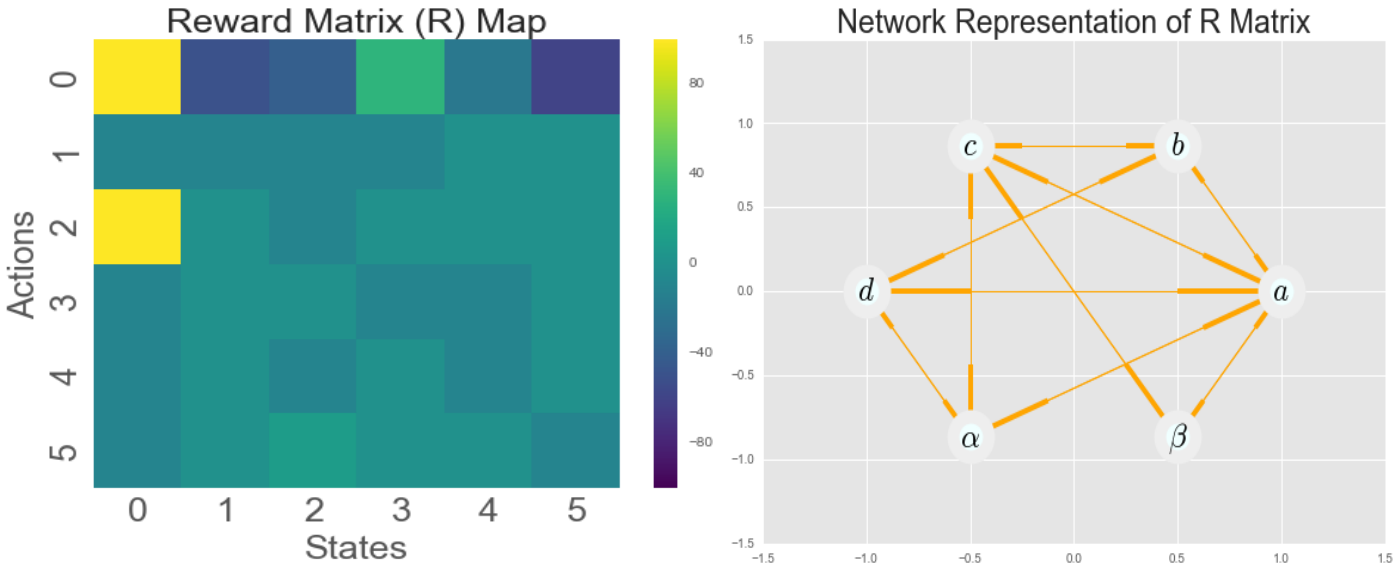

Figure 2: (Left) Map of Reward values showing positive and negative value areas. (Right) Network representation of our R matrix with the terminal state being represented by node 'a'. Also the directed edges shows that the matrix produces a fully connected graph with all states being reachable. The thickened arrows at the incidence of the nodes represent directions.

Table 1: shows our R matrix which contains all the instant rewards that are available to the agent. As it was graphically represented above, our agent can go to any state from any starting position since there are not any states without a link. The rows of Table 2 represent the current state S_t and its columns the next available state S_{t+1}.

	0	1	2	3	4	5
0	100	-50	-40	-30	-20	-60
1	-10	-10	-10	-10	0	0
2	100	0	-10	0	0	0
3	-10	0	0	-10	-10	0
4	-10	0	-10	0	-10	0
5	-10	0	10	0	0	-10

Table 2: State transition function

State, S_t		State, S_{t+1}
0	->	0 or 1 or 2 or 3 or 4 or 5
1	->	0 or 1 or 2 or 3 or 4 or 5
2	->	0 or 1 or 2 or 3 or 4 or 5
3	->	0 or 1 or 2 or 3 or 4 or 5
4	->	0 or 1 or 2 or 3 or 4 or 5
5	->	0 or 1 or 2 or 3 or 4 or 5

For the Q learning algorithm, we use the following parameterisation:

- Learning rate, alpha defined in Fig 1
- Reward Matrix R, defined in Fig 2
- Discount factors, gamma = 0.2
- Maximum number of iterations, n=1200

Q LEARNING: BASIC CASE: epsilon-greedy + γ = 0.2

The parameter γ determines the value of future rewards. It affects the learning of the agent and can be static or dynamic. For the small grid, γ was set equal to 0.2 for the first trial and then it was increased to 0.8, while the rest of the variables remained fixed. This was done in order to compare the magnitude of gamma to the learning process. It should be mentioned that gamma can take values between 0 and 1 and has two extreme cases. For γ = 0, the agent will be myopic with the γmax Q(s, a) becoming 0, so it will not consider future rewards but only the immediate rewards. This approach can only be useful when the reward function is described in great detail and at a later section we will present the findings of this trial. A value of γ=1 will make the agent value the future rewards same as the current ones. This means that there is practically no difference in selecting an action now and in 5 moves. This practice was tested and it is presented later on, but as it will be shown, the learning process doesn't work properly with extremely high gamma values.

Q Matrix initialised as zeros

	0	1	2	3	4	5
0	0	0	0	0	0	0
1	0	0	0	0	0	0
2	0	0	0	0	0	0
3	0	0	0	0	0	0
4	0	0	0	0	0	0
5	0	0	0	0	0	0

Q learning matrix after an episode

	0	1	2	3	4	5
0	0	0	0	0	0	0
1	0	0	0	-7.011	0	0
2	70.71	0	0	0	0	0
3	0	0	0	0	-9.142	0
4	0	0	-7.071	0	0	0
5	0	0	7.071	0	0	-9.142

Table 3: Q Matrices for the small grid before and after an episode

The Q matrix represents the "brain" of the agent. Initially was full of zeroes because the agent knew nothing about the environment. Its size was the same with the R matrix, (6, 6). Table 3 (left) shows the original Q matrix and (right) its values after the first episode.

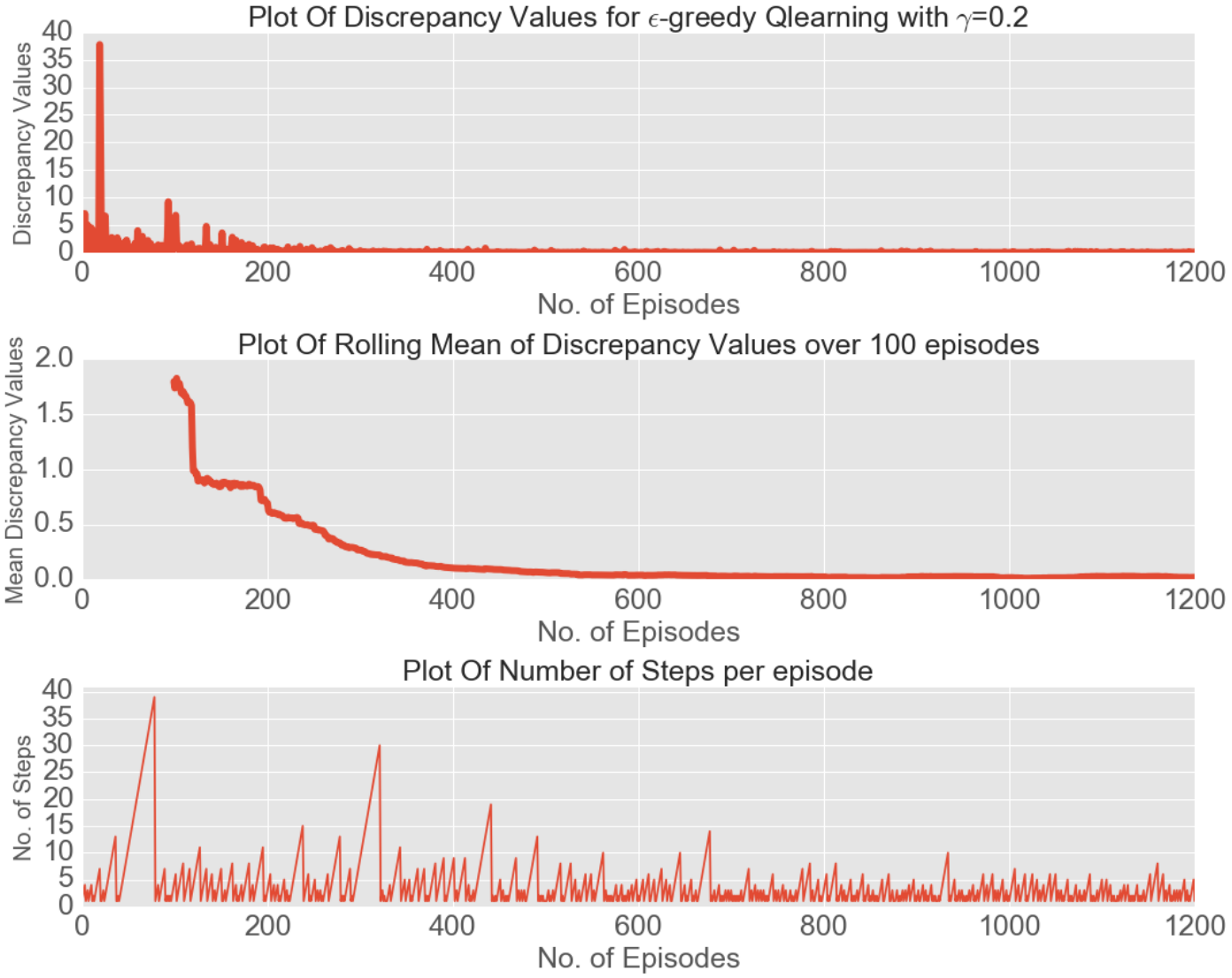

Figure 3: (Top) Plot of discrepancy values for each of the 1200 episodes learnt by the agent calculated according to Equation 3. (Middle) Plot of a rolling with a period of 100, showing the smoothed trend over 100 episodes of the discrepancy values. We observe that the agent converges around 400 episodes without any significant large changes to the Q matrix after this. (Bottom) Plot of the number of steps taken by the agent during each episode. We note that initially the agent explores more so there are a greater number of steps but this drops around 400 episodes and the number of steps reduces due comparatively due its greedy nature.

Figure 4: (left) Final Q Matrix for the small grid case with epsilon greedy policy and discount factor = 0.2. The yellow square represents our terminal state and this corresponds to our R, matrix in Fig 1. The agent has managed to successfully find its way to its destination and also notable in its absence is the lack of large negative values for the Q Matrix. This shows that the reinforcement signal indeed modified the agent's behaviour in avoiding these areas.

Final Q Matrix after 1200 episodes for small grid					
0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
-5.6366	-6.1608	9.9991	-6.8325	0.9279	5.9923
100.0000	1.6172	9.7332	3.8612	1.1246	5.9384
-9.4915	1.1630	20.0000	-6.5812	-6.1124	5.0640
-7.9945	1.0296	4.3090	2.6536	-9.0312	6.0000
-9.7212	1.3623	30.0000	3.7607	1.1401	-4.4128

Table 4: Final Q matrix learnt by agent after 1200 episodes.

Q LEARNING: epsilon-greedy + γ = 0.8

The above results are with an agent that is very short term in nature as defined by its discount factor being so low. So in order to make the agent a more of a long term thinker we increase the gamma value to 0.8 now and consider the results. Since we have shown that the agent behaves as expected in the following sections we present final results only and skip intermediate results.

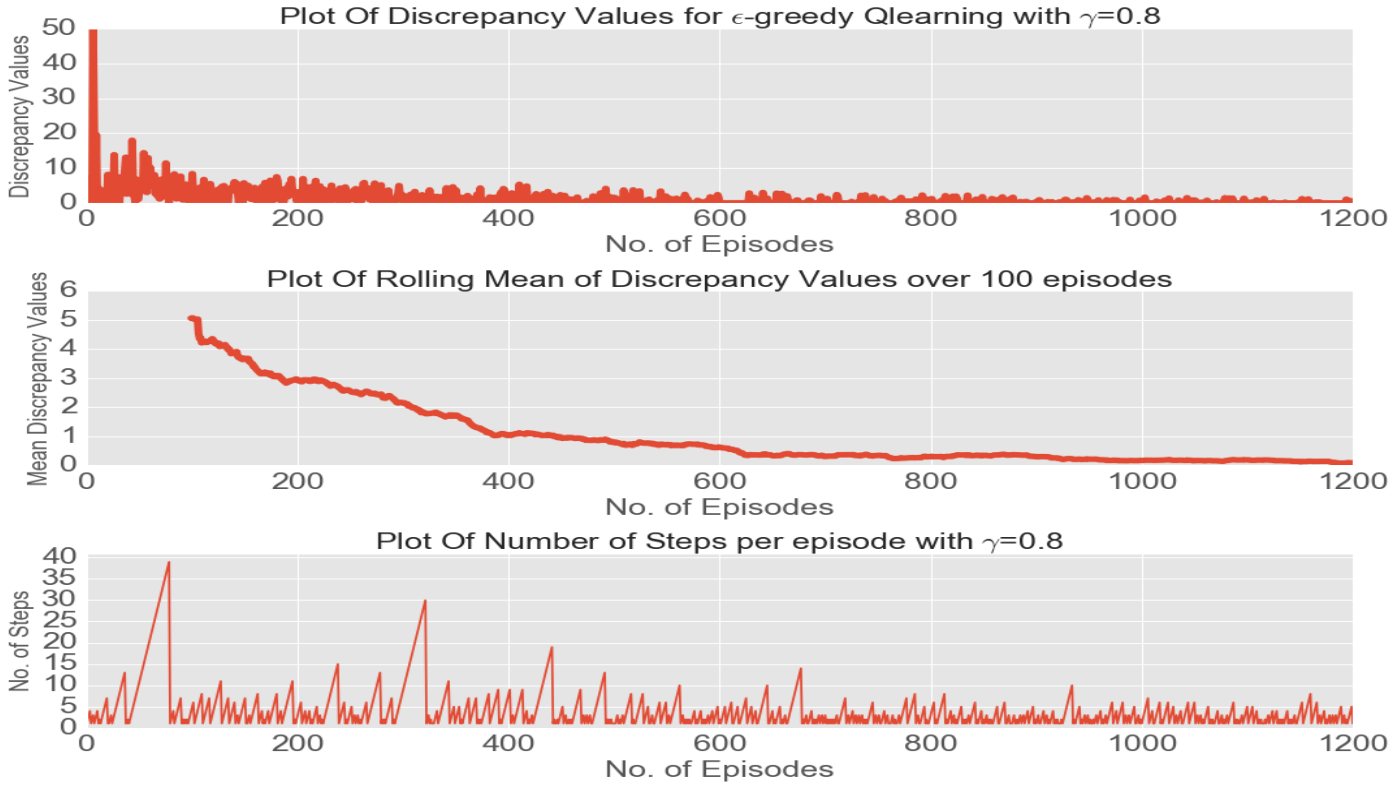

Figure 5: Performance measures for our agents but this time with an increased learning rate. The first point to note is that the discrepancy values are not as smooth as in Fig 3. In the prior case we noticed convergence around 400 episodes but in this case we see that there are significant amount of spikes well into the 600 episode mark. It appears to reach convergence at around 1000 episodes compared to 400 before. The mean plot here shows a much gentler decline over a much greater number of episodes than before. Also the bottom plot of our steps we see that there is a greater amount of steps even in the later episodes in the higher gamma case than the lower gamma case. This indicates that higher gamma values lead to slower convergence due to the long term nature of the agent.

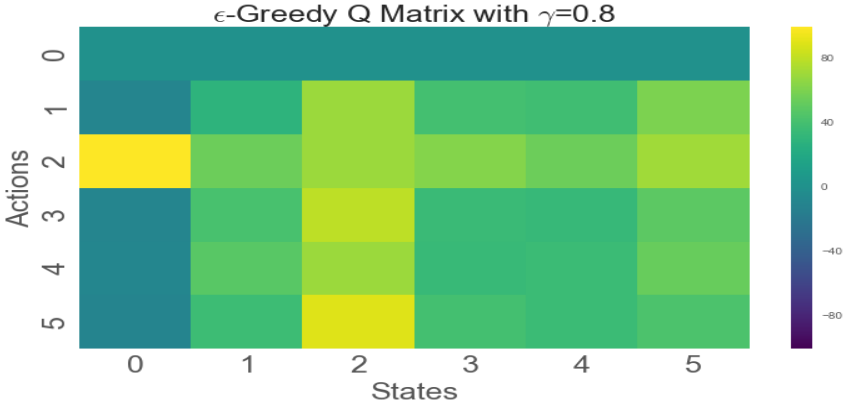

Figure 6: (left) Final Q Matrix for agent after 1200 episodes with gamma = 0.8. The agent eventually finds its way to the terminal state but spends a lot of time in the rest of the grid as part of its long term approach. Hence it accrues higher Q values in some of the other areas defined by the R matrix than the agent in the previous case.

Advanced Case 3: Epsilon greedy + different α

In the previous section we have shown that high gamma value causes the epsilon greedy agent to take longer to converge while the lower gamma values causes faster convergence. So for this set of testing we kept gamma values constant at 0.2 and used three different settings for the learning rate. Firstly, we set alpha as in Equation 4 and then we set it to a fixed value of 0.2 and 0.8 to assess its impact on the agent.

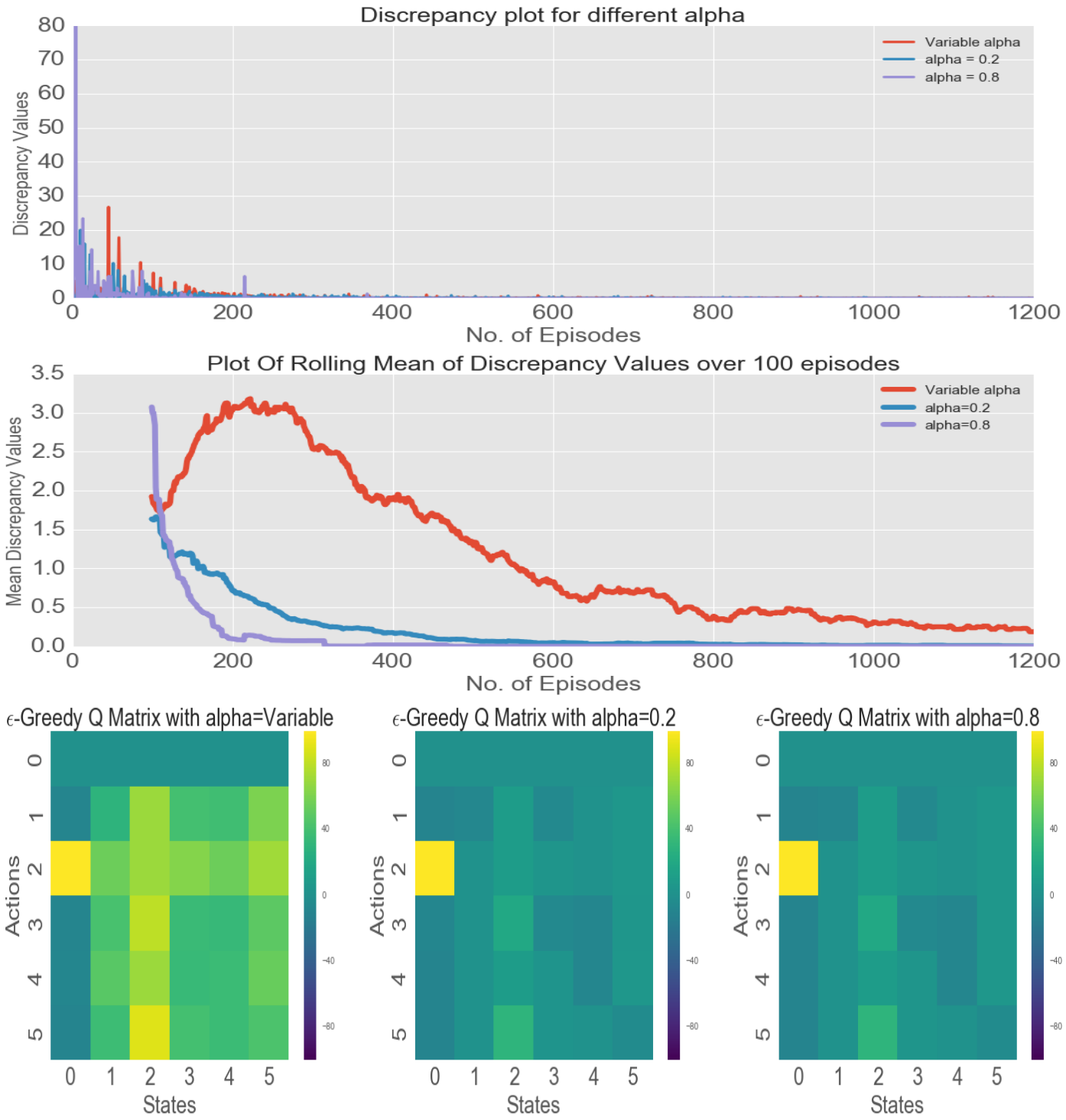

Figure 7: (top) Performance measures for the different learning parameters for the agent and (bottom) associated final Q matrices.

We see that for the different values of alpha the epsilon greedy agent produces fairly consistent behaviour as we can see from the final Q matrix plots. They are nearly identical suggesting that the learning does not have as big of an impact as the gamma parameter which we saw produced much greater variation in the final Q matrix. This is not surprising given how close the trend in the discrepancy plots are. However, it is interesting to see the trends in the rolling mean plot of the discrepancy. Here we see the fixed alpha values converging faster than the variable alpha and higher alpha seems to converge faster than the lower and variable alpha. This could be explained by the update rule for Q learning where the alpha term is used to damp the update terms bigger terms mean bigger damping and while lower terms would cause slower damping of the terms. This is essentially what the plots show so gives us confidence in our implementation of the Q learning algorithm and suggests that despite the variation in the update values the end result for different alphas look similar so there is not much difference to the final solution as to which version of alpha chosen. The only difference would be the convergence time. Even in this case we find that our upper bound of 1200 iterations is sufficient for this to converge.

Advanced Case 4: Softmax Policy + γ = 0.2

Softmax Policy: The softmax function is used in order to pass from a set of values to a list of probabilities to take an action. These probabilities are calculated with the following formula:

$$P_t(a) = \frac{\exp(q_t(a)/\tau)}{\sum_{i=1}^{n} \exp(q_t(i)/\tau)}$$

Equation 5: The calculation of softmax probabilities given an input. The $q_t(a)$ is the expected value of the reward that will be earned if the agent follows the action a. As it can be observed, the action with the highest value will have a probability closer to 1. The second parameter that we have to mention is the temperature, τ, and parameter. For low temperatures, the expected rewards affect the probability while for high temperature values, all the possible actions have quite similar probabilities.

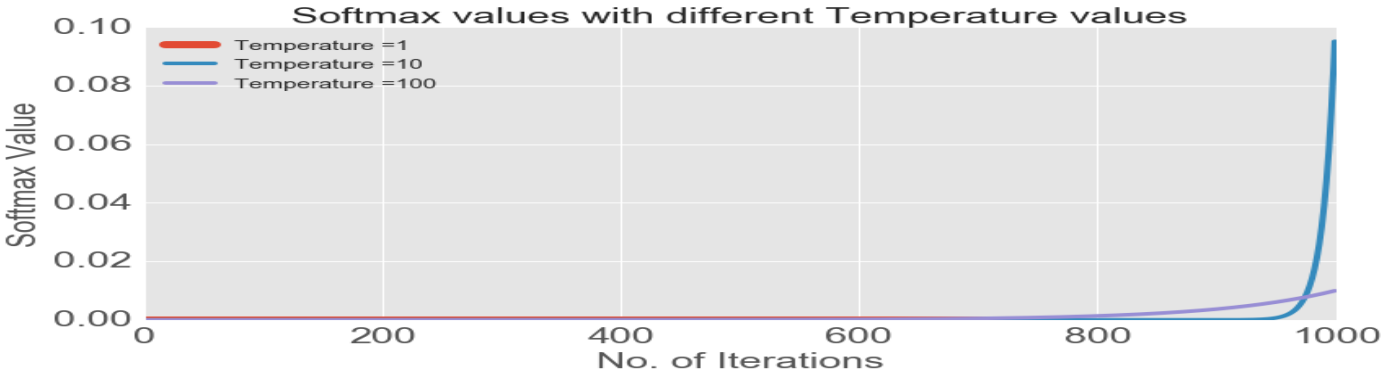

Figure 8: Impulse response for the softmax function with different temperature values. We see that a value of 1 has no effect, while 100 causes a very gentle increase but our chosen value of 10 gives a good balance of increasing probabilities for the maximum number of iterations we have chosen.

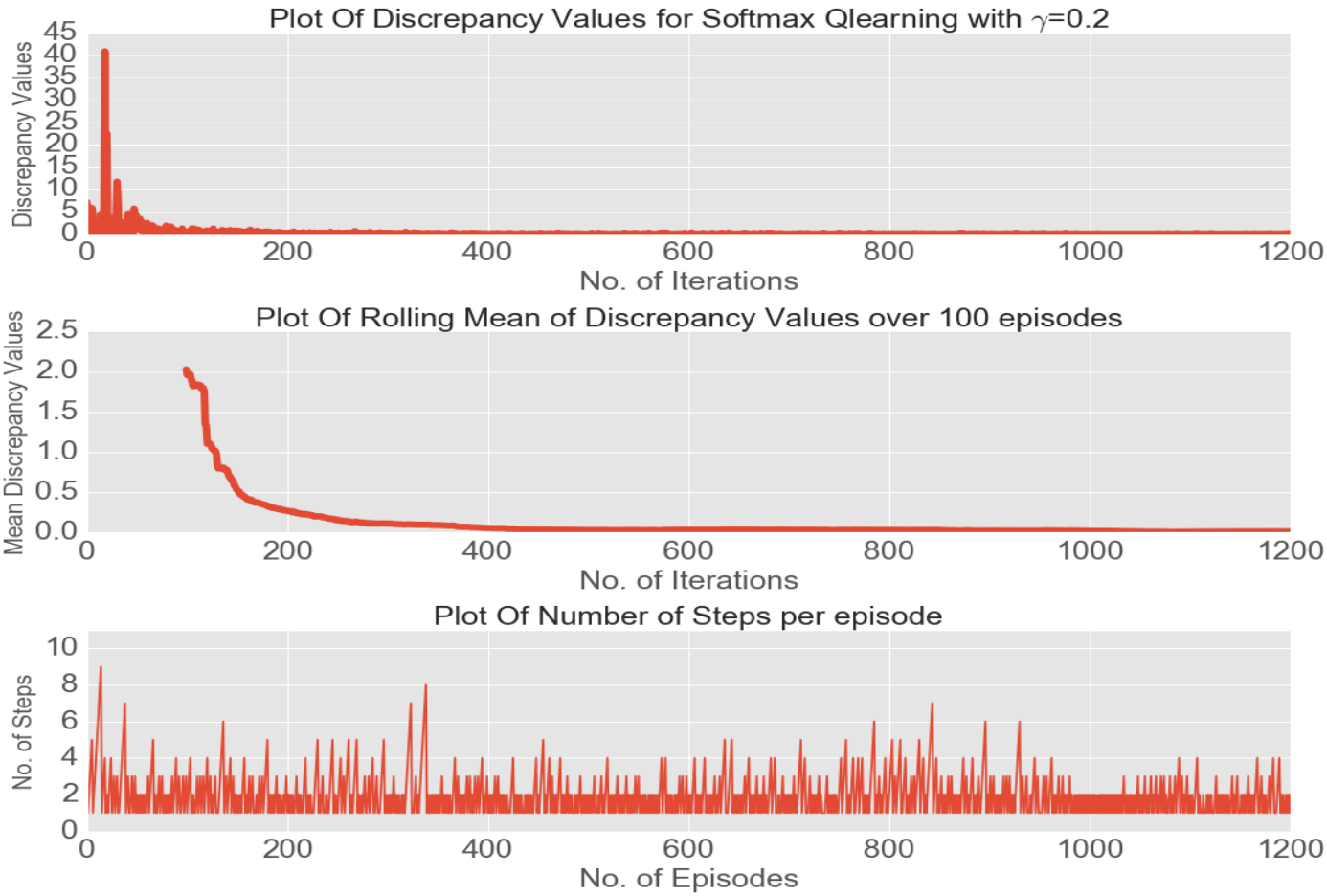

Figure 9: Softmax policy performance with gamma = 0.2. We see that this policy is clearly superior to the epsilon-greedy approach for this problem. While the agent with the epsilon greedy and gamma = 0.2 converged around 400 episodes we see that the agent here converges around 250 episodes in the mean discrepancy plot. Also from the top plot, the agent learns the highest Q values in under 100 episodes.

Advanced Case 4: Softmax Policy + γ = 0.8

To enable us to compare with the epsilon greedy gamma = 0.8 case we repeat the previous experiment with the icnreased gamma value.

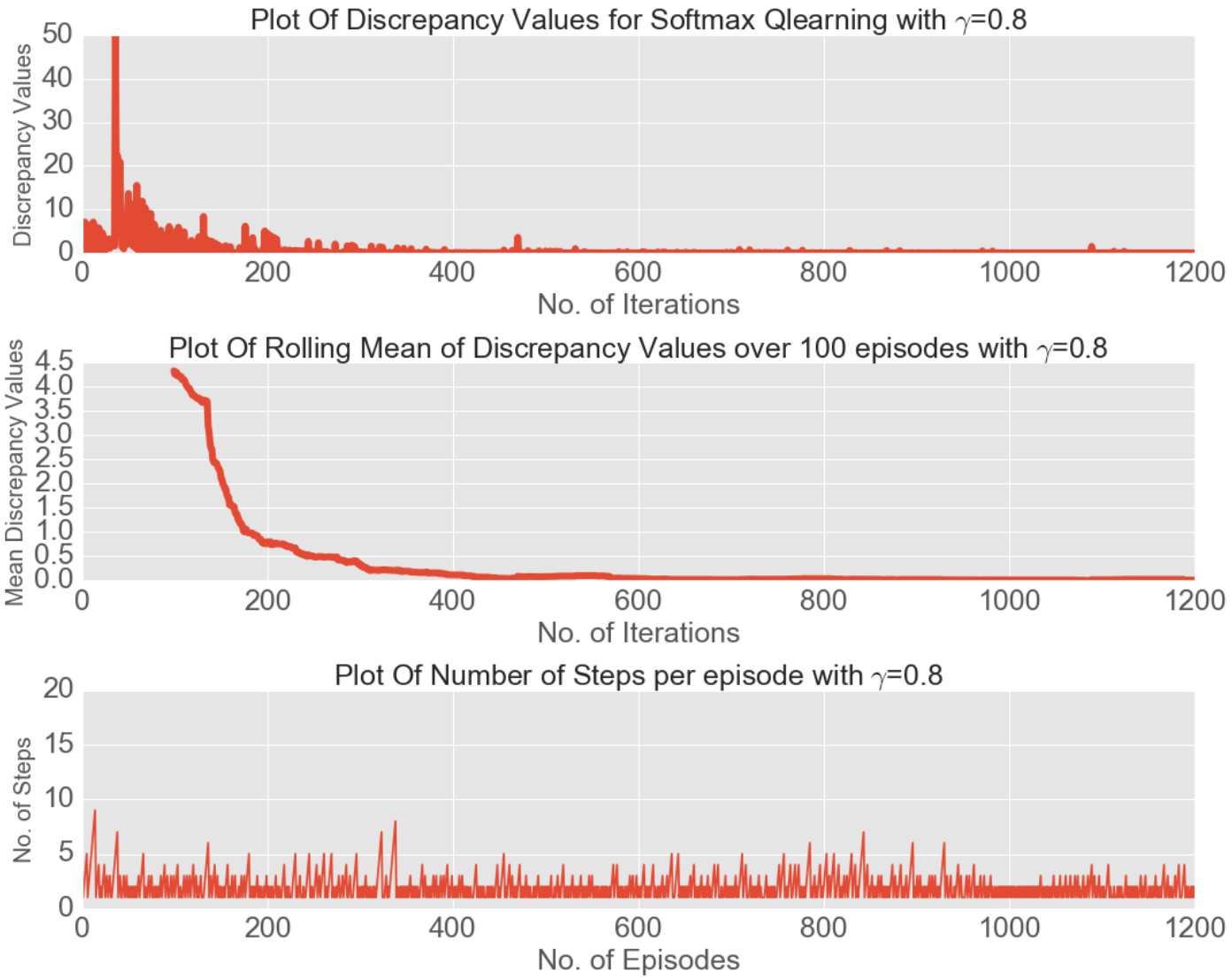

Figure 10: Plot of performance measures for Softmax with increased gamma values. Here the softmax really shows its stability over the epsilon greedy method. With the similar epsilon greedy case we saw that the convergence was pushed back to around 1000 episode mark but even with the increased gamma the softmax converges around 300 episodes. This is much less than the increase observed in the epsilon greedy case. The trends in the discrepancy and mean discrepancy are consistent with Fig 9.

Advanced Case 4: Comparison of Softmax and Epsilon-greedy

In this section we consider the two policies that we have tested so far in our small grid problem: epsilon greedy and softmax. We analyse the gamma value 0.2 and 0.8 cases for each policy together.

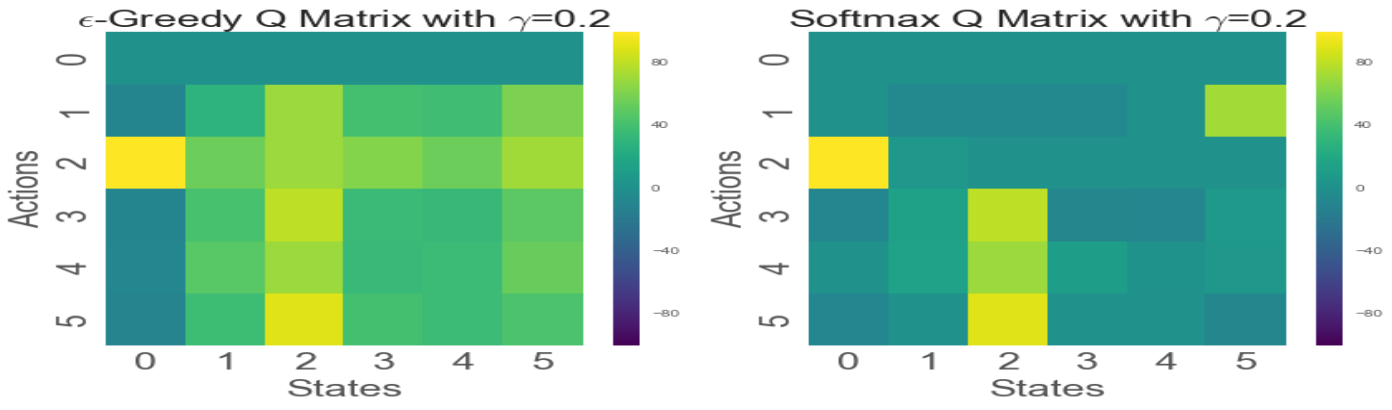

Figure 11: Comparison of the final Q matrices learnt by the agent with the different policies for the same gamma values. We do not observe much difference here and in both cases the agent finds the terminal state although it does so faster with softmax as shown in Fig 9,10.

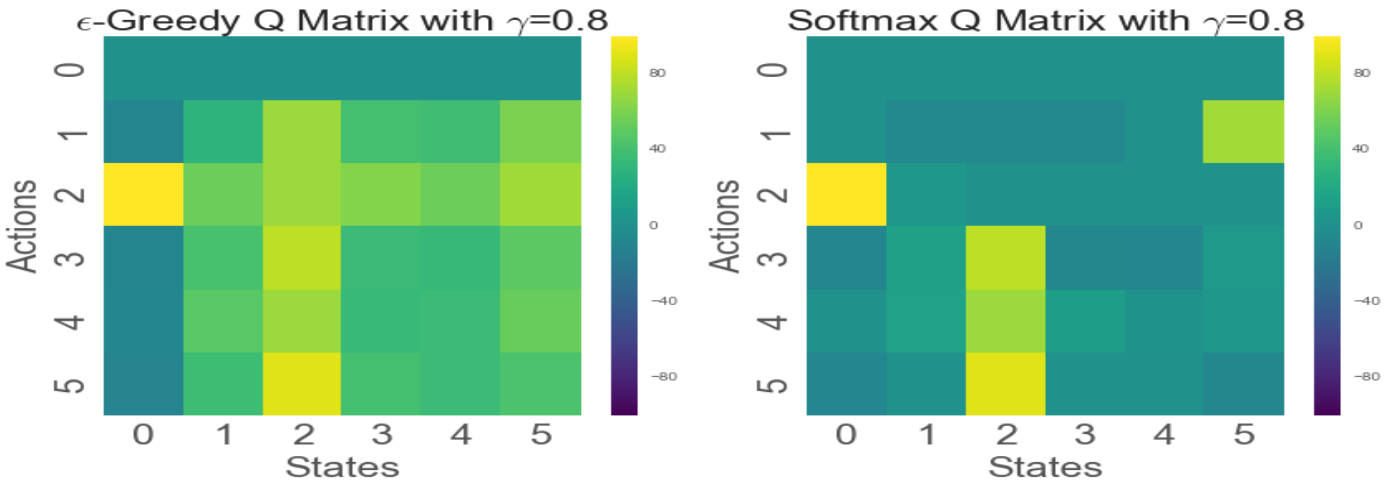

Figure 12: Comparison of final Q matrices for the agent with gamma = 0.8. The divergence between the two policies is much clearer at this value of the discount factor. We see that in both cases there is more exploration by the agent of the rest of the grid and we see more bright spots in the Q matrix as a result of this. This is to be expected as the agent is taking a more long term view of rewards but amount of the bright spots in the map is of interest as we see that the softmax policy is much better. The agent with softmax explores but its exploration is not erratic and all over the map as it is with the epsilon greedy policy.

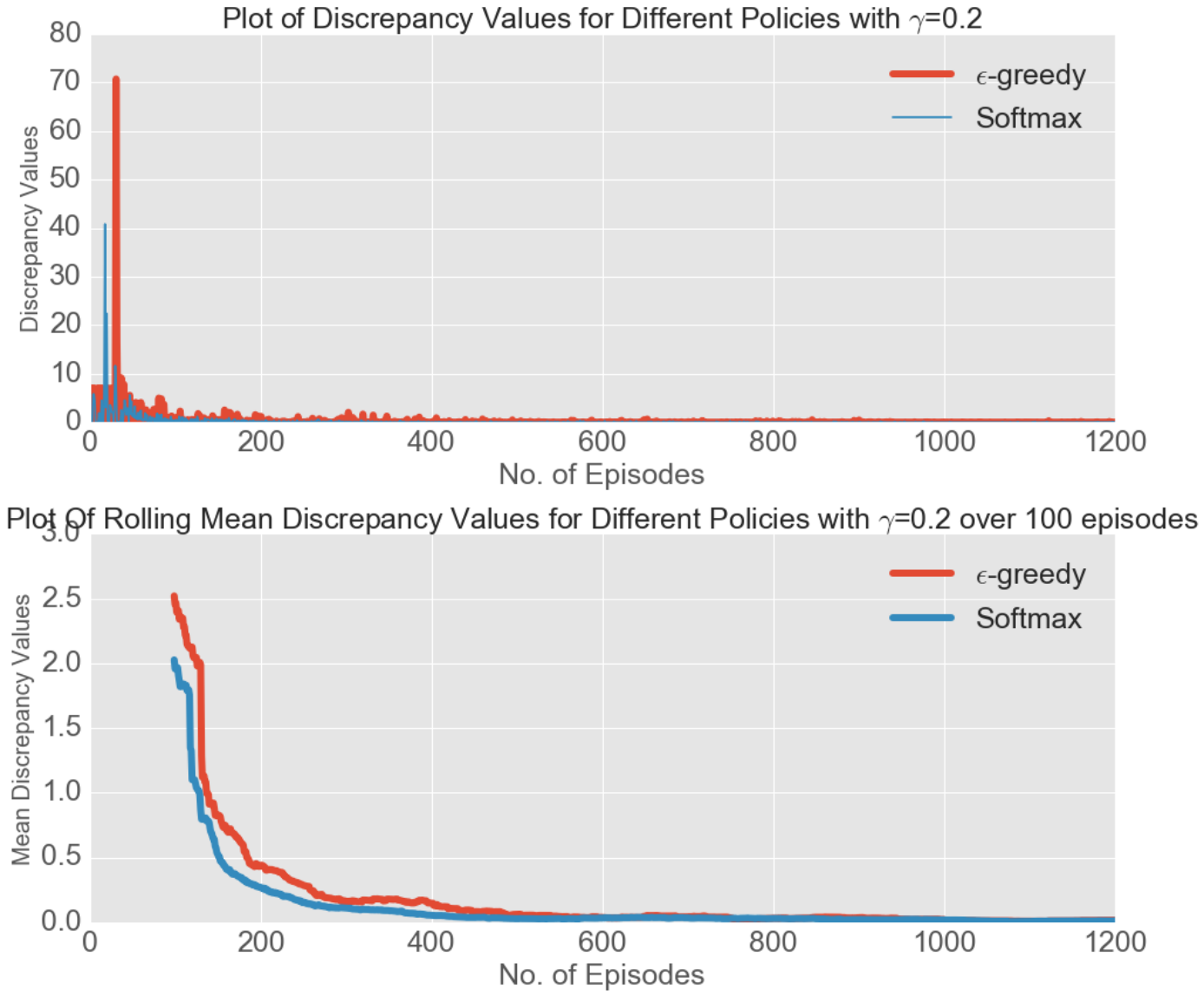

Figure 13: Performance comparison for epsilon greedy and softmax policy for gamma = 0.2. We see that the performance is similar for both policies with this gamma values however, the softmax policy has Q values of smaller magnitude in comparison to the epsilon greedy and converges faster as noted previously.

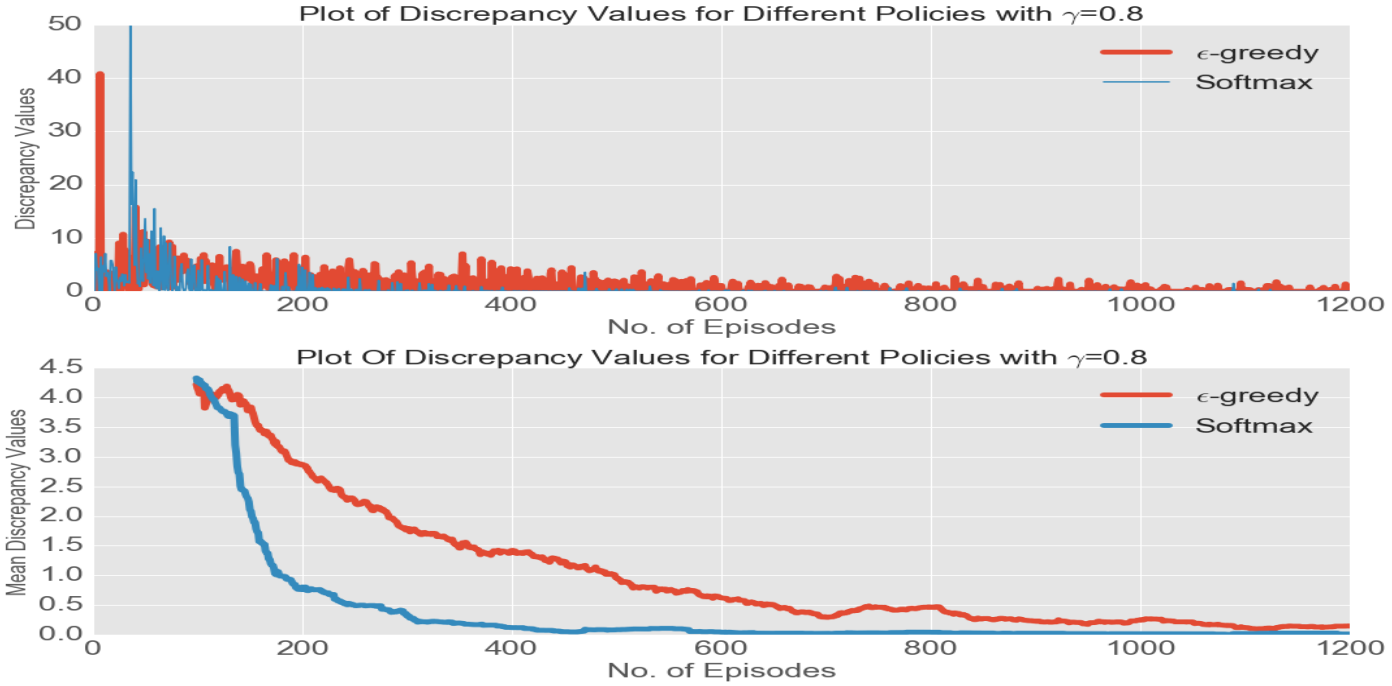

Figure 14: Performance plot of epsilon greedy and softmax policies with gamma = 0.8. As noted we see clearly in the plot of the mean discrepancy that the softmax does not require more episodes to converge as the epsilon greedy does with higher gamma values but still converges on a similar number of episodes. From this we can say that an epsilon greedy policy is more susceptible to gamma parameter changes than the softmax policy.

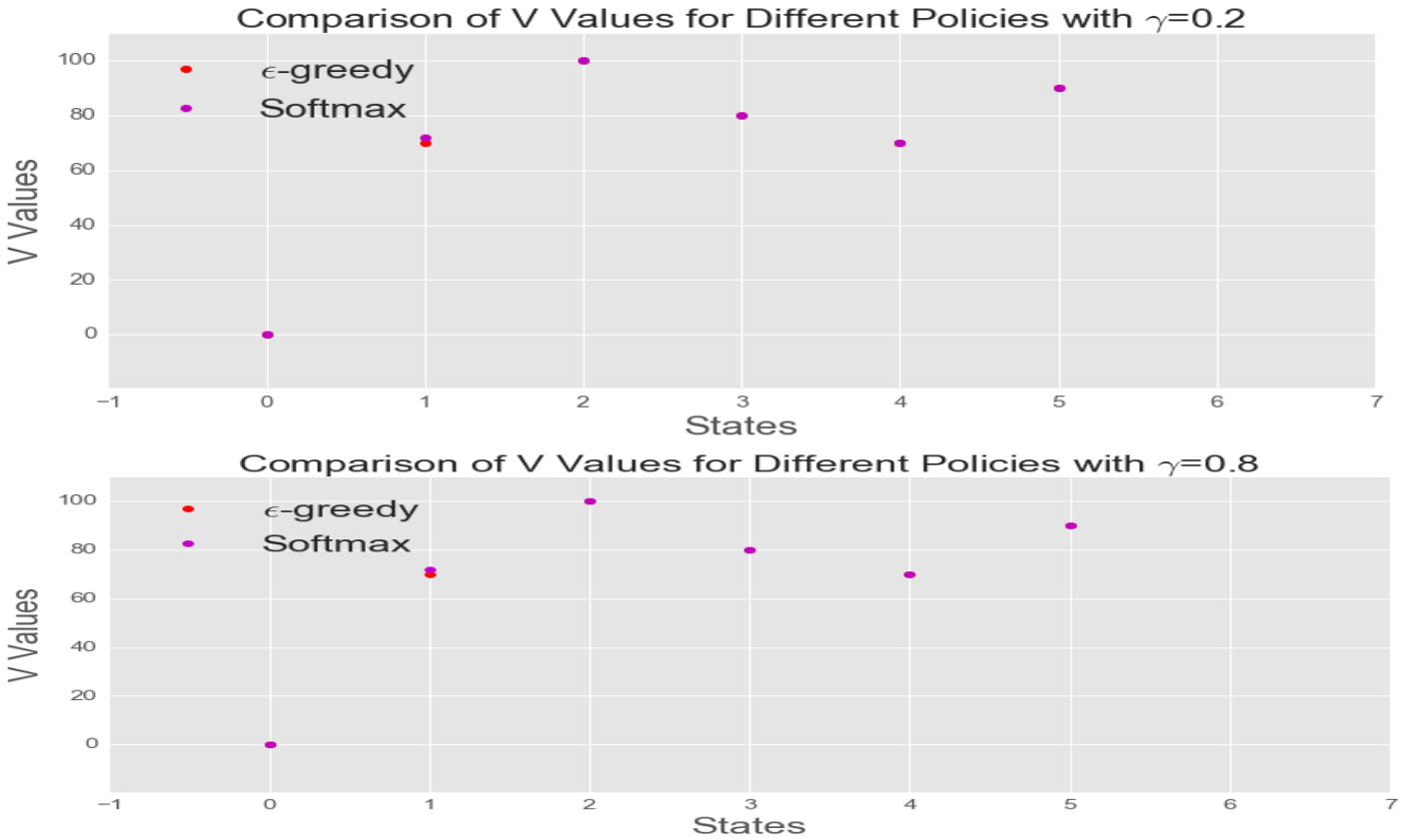

Figure 15: (top) V values for both policies with gamma = 0.2 and (bottom) with gamma = 0.8. We define the V values as the maximum Q values along the column direction. The softmax learns mostly similar V Values as the epsilon greedy with only one point being significantly different for small gamma. With the higher gamma value however we more variation in the V values of the epsilon greedy policy. This again gives evidence that gamma affects epsilon greedy policy more than softmax policy as the V values for the softmax are fairly similar in both cases. This could be explained by the fact that the softmax assigns higher probabilities for actions over increasing episodes which favours exploitation rather than exploration. With the epsilon greedy policy the greedy nature of the policy probably does not complement the longer term outlook we are trying to encourage the agent to take with the higher gamma values. Therefore it ends up being greedy in its exploration rather than the exploitation which was the case with low gamma.

Q LEARNING: STOCHASTIC CASE

This set of analysis is conducted with the MDPTOOLBOX. Also, the MDPTOOLBOX represents a state transition probability matrices and the reward matrices as a 3 dimensional matrix of the form (Action, State, State). When these hypothesis spaces are visualised, we stack them in the vertical direction to allow representation onto a 2D plane. Also the default number of iterations is 10K and we use this for all the analysis unless stated otherwise. The alpha value is defined as a function of the number of episodes as given in Equation 4. The policy selection is epsilon greedy with increasing probability also defined as a function of the number of episodes.

$$\epsilon = 1 - \frac{1}{\log(n+2)} : n \ = number\ of\ episodes$$

Equation 5: Epsilon greedy policy selection as function of the episode number.

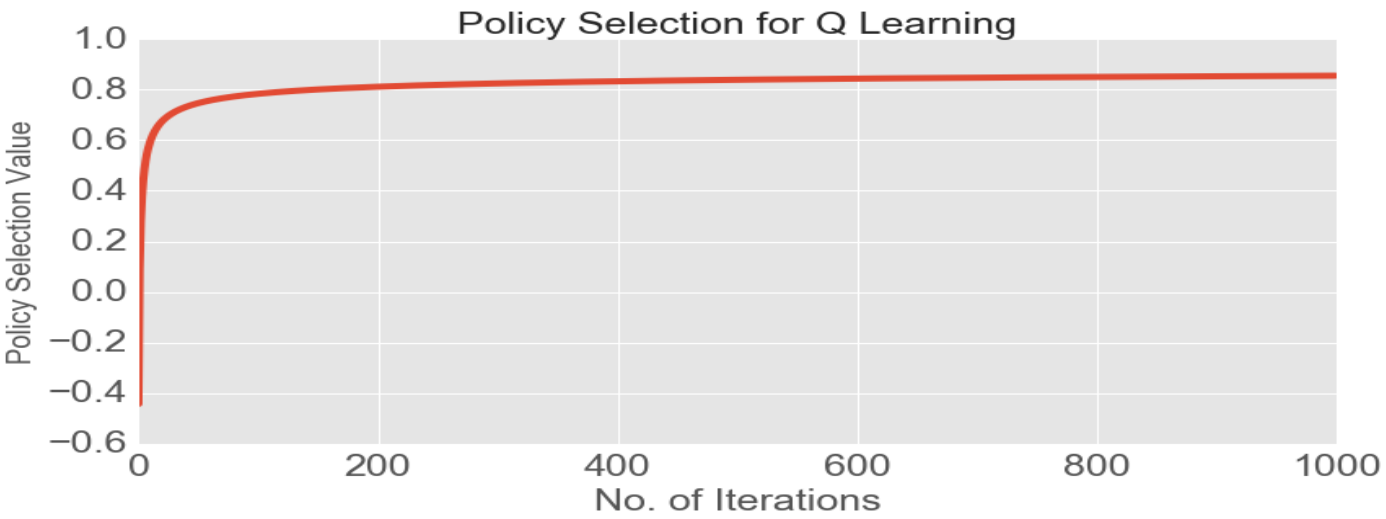

Figure 16: Impulse response for the epsilon greedy policy selection for the stochastic case

We see that the value increases over time meaning our agent gets greedier as the number of training episode increases so we move from more exploration to more exploitation. This is in our opinion better than having a static value of epsilon as it allows for a good balance between exploration and exploitation. This strategy allows the agent to follow exploratory moves at the first episodes and as it approaches the end of the iterations, its behaviour becomes greedier. In this way, the agent learns an optimal solution.

How it works

This package solves for discrete time Markov Decision Processes. It also features a rich set of algorithms such and utilities to solve these problems. We will use the Q Learning implementation from here to analyse the performance our agent in a stochastic environment of increasing dimensionality. The package features an 'examples' method which allows the user to generate a valid MDP consisting of valid state transition probabilities and associated reward functions. We use the random function from the example class to generate an MDP of the desired size. So the size of our stochastic grids increase from (6, 6) to (10, 10) and finally to (100, 100).

This function takes as an argument two integers for the number of states and actions in addition to return a sparse matrix. We choose the dense matrix representation of our problem. Based on the input number of states and actions the state transition probability matrix, P is initialised with zeros in the form (Action, State, State) and the Reward matrix, R is also initialised as zeros in the form (Action, State, State). Then for each value in the range of actions and states the function iterates over the initialised matrices to populate them. The values are draw from a continuous uniform distribution over a stated interval. For the P matrix this is [0, 1] because it represents probabilities. For the R matrix, the range is [-1, 1] because we want to reinforce our agent to avoid some areas while reaching high probability areas. The function also checks the P matrix to ensure that each state has at least 1 transition so the agent does not get stuck. The P matrix is then normalised to ensure that the probabilities are valid probabilities and the R matrix is checked to ensure that its values lies in the [-1, 1] interval. The output is a dense P and R matrix. We use this function to generate valid probabilistic MDP of dimensions (6, 6), (10, 10) and (100, 100).

So far we have presented results of a deterministic Q learning agent and have seen that the discount factor has a strong impact on its performance. We would like to assess whether we see the same pattern in our stochastic case. To do so we consider a (6, 6) grid again and consider gamma values of 0, 0.5 and 1. The rationale for choosing these values are to see how the agent performs at the edge of myopia and long term outlook.

This section is organised as follows. We first present the case where we evaluate the extreme gamma values and its impact on a (6, 6) stochastic grid. We then evaluate a larger stochastic grid, hence a different transition and reward matrix and evaluate the effect of increasing the number of iterations from 10K to 20K on the performance of the agent. Then we expand the scope of our problem by considering a Big Grid of (100, 100) and evaluate the agent with gamma = 0, 0.5 and 50K iterations. Lastly, we conclude by comparing all the different cases and presenting our conclusions.

Q LEARNING: Advanced Case 2: Different γ on stochastic grid

We first present our Transition Probability Matrix, P and the Reward Matrix, R for the stochastic case in Fig 17. We only have positive values in the P matrix as they represent probabilities.

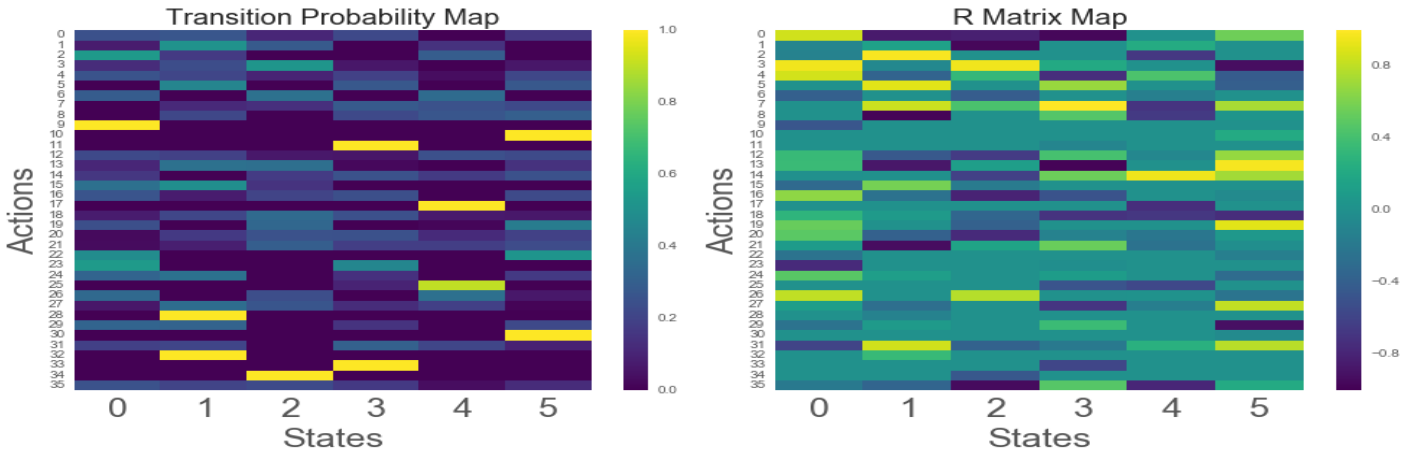

Figure 17: (Left) Probability Transition Matrix, P for our grid problem. The light colours represent areas of high probability and dark colours represent areas of low probability. (Right) Reward Matrix, R through which we tell our agent how to determine the optimal policy. Here the light areas represent where we would like our agent to go and the negative values or dark areas where we would like the agent not to go.

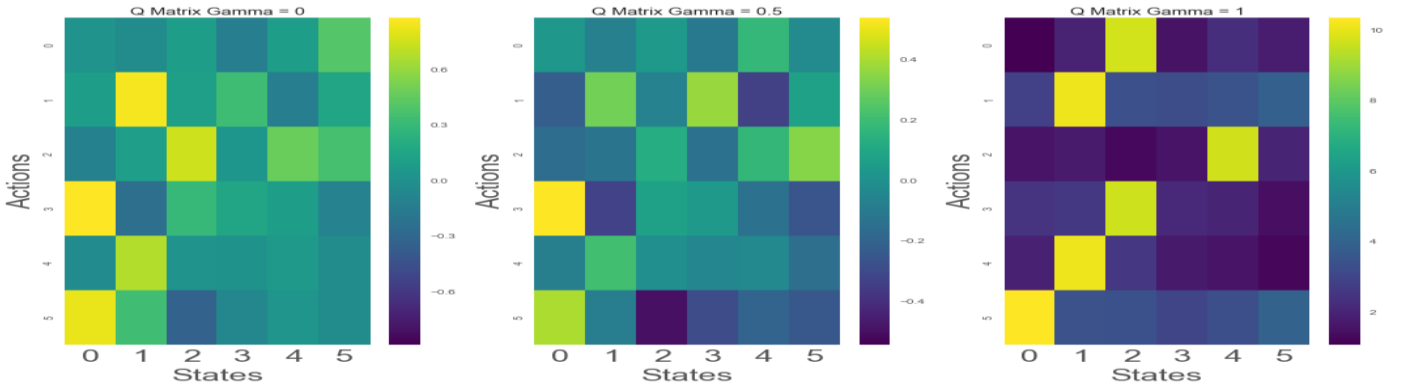

Figure 18: Final Q Matrix comparison plot for Q Learning with different gamma values.

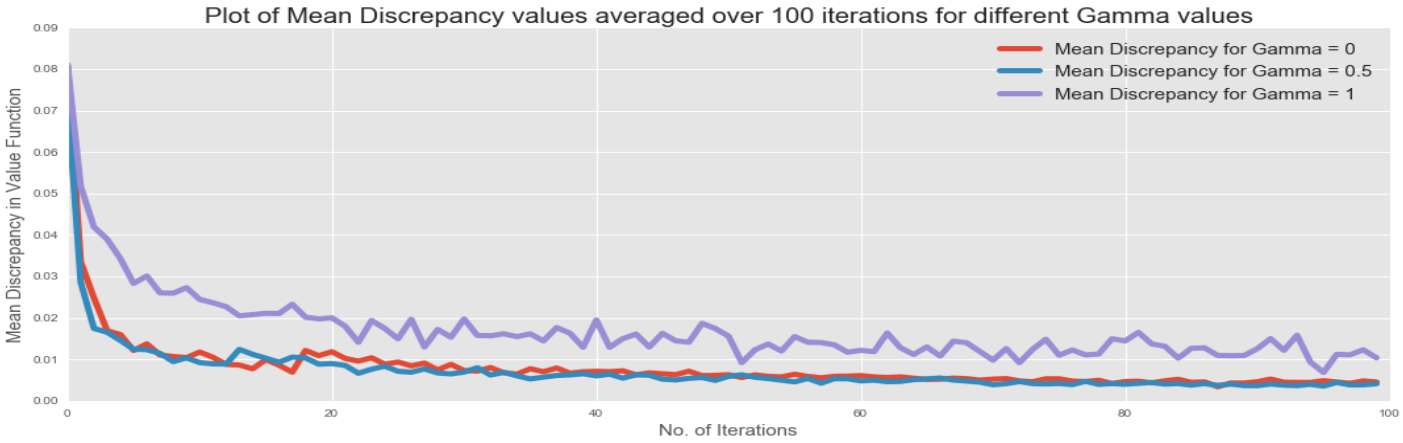

Figure 19: Mean Discrepancy plot for 3 runs of Q Learning with different gamma values. We observe that the 0 and 0.5 cases achieve convergence around 40 while the 1 does not appear to converge. When gamma parameter is equal to one, the agent weights future rewards equal to the current ones. This means that there is no difference between the reward of an action taken now and taken some moves later. In this extreme case, learning does not work and consequently, it fails to converge. **The mean discrepancy is averaged over (n/100) episodes**.

We see that there is quite a lot of variation between the final Q matrices the agent learns with the different gamma values. Of the different Q matrices we see that the 0 and 0.5 case are fairly similar in the sense that they learn almost similar areas of high Q values although the magnitude in the 0.5 case is slightly smaller. But the gamma = 1 produces a completely different map compared to the rest so we can assume that this value failed to converge. Again, we have strong evidence that gamma values have a strong impact on an epsilon greedy learning agent and that higher values lead to longer convergence times and where it approaches 1 it fails to converge. The curious case is that of gamma = 0 where we would have expected to observe similarly extreme behaviour. But the agent is more stable in its myopia than its long term outlook. This could be down to the probabilistic nature of the P and R matrices for this particular analysis.

Q LEARNING: Advanced Case 5: Different state and reward functions

Here we generate a (10, 10) P and R matrix. Since the dimensions are different the state and reward functions are also different compared to the previous (6, 6) case. In this section we run our Q Learning analysis again with the gamma value of 0.5 and conduct our experiment on a larger grid. We choose 0.5 as this gives a stable result as noted in the previous section. Here, we start with a (10, 10) state transition probability matrix and reward matrix. We run our Q learning algorithm with 10K and 20K iterations and compare the results. We see that both the cases converge to a similar solution after around 5000 episodes. Note that **the mean discrepancy is averaged over (n/100) episodes.**

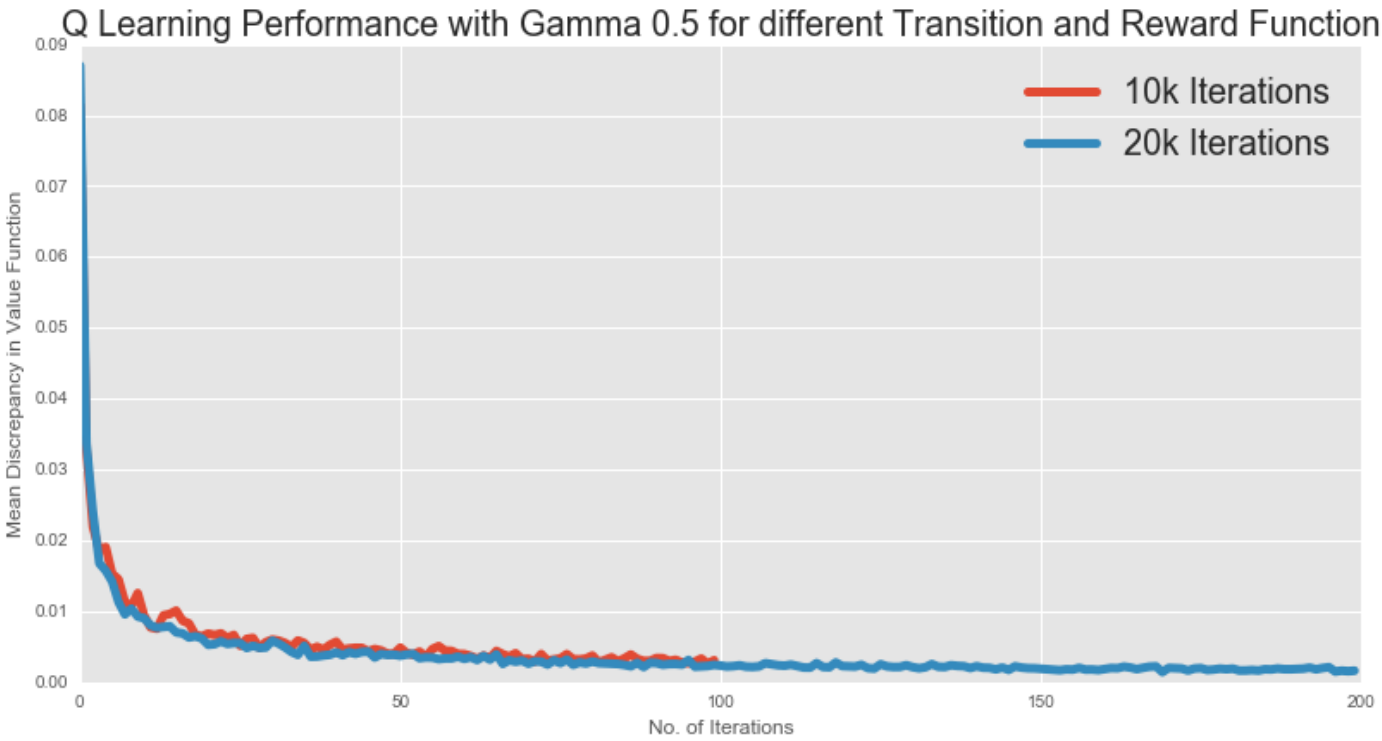

Figure 20: Mean discrepancy plot for Q Learning on a (10, 10) grid with gamma = 0.5. The 20K iterations has more iterations hence the scale for this is longer than the 10k case. Note that these values are smoother over 100 episodes. Both runs seem to achieve a stable solution.

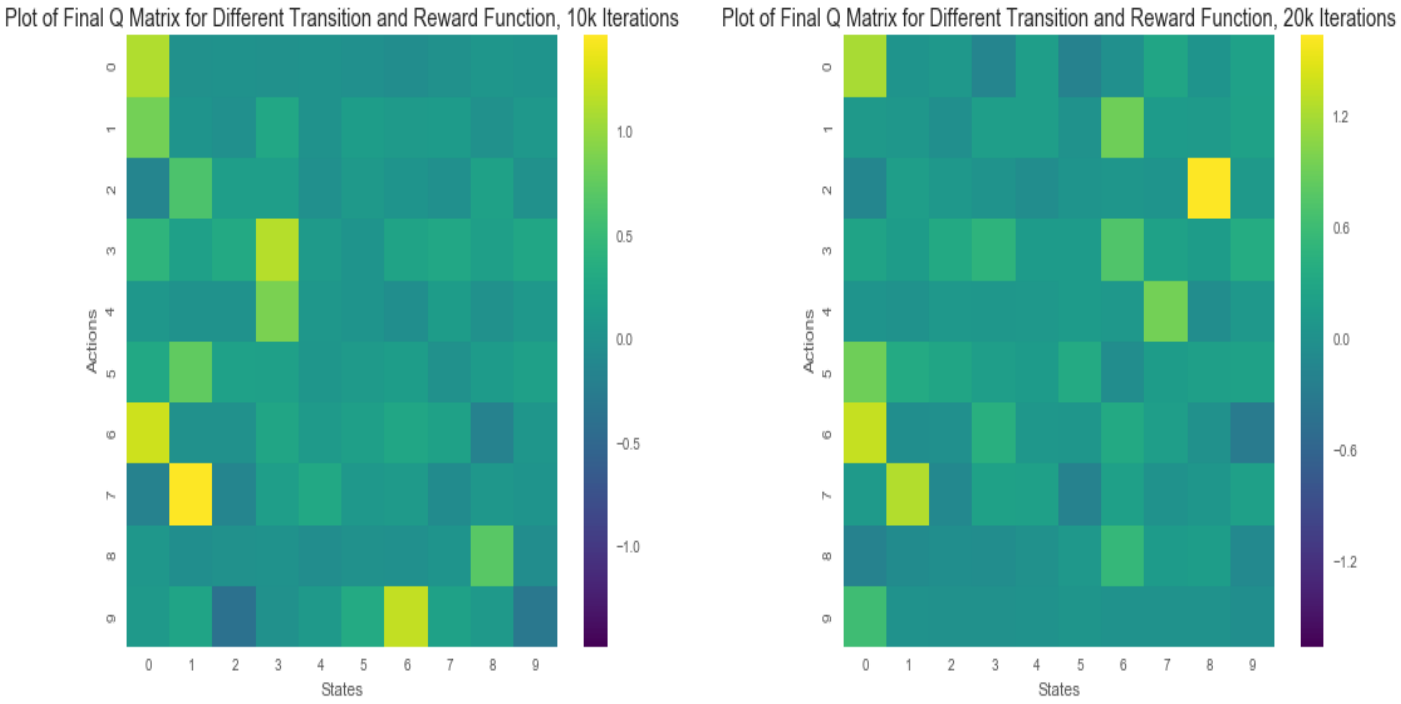

Figure 21 Final Q Matrices for the (10, 10) grid case. We see that some areas have high values and are common to both while the 20K case finds other high value regions.

Q LEARNING: Extra 1: Expanding the scope of the problem

We expand the scope of our problem by considering a Big Grid of (100, 100) and evaluate the agent with gamma = 0, 0.5 and 50K iterations.

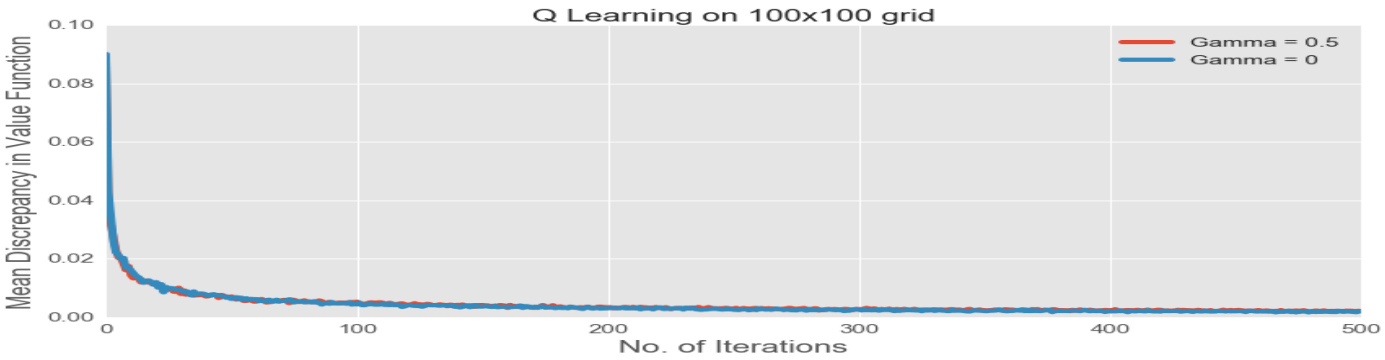

Figure 22: Plot of performance for the agent in a 100,100 grid with discount factor or 0 and 0.5. We see that over 50,000 iterations both the solutions achieve a fairly similar convergence.

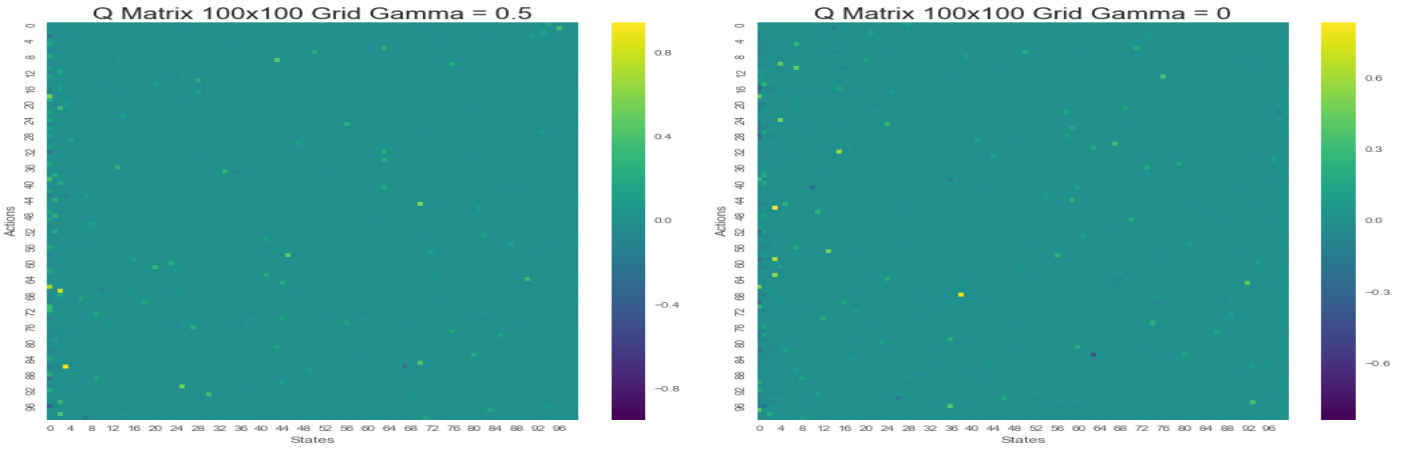

Figure 23: The final Q Matrix for the different cases. The agent with a discount factor of 0.5 in general learns Q values overall of a higher magnitude than the agent with discount factor 0. But we see bright spots on Q matrix for the agent with discount 0 than that of agent with discount 0.5.

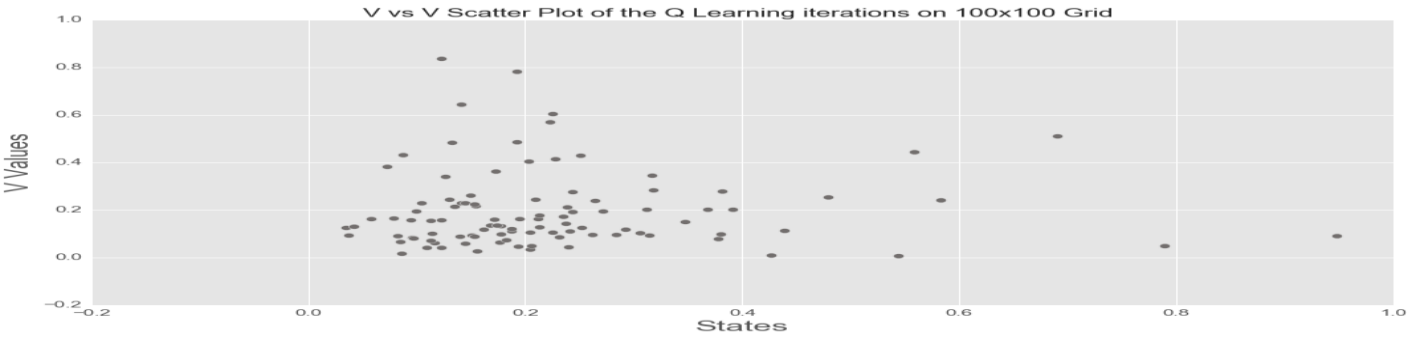

Figure 24: Plot of V values for both agents showing that in both cases they learn comparable V values due to the fairly close clustering of points with a few outliers present. These could represent the high probability areas that they find.

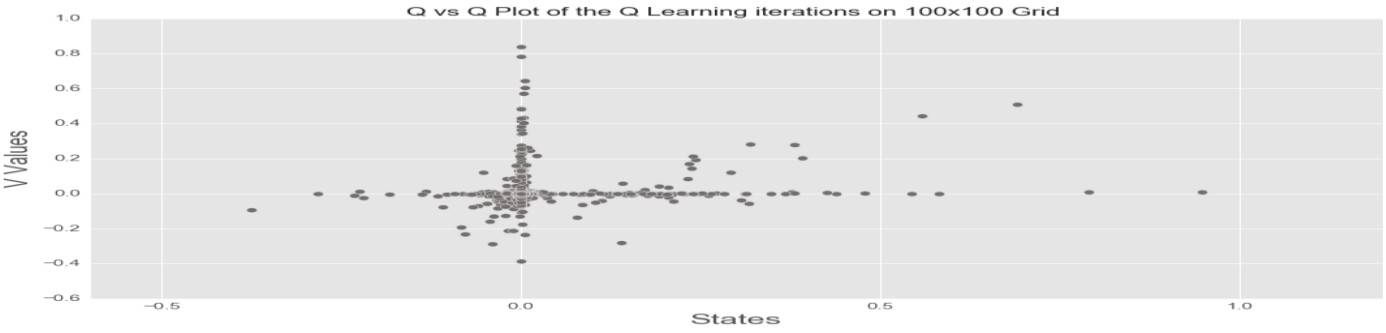

Figure 25: Plot of Q values for both agents showing that most values are clustered close to zero but there is large amount of variation in the Q values learnt. We can also see the direction the agent was going based the trajectory of the Q values.

Q LEARNING: Extra 2: Comparisons Stochastic Cases

Firstly, we consider the different policies that is learnt by the agents under different circumstances in the different hypothesis spaces show in Fig 26. We see when our problem space is fairly small and we vary the discount parameter. The optimal policy learnt by the agents with discount factor 0 and 0.5 are similar while that learnt by agent with discount 1 is highly variable. This suggests that with this value the algorithm probably does not converge within the maximum number of iterations hence could be a reason for the variability.

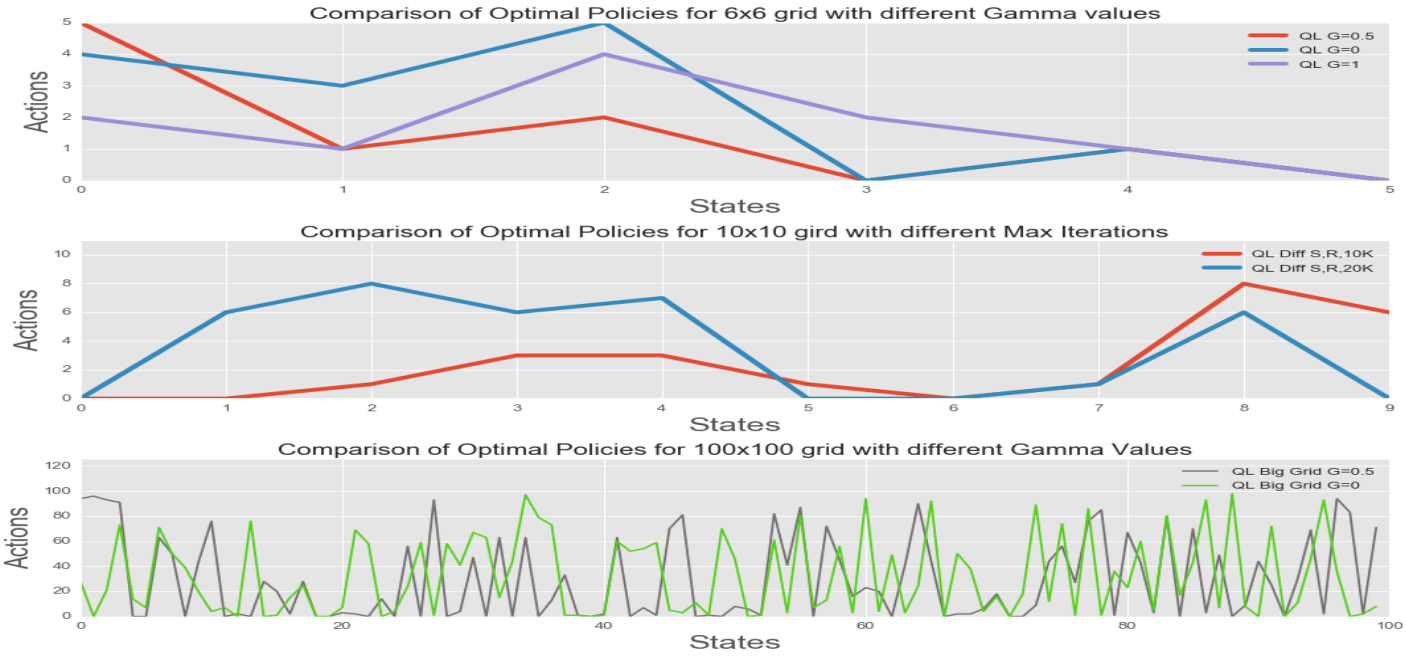

Figure 26: Comparison of optimal policies learnt by the agent for the different cases considered

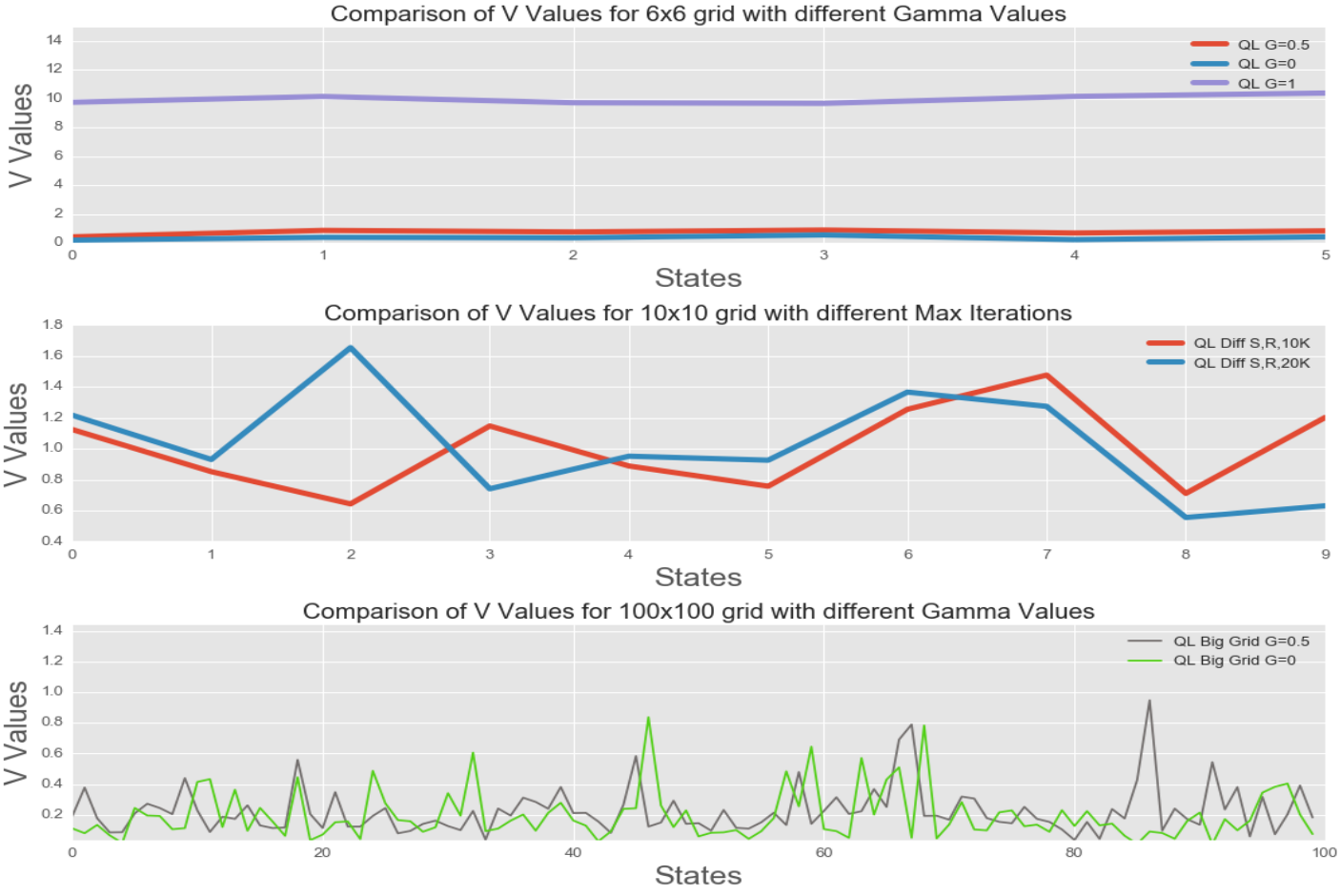

Figure 27: V Value comparison across the different runs. The V values learnt by the agent across the different cases we considered. The overall trend is that the agent learns similar V values in all cases where the discount factor is 0 and 0.5 but the case where it is 1 the values diverge sharply. This could be due to the convergence issue we have identified.

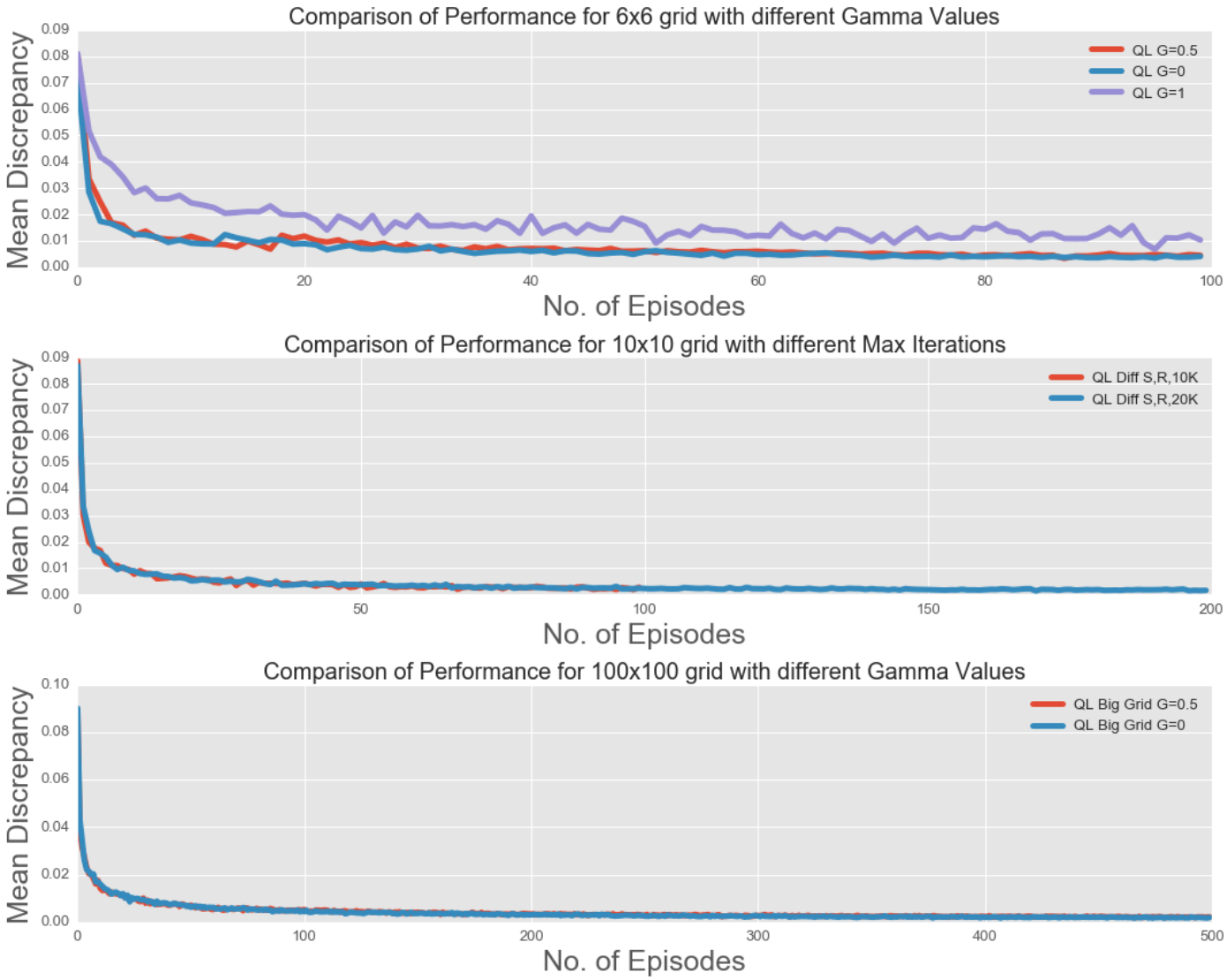

Figure 28: Mean discrepancy over the different runs. The mean discrepancy over the different training runs converge and again we see evidence that discount factor 0 and 0.5 give most stable results even when we vary the number of iterations and enlarge the hypothesis space. The discount factor 1 shows a lack of convergence. This is one the reasons we opted to perform the remaining experiments with discount factor other than 1.

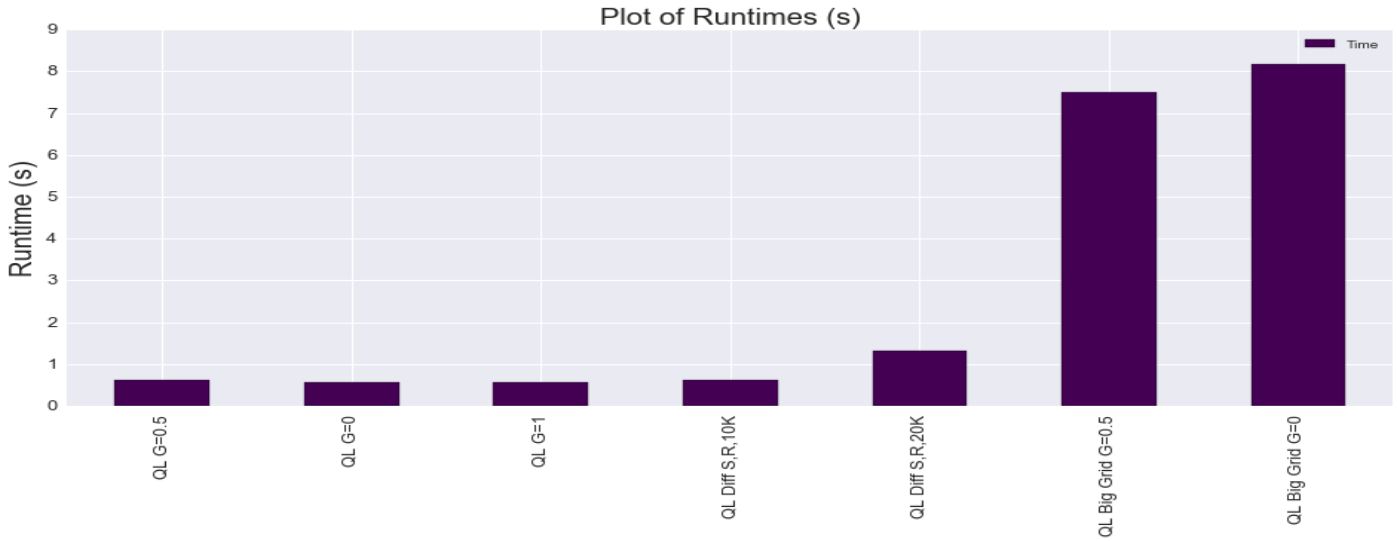

Figure 29: Plot of runtime in seconds for the different experiments. As expected we see that 20K iterations, take longer than 10K iterations and 50K iterations. However we note that discount factor of 0.5 for the 10K iterations was faster than the other values. From a performance perspective it is important to note that a 5 times increase in the number of iterations leads to almost an 8 time increase in the runtime. This suggests nonlinear complexity and could be critical when considering larger hypothesis spaces.

DISCUSSION

Prior to our experiments, we believed that the Q Learning algorithm would be equally affected by changes in parameterisation and that some parameters were not more significant than others. So we expected for a given policy that the agent would behave differently for similar values of gamma and alpha. This intuition is based on Equation 1, because a higher gamma value makes the agent take a longer view but a higher alpha value cause the agent to become greedy as a larger term implies a larger decay of the update value. We also this to be the case in our experiment as Fig 7, shows that for a fixed value of 0.8 we achieve faster convergence then when the value is low or decaying over the number of episodes. Depending on the nature of the problem we would recommend starting with a moderate value of alpha and then tune upwards or downwards as necessary. For our application a high learning rate seems to be beneficial in terms of number of episodes required before convergence. We would recommend a starting value between 0.5 and 0.8 for future applications for similar problems.

With respect to the discount factor we observe some deviations from our expectations. We expect that the agent will take a longer term view when this value is high and act greedily when this value is low. But in our stochastic case with variable gamma values we see that the gamma = 1 case fails to achieve convergence. However, it should be mentioned that what we have deemed to be a lack of convergence over 10K episodes might converge over a larger range. This could be investigated in further work. But the other surprise is the similarity in performance between the gamma =0.5 and 0 cases. For discount factor 0 we expect the agent to be myopic and only consider the immediate rewards. However, in practice for our experiment this is not the case. One possible explanation for this deviation is potentially due to the probabilities associated with the P and R matrix as generated by the package. We find that in general for the discount factor, $0 < \gamma < 1$ are stable. We would recommend a gamma = 0.5 as a good starting value to avoid convergence issues or making the agent far too myopic.

We did not expect that the effects of the learning rate and discount factor to be so policy specific. We have shown that there is almost a coupling between epsilon greedy performance and these parameters. But the same is not true for the softmax policy which appears much more invariant to parameter changes and produces consistent results despite parameter variations. So we recommend using softmax as a policy because it achieves faster convergence and shows greater invariance to parametrisation than epsilon greedy.

We have presented analysis of the Q learning algorithm in both its stochastic and deterministic case. We have compared the parameter variations between them and have gained some interesting insights. With regards to policy, it appears that a softmax policy is much more stable than an epsilon greedy policy at both high low values of the discount factor. We have shown that for an agent with an epsilon greedy policy the convergence breaks down as the discount factor approaches 1 but the converse is not found to be the case in our analysis. We find that low to 0 discount factor values produce stable results which high values and 1 gives very different results. This leads us to conclude that the Q learning agent performs better in our experiments being myopic rather than a long term thinker. We find that low discount factor values are scalable and stable even when the hypothesis space is large as shown by our big grid case.

The learning rate is found to have a smaller impact as we have shown in the deterministic case that varying the alpha or leaving it fixed leads to almost identical final Q matrices and the performance measures show a similar pattern. This is only the case when a low discount factor is used. Given the impact we have observed of high gamma values we do not expect the alpha variability will produce the same impact because the gamma value would lead to convergence issues for the agent. These are only applicable for the epsilon greedy case. For the softmax case we find that varying the temperature parameter by small amounts hardly leads to much difference and that by increasing the order of the magnitude of the temperature parameter we can produce steeper or gentler profiles for the softmax probabilities. A fixed value of 10 is found to be good value for our softmax policy.

For both types of policies considered the agent is able to arrive at the optimal policy in the deterministic case and learns very similar policies with stable gamma values even when the scope of the problem is greatly increased. Another insight is the fact that the Q Learning algorithm has nonlinear complexity. As we see that increasing the number of iterations by 5 leads to an 8 time increase in the run time. This is a very important consideration when considering very large problems.

Also, we propose an original way to visualise the performance measures by plotting a series of transforms of the performance measure shown in Fig 30. We plot a Hilbert transform which applies a linear operator and keeps the function in the same domain. It derives an analytical representation of the signal by extending it into the complex plane. This will help us identify potential periodicity in the data. This is useful because it is indicative of the agent potentially being stuck in a local minimum. The double Hilbert transform reverses the original signal and is perhaps less useful in this case. The Fourier transform gives the frequency spectrum of a given signal and shows dominant frequencies in the data. From this we can easily see if some values are more common and repeatedly learnt by the agent. This could highlight potential convergence of being stuck in a local minima. The Double Fourier Transform gives a spectrum of a spectrum and picks out values not immediately obvious from the first spectrum. The purpose of these are to add to our analysis toolkit and we include them here as potential extensions. The example presented is of the epsilon greedy agent with gamma = 0.2 the FFT and Hilbert plot both show that this agent achieves convergence fairly quickly and is stable from episode to episode without any spikes. We concluded this based on the Q matrix plots and rolling means previously but highlights the utility of these plots.

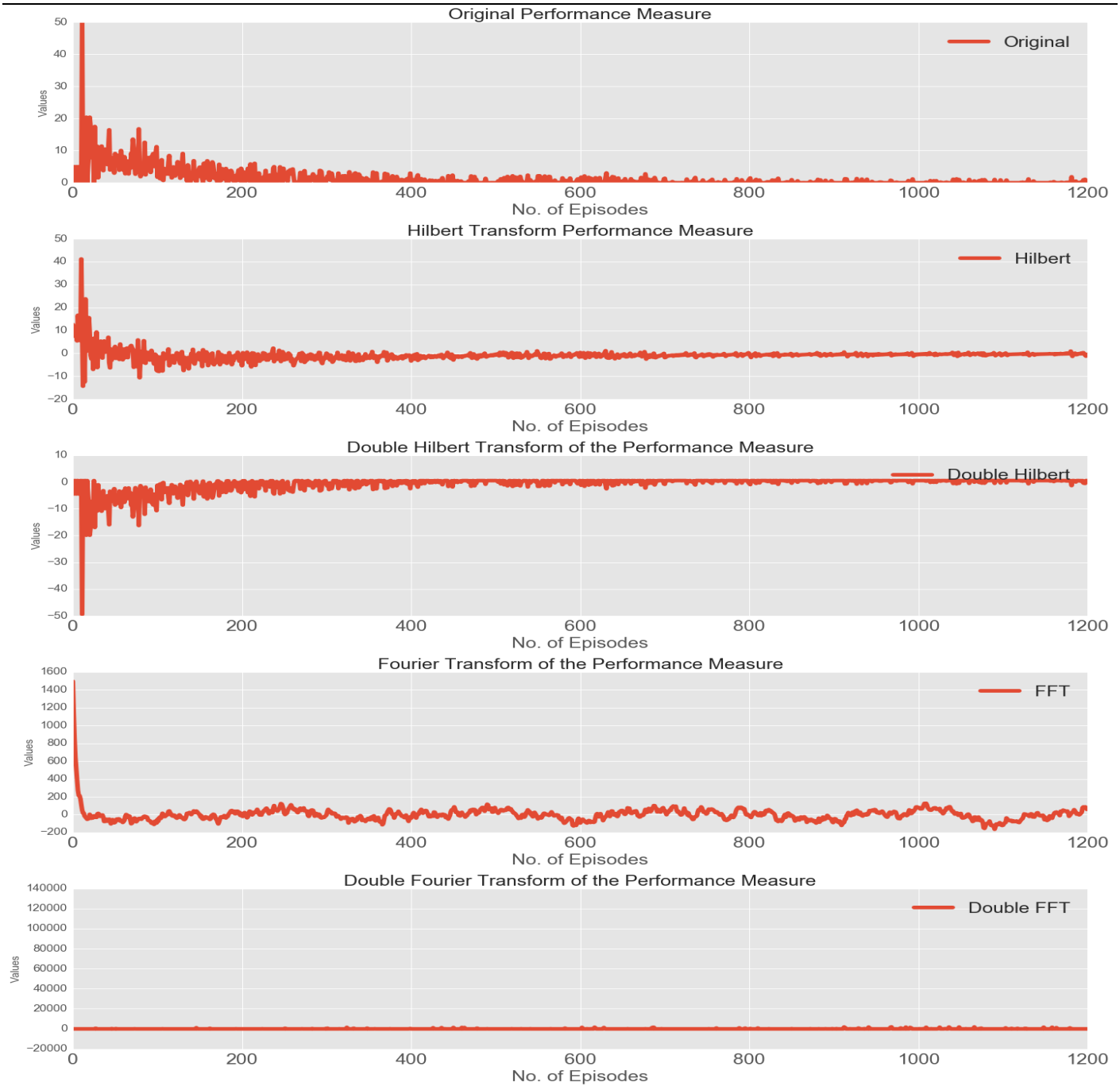

Figure 30: Some suggested alternative visualisations and analysis of performance measures.

FURTHER WORK

We suggest trying more creative approaches to defining policies and learning rate functions to assess their impact on the learning agent. In addition we suggest alternative visualisation of the performance measure such as by applying a Fourier Transform, Hilbert Transform, Double Hilbert and Fourier Transforms shown in Fig 30. In addition future work could use cosine similarity or other distance metrics between the performance measures to compare the performance of various Q learning parameter setups easily and consistently.

CONCLUSIONS

We have found that the Q learning algorithm is good choice for solving MDP problems and that it has a strong dependence on parameterisation for its final performance. The choice of policy is crucial as the epsilon greedy policies and discount factors have a strong coupling this is less the case with the softmax policy. This method is scalable to larger MDP's but its nonlinear complexity could inhibit its use for very large problems without suitable computation capacity. Overall, the Q Learning algorithm is a very good approach for solving MDP problems and lends itself to easy understanding and extension as we have demonstrated. However, care must be taken to choose a policy wisely to avoid unnecessary parameters dependence.

REFERENCES

[1] Richard S. Sutton and Andrew G. Barto, R. S. Sutton, and A. G. Barto, "Introduction to Reinforcement Learning," *Learning*, vol. 4, no. 1996, pp. 1—5, 2005.

[2] "Reinforcement Learning Introduction." [Online]. Available: http://reinforcementlearning.ai-depot.com/. [Accessed: 20-Mar-2016].

[3] F. Pérez and B. E. Granger, "IPython: a System for Interactive Scientific Computing," *Comput. Sci. Eng.*, vol. 9, no. 3, pp. 21—29, May 2007.

[4] S. van der Walt, S. C. Colbert, and G. Varoquaux, "The NumPy Array: A Structure for Efficient Numerical Computation," *Comput. Sci. Eng.*, vol. 13, no. 2, pp. 22—30, Mar. 2011.

[5] J. D. Hunter, "Matplotlib: A 2D graphics environment," *Comput. Sci. Eng.*, vol. 9, no. 3, pp. 90—95, 2007.

[6] "Seaborn: statistical data visualization – seaborn 0.7.0 documentation." [Online]. Available: https://stanford.edu/~mwaskom/software/seaborn/index.html. [Accessed: 20-Mar-2016].

[7] W. McKinney, "Data Structures for Statistical Computing in Python," in *Proceedings of the 9th Python in Science Conference*, 2010, pp. 51—56.

[8] I. Chadès, G. Chapron, M.-J. Cros, F. Garcia, and R. Sabbadin, "MDPtoolbox: a multi-platform toolbox to solve stochastic dynamic programming problems," *Ecography (Cop.).*, vol. 37, no. 9, pp. 916—920, Sep. 2014.