INDIAN INSTITUTE OF TECHNOLOGY,
MADRAS

REINFORCEMENT LEARNING PROGRAMMING
ASSIGNMENT 1

CS6700

Multi-Armed Bandits

Author
DODIYA KEVAL

Roll Number
CS19M023

February 4, 2020

1 Problem 1

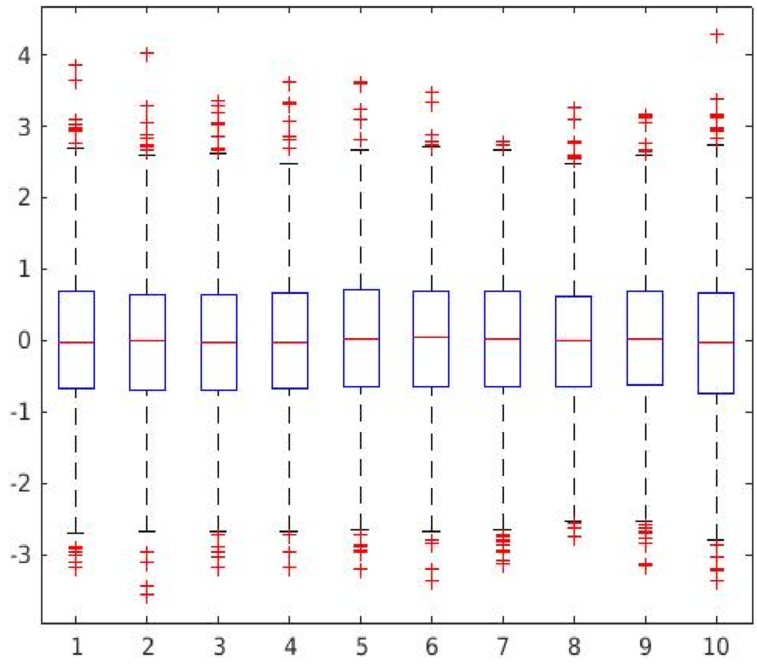

Figure 1: Testbed ,Xlabel: number of bandits, Ylabel :reward

figure .1 indicates testbed of 10 different arms. those arms are sampled from normal distribution with $\mu = 0$ and $\sigma^2 = 1$

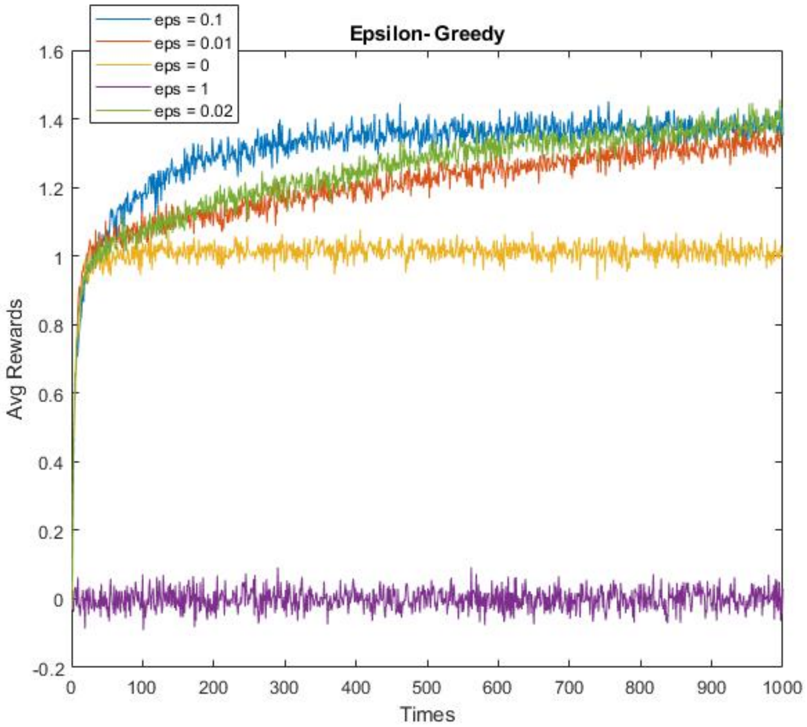

Figure 2: Average reward Vs Time for ϵ- greedy with different value of ϵ for 2000 bandits problem

By observing fig 2 we can say that when epsilon increases regret got decreases. so with higher epsilon converges(eps = 0.1) become faster.and with ϵ =0 which means fully greedy(no exploration) algorithm exploits only one arm and hence its avg reward is stuck at zero most of the time.

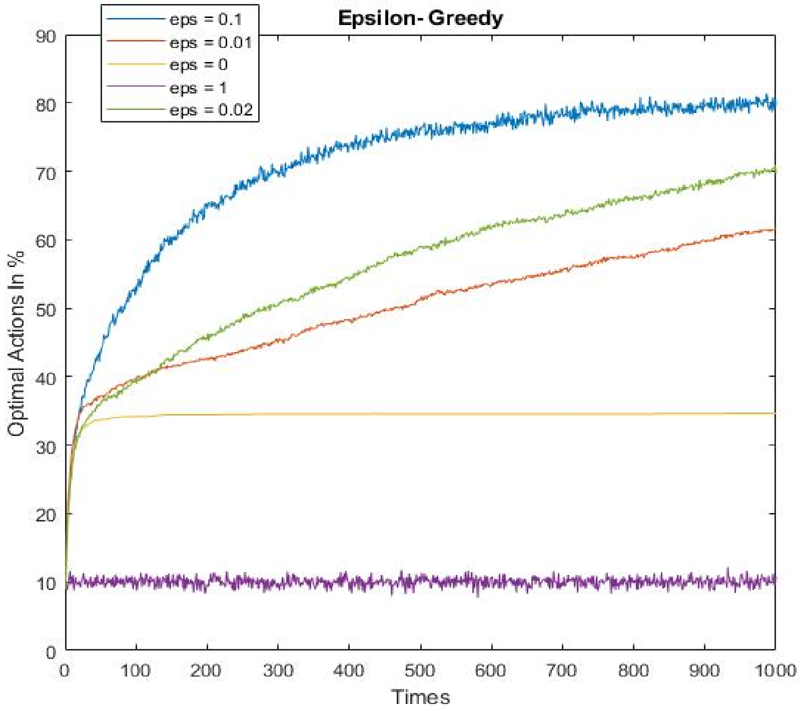

Figure 3: optimal Action Percentage Vs Time for ϵ- greedy with different value of ϵ for 2000 bandits problem

one can infer from figure 3 that with higher value of ϵ the amount of exploration of all arms increases so with 0.1(ϵ) optimal action percentage become around 80% whereas with fully greedy algo. (ϵ = 0) optimal action stuck at one arm gives least amount of exploitation of optimal arm.

2 Problem 2

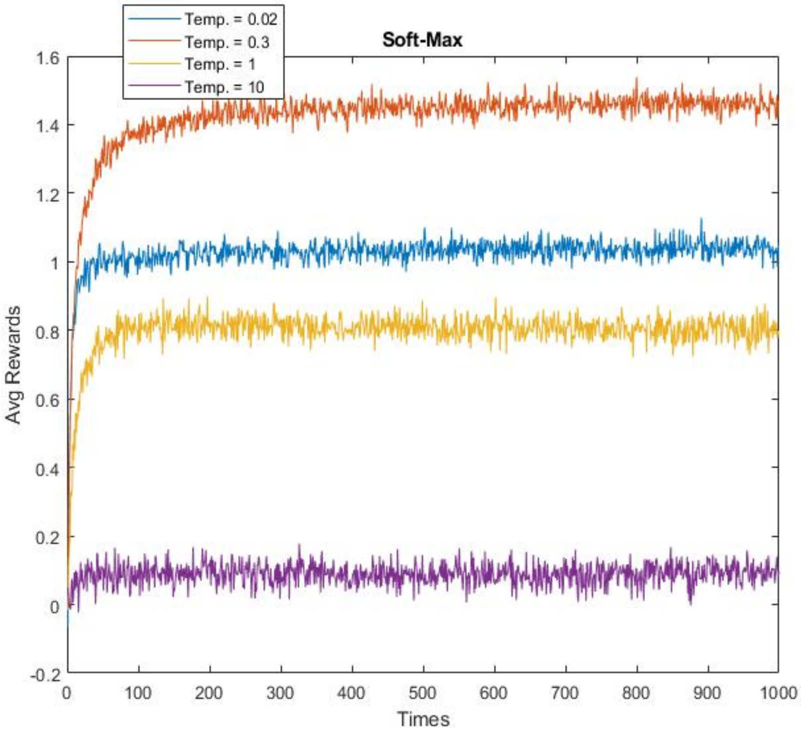

Figure 4: Average reward Vs Time for SoftMax with different value of Temperature for 2000 bandits problem.

In the softmax sampling method as temperature increases the probability of all arms becomes nearly equals so the rate of selection of optimal arm decreases implies avg reward is very low for temp = 10. and for smaller value of temp. method acts as greedy hence avg reward becomes around 1.4.

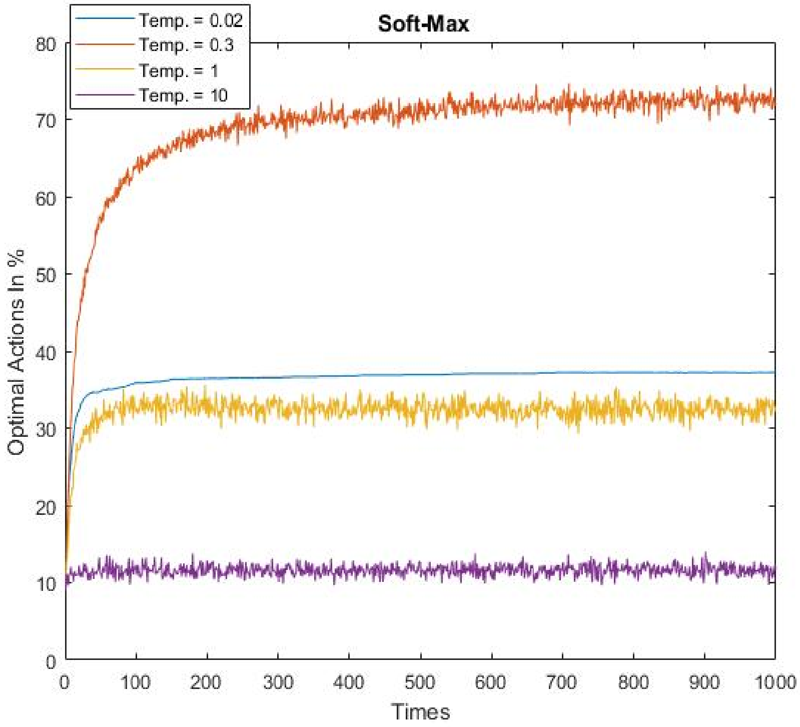

Figure 5: Optimal actions in Percentage Vs Time for SoftMax with different value of Temperature for 2000 bandits problem.

As mentioned above with higher value of temperation probability of all arms become near to equal so due to randomness optimal arms selection rate decreases.

3 Problem 3

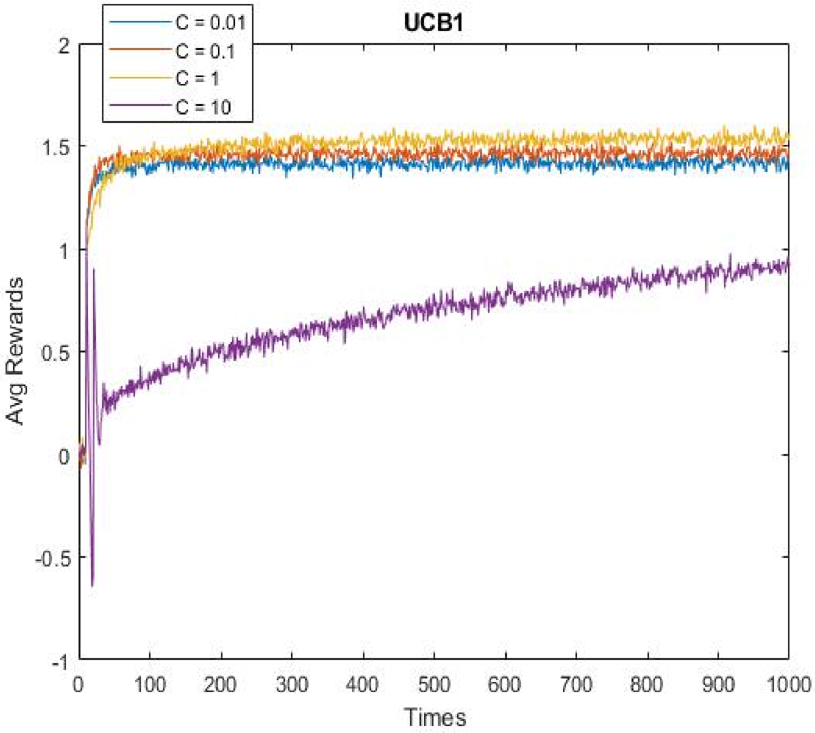

Figure 6: Average reward Vs Time for UCB1 with different Learning Rates(C) for 2000 bandits problem.

- For UCB1 method 'C' is the Learning Rate or certainty of estimation of mean. so if 'C' gets higher confidence interval wil get larger implies less certainty about the estimate.
- therefore this method selects suboptimal arm most of the time hence avg reward lower.
- same for lower value of 'C' confidence interval less implies selection rate of optimal arm is higher so avg reward become higher
- And By figure 6 conclusion will be at C =1 method performs best.

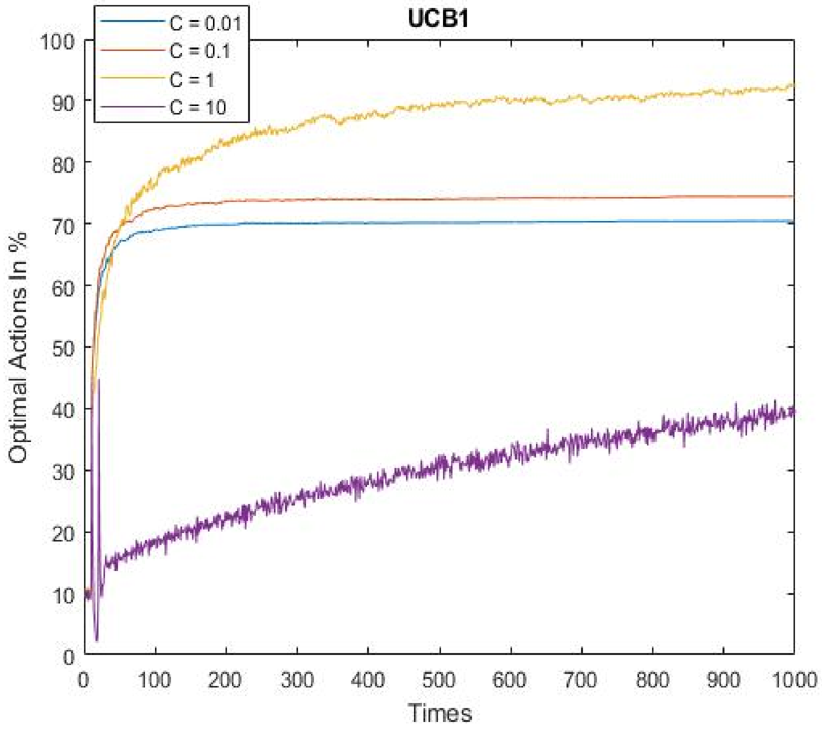

Figure 7: Optimal actions in percentage Vs Time for UCB1 with different Learning Rates(C) for 2000 bandits problem.

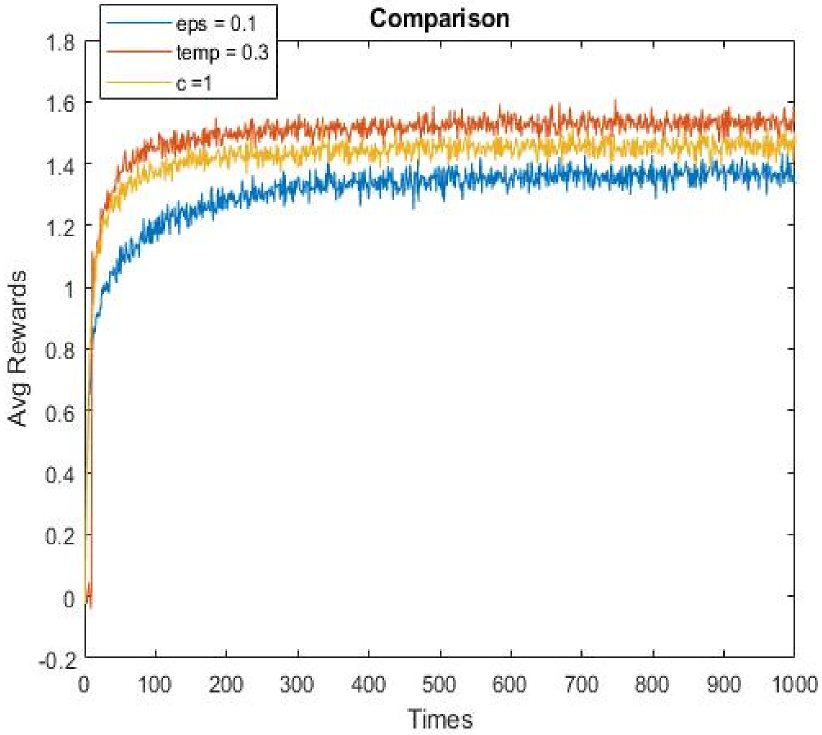

Figure 8: Average reward Vs Times comparison among ϵ-Greedy, SoftMax, UCB1 method with different parameters

- here ϵ = 0.1, temp = 0.3 and c = 1, UCB1 and softmax perform around same,while epsilon-greedy gives less reward and higher regret
- so in conclusion UCB1 and softmax is the better compares to epsilon-greedy.

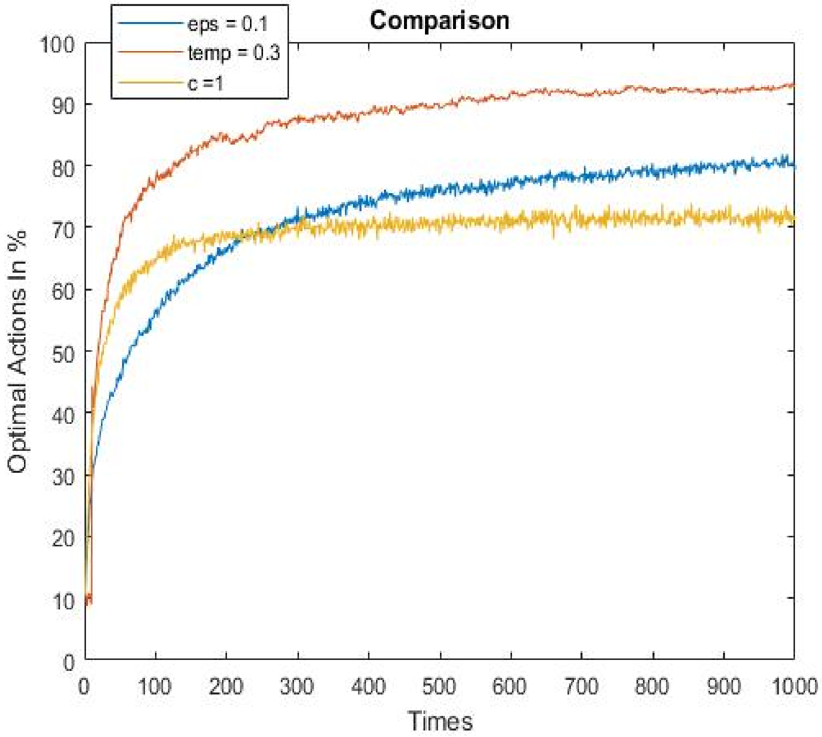

Figure 9: Optimal Actions in % Vs Times comparison among ϵ-Greedy, SoftMax, UCB1 method with different parameters

4 Problem 4

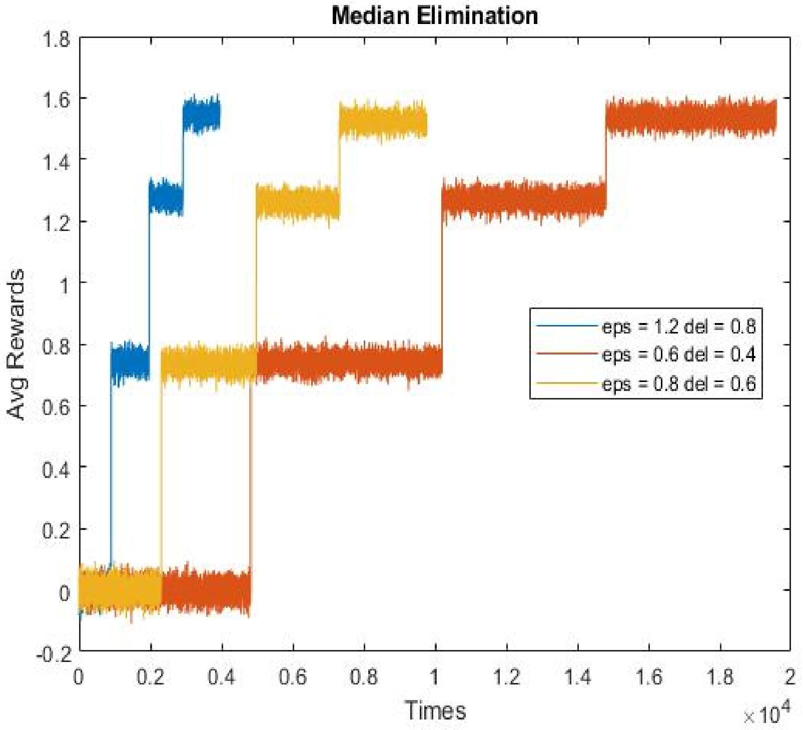

Figure 10: Average reward Vs Times for MEA method with different ϵ and δ for 2000 bandits problem.

- here number of samples required to be taken for each arm is square of $\frac{2}{\epsilon^2}log(\frac{1}{\delta})$, so for higher values of ϵ and δ, Median Elimination converges faster.we can infer this by above figure.
- The time required to compute the median of k arms of a single bandit problem is O(k*log k + O(1)) by sorting and finding. and in performance MEA takes time around 3 - 6 seconds that is lesser than other algortihm so median calculation is not the rate determining step.
- While, comparing Median Elimination with ϵ-greedy, softmax and UCB, the convergence of median elimination is slow(large regret) . but in optimal action in % median elimination is better than softmax, $epsilon$-greedy and around same with UCB1.

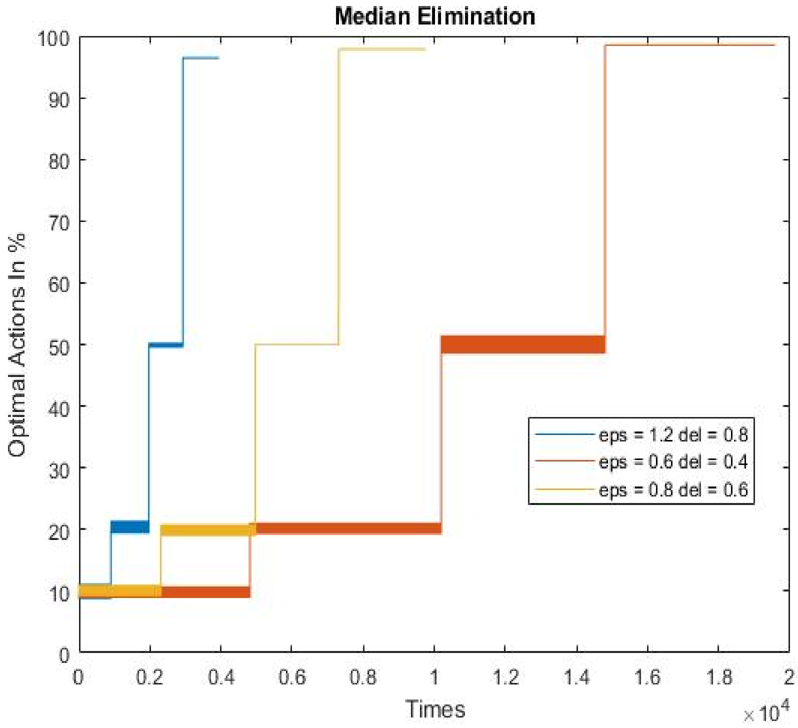

Figure 11: Optimal Actions Vs Times for MEA method with different ϵ and δ for 2000 bandits problem

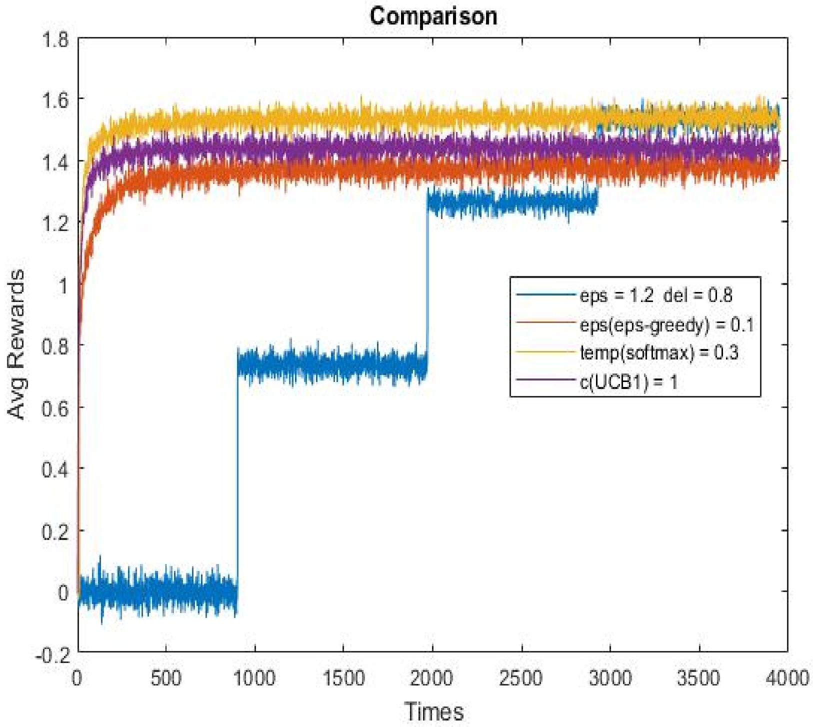

Figure 12: Average reward Vs Times comparison among ϵ-Greedy, SoftMax, UCB1,MEA method with different parameters.

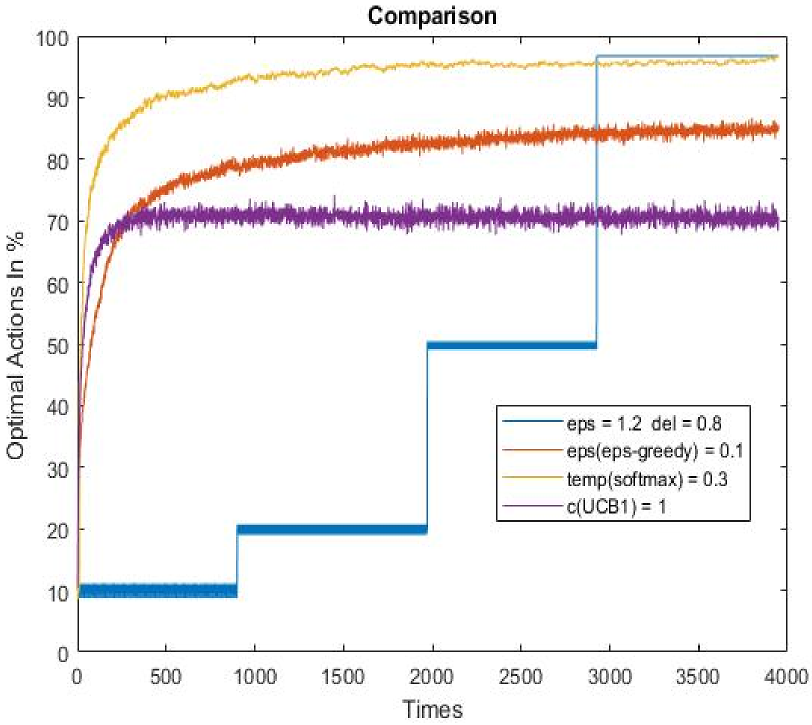

Figure 13: Optimal actions in % Vs Times comparison among ϵ-Greedy, SoftMax, UCB1,MEA method with different parameters.

5 Problem 5

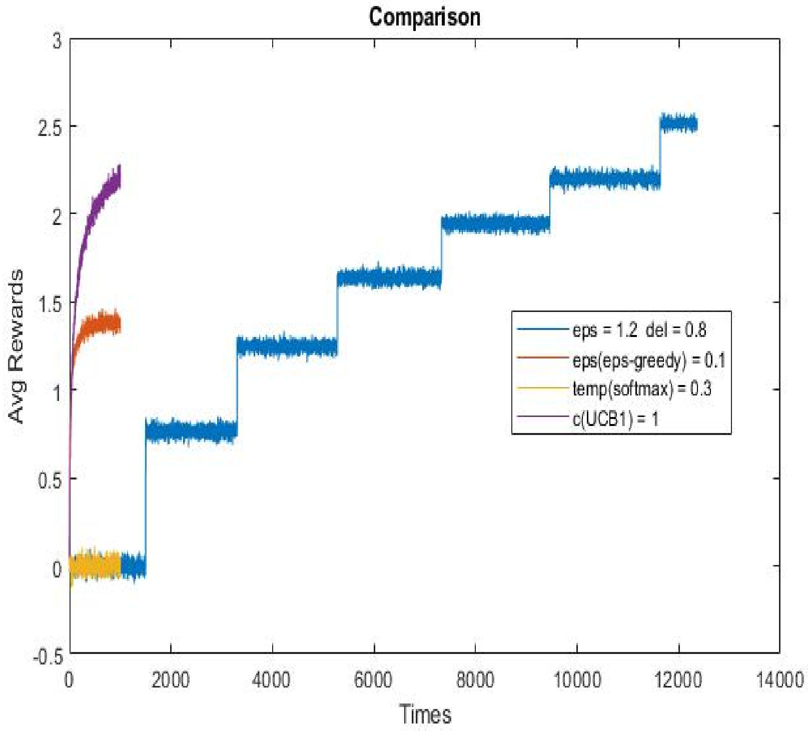

Figure 14: Average reward Vs Times comparison among ϵ-Greedy, SoftMax, UCB1,MEA method with different parameters and 1000 arms ,2000 bandits problem.

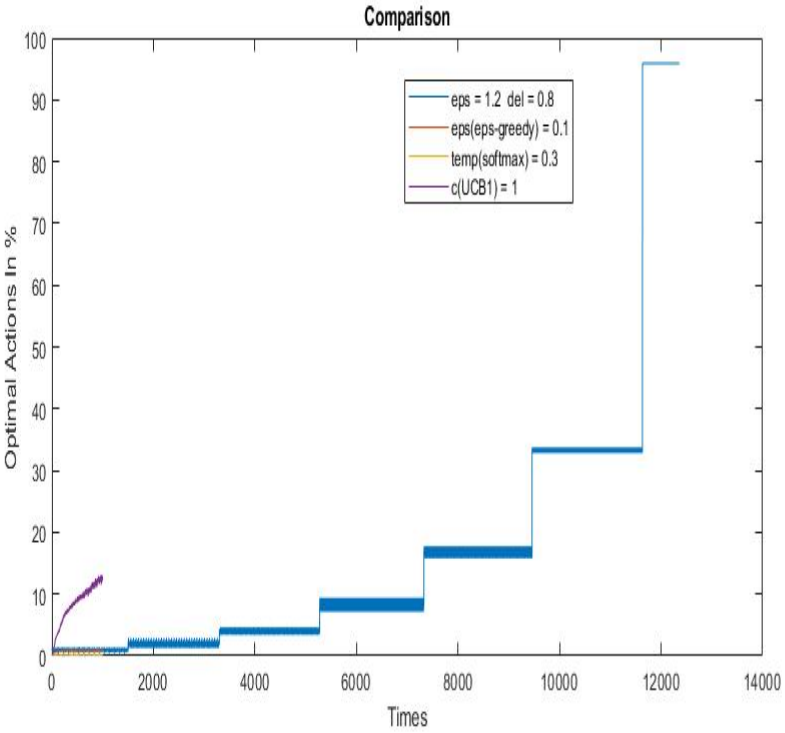

Figure 15: Optimal Actions in % Vs Times comparison among ϵ-Greedy, SoftMax, UCB1,MEA method with different parameters and 1000 arms ,2000 bandits problem..

- here with 1000 arms MEA takes 10^5 order steps .so uploaded the 100 arms MEA and compares with different algorithms.
- here by above figures 14,15 with large number of arms convergence time of MEA increases. we can check the performance with 2000 steps.