Homework-1

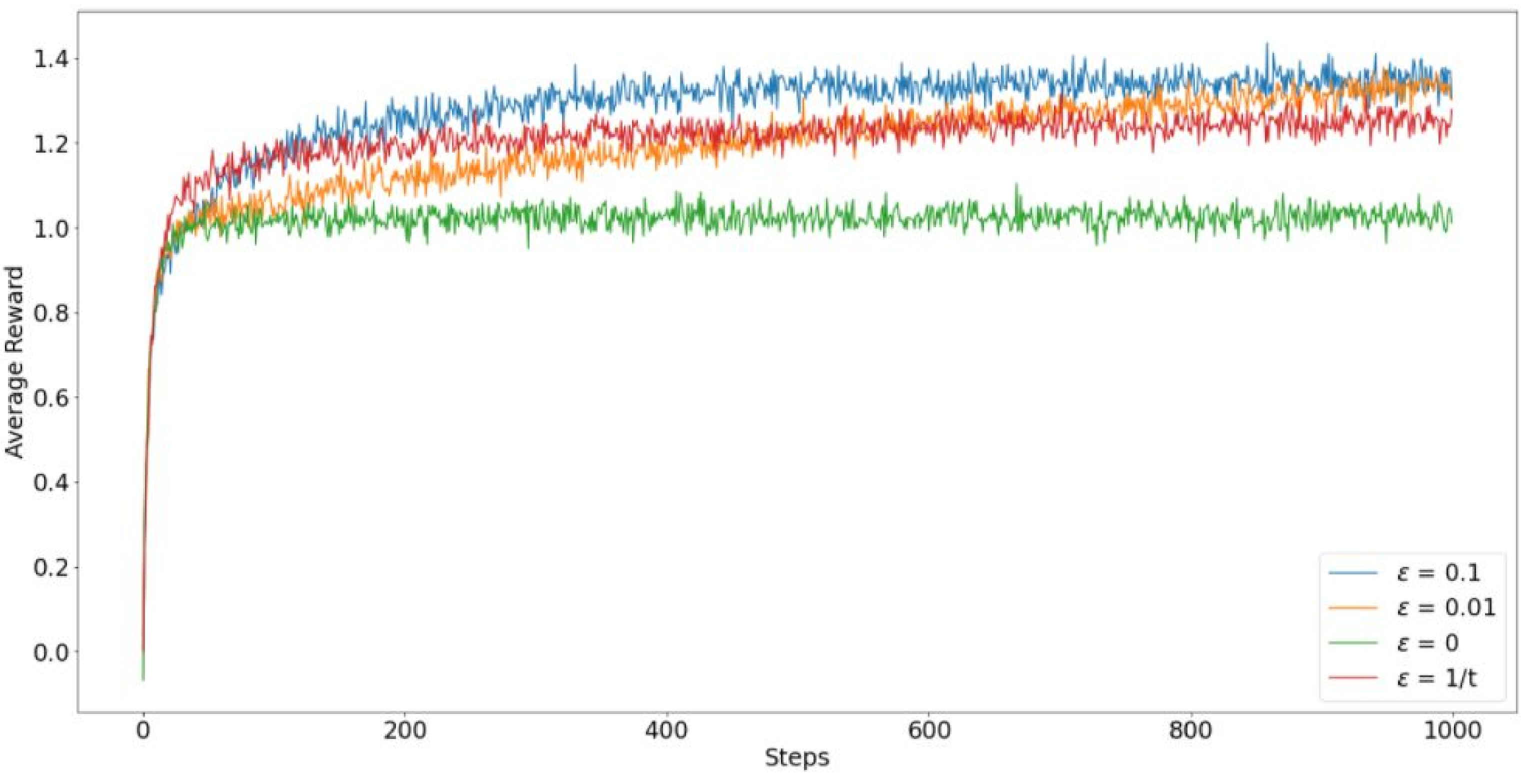
Question 1:

The 10-armed testbed with $q^*(a)$, a=1,2,.......,10 selected according to a normal distribution with mean 0 and variance 1.

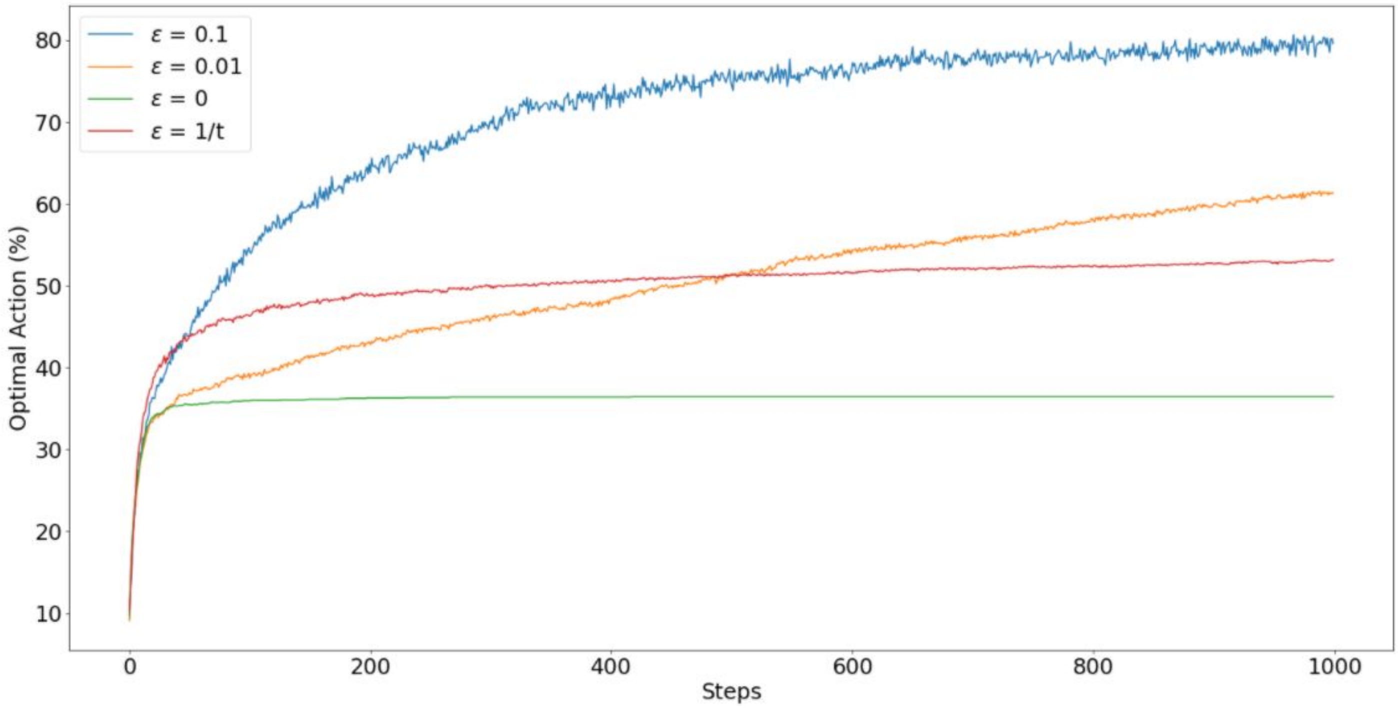
Rewards for each arm are selected using normal distribution with mean $q^*(A_t)$ and variance 1.

Graph drawn for ε = 0.1, 0.01, 0 and 1/t.

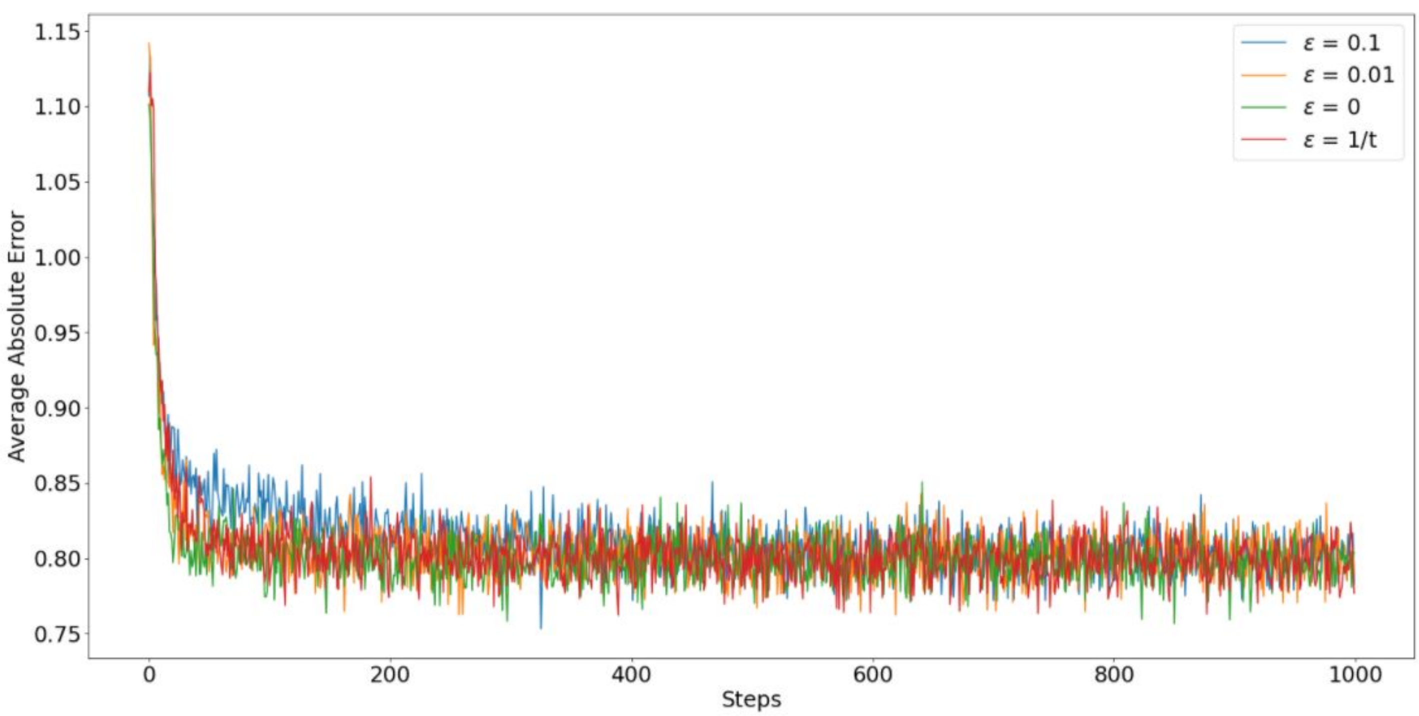
Time steps=1000, Number of simulations=2000



Average Reward



Optimal Action (%)



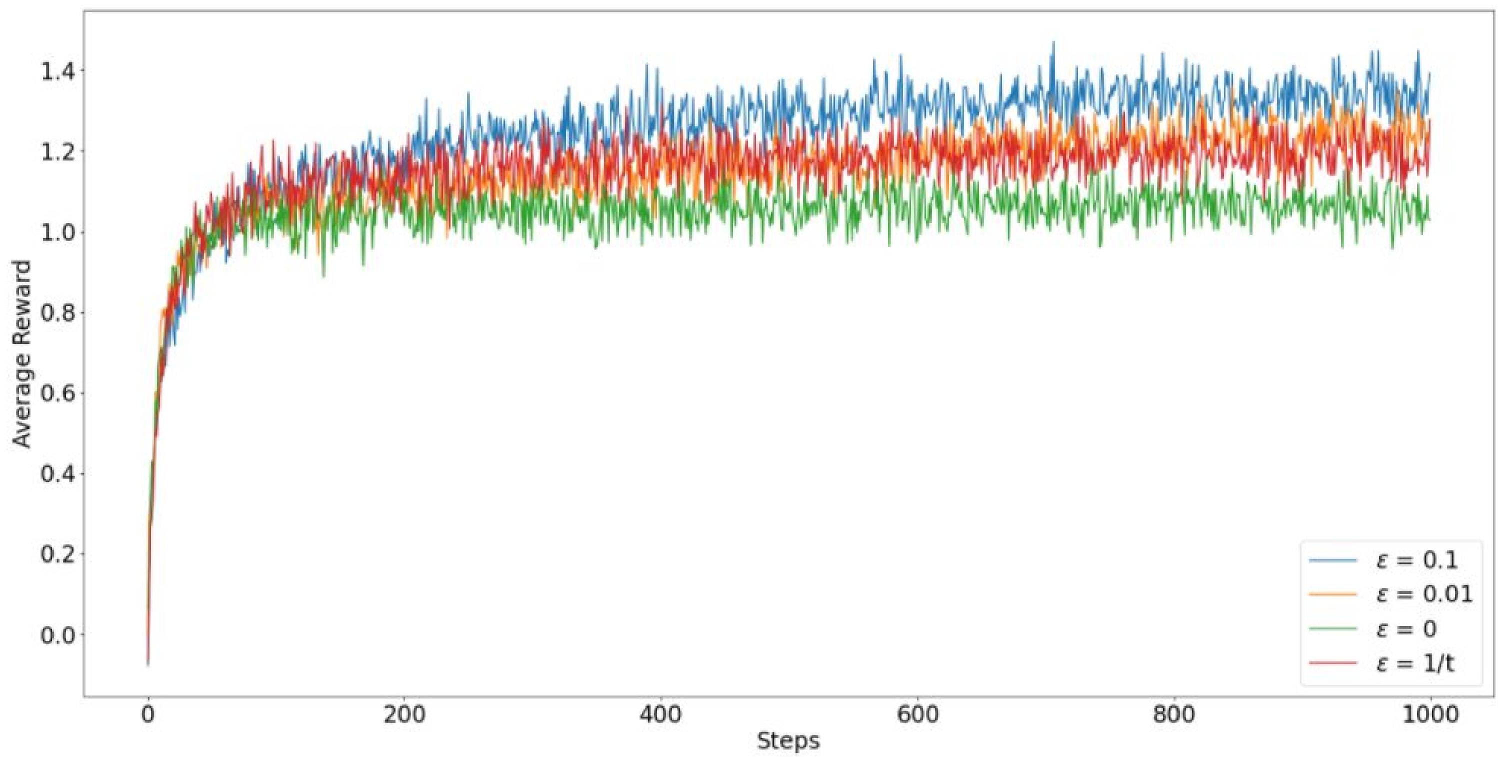
Average Absolute Error

Result: For time steps=1000, ϵ = 0.1 gives the better result but this is not true for long runs as then ε = 0.01 gives the better result as it picks up optimal action throughout the simulation with greater probability as compared to ε = 0.1.

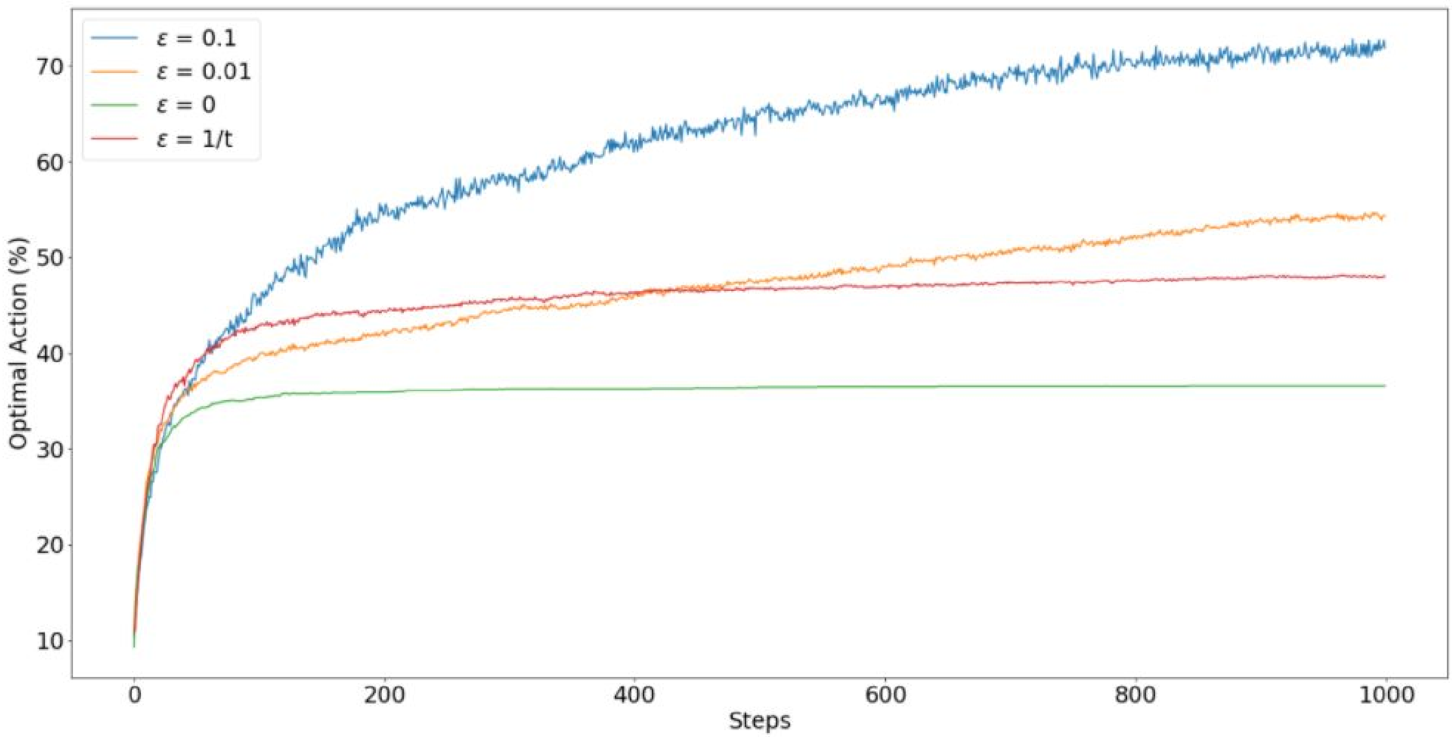
Also, for long runs, ε = 1/t gives us the best result as it increases its probability to exploit as t increases thus motivating exploration early on to find optimal estimate and then exploiting once it has been found (as $t \rightarrow \infty$, $\varepsilon \rightarrow 0$)

Question 2 :

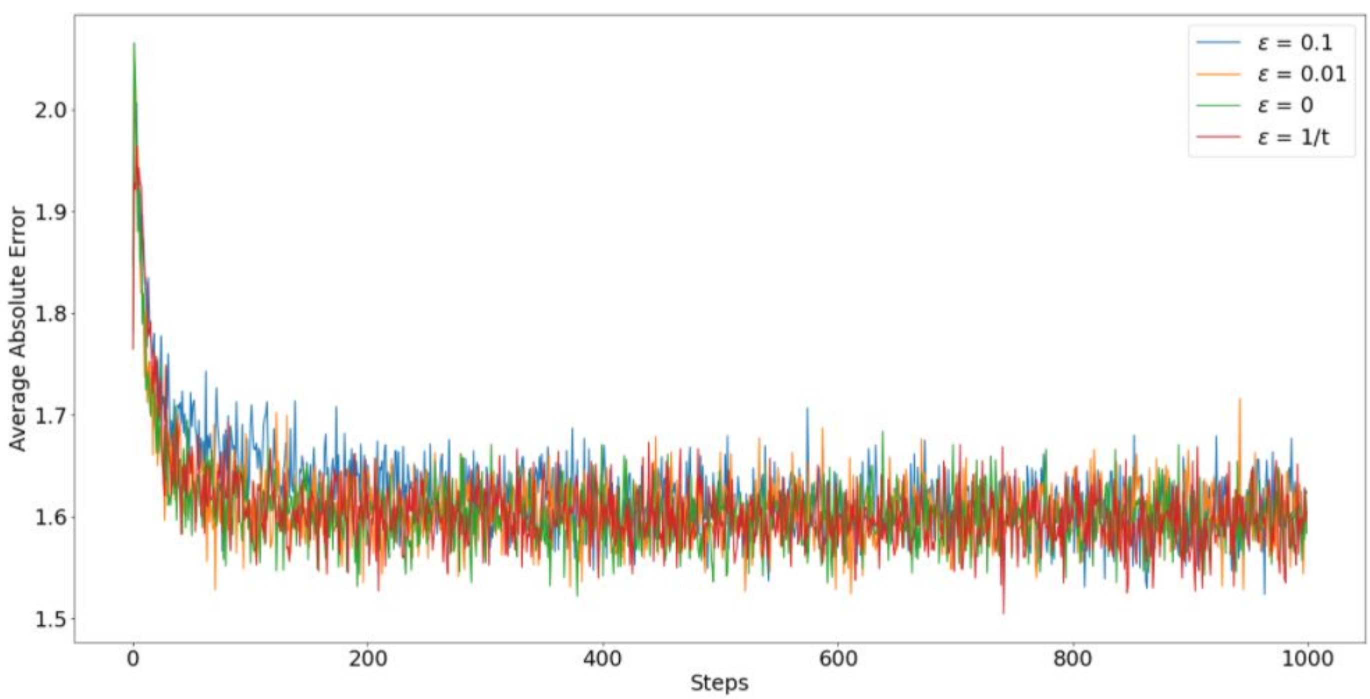
Same as Question1 but now rewards for each arm are selected according to a normal distribution with mean $q^*(A_t)$ and variance 4.



Average Reward



Optimal Action (%)



Average Absolute Error

Result: The same trend follows as in Question1 for different values of ε. But, now the rewards and estimates are much noisier and variance= 4 also negatively effects the optimal action for time steps=1000 (can be seen through the range of values on the y-axis)

Page No.
Date

Question 3

In the long run, $Q_t(a)$ for $\varepsilon = 0.01, 0.1$ and $\frac{1}{t}$ would converge to become $Q^*(a)$. (following sample-average method).

- This is not the case for $\varepsilon = 0$ as it always picks the greedy estimate hence it performs the poorest.

- $\varepsilon = 0.01$ in the long run performs better than $\varepsilon = 0.1$ as:

- for $\varepsilon = 0.01$

$$P[\text{picking greedy action}] = 1 - \varepsilon + \frac{\varepsilon}{|A|}$$

$$= 0.991$$

- thus it exploits and picks up the optimal arm with more probability when compared to $\varepsilon = 0.1$ as $P[\text{picking greedy action} \mid \varepsilon = 0.1] = 0.91$

- ~~This is~~ It may be possible that initially $\varepsilon = 0.01$ takes more time to explore when compared to $\varepsilon = 0.1$ case but since $t \to \infty$, ~~thus it can be said~~ so now $\varepsilon = 0.01$ would come up with a better estimate which it now chooses with greater probability.

★ For $\varepsilon = \frac{1}{t}$,

Our probability for exploring is initially high and reduces over time. This help in knowing a better estimate of Q. and as t increases, we shift our focus to exploitation as till then a pretty good estimate of Q_t that is closest to Q^*_t would be found.

Also as $t \to \infty$, $\varepsilon = 0$ and thus when Q^*_t is approached then we always exploit.

★ Thus, overall $\varepsilon = \frac{1}{t}$, will perform best in the long run. (may not be the case for small time steps)

Page No. ____
Date

Q Question- 4

→ Sample mean

- It is not influenced by initial choice of $Q_1(a)$ as soon as all the arms get selected atleast once. from then onwards, $Q_1(a)$ has no effect on Q_n.

$$\Rightarrow \quad Q_n = \frac{1}{n-1} \sum_{i=1}^{n-1} R_i$$

$$= \frac{1}{n-1}\left(R_{n-1} + \sum_{i=1}^{n-2} R_i \right)$$

$$= \frac{1}{n-1}\left(R_{n-1} + \frac{(n-2)}{(n-2)} \sum_{i=1}^{n-2} R_i \right)$$

$$\Rightarrow \frac{1}{n-1}\left(R_{n-1} + (n-2)\, Q_{n-1} \right)$$

$$\boxed{\text{as } \quad Q_{n-1} = \frac{\sum_{i=1}^{n-2} R_i}{n-2}}$$

$$\Rightarrow \frac{1}{n-1}\left(R_{n-1} + (n-1) Q_{n-1} - Q_{n-1} \right)$$

$$\boxed{Q_n = Q_{n-1} + \frac{1}{(n-1)}\left(R_{n-1} - Q_{n-1} \right)}$$

- No dependence on Q_1 can be seen for $\forall n$ apart from $n=2$.

Page No.
Date

→ Constant step-size α

This method aims at weighing past rewards with decreasing weight as t increases. thus even though t becomes very large, there is some dependence seen on $Q_1(a)$.

This can be shown as:

$$Q_{n+1} = Q_n + \alpha [R_n - Q_n]$$

$$= \alpha R_n + (1-\alpha) Q_n$$

$$= \alpha R_n + (1-\alpha) [\alpha R_{n-1} + (1-\alpha) Q_{n-1}]$$

$$= \alpha R_n + (1-\alpha) \alpha R_{n-1} + (1-\alpha)^2 Q_{n-1}$$

$$= \alpha R_n + (1-\alpha) \alpha R_{n-1} + (1-\alpha)^2 \alpha R_{n-2}$$

$$+ \cdots + (1-\alpha)^{n-1} \alpha R_1 + (1-\alpha)^n Q_1$$

$$\boxed{Q_{n+1} = (1-\alpha)^n Q_1 + \sum_{i=1}^{n} \alpha (1-\alpha)^{n-i} R_i}$$

- Here, clear dependence on Q_1 can be seen with weight $(1-\alpha)^n$.

Also, if α is smaller than ~~too~~ though $(1-\alpha)^n$ continues to decrease over

Page No.
Date

time, when compared to case where α is larger, it is still larger.

$$(1-\alpha_1)^n > (1-\alpha_2)^n$$

if $\alpha_1 < \alpha_2$

→ we can propose our step size to be:

$$\alpha_{new} = \frac{\alpha_{old}}{\bar{\theta}_n}$$

wherein $\bar{\theta}_n = \bar{\theta}_{n-1} + \alpha_{old}(1 - \bar{\theta}_{n-1})$

for all $n \geq 0$
and $\bar{\theta}_0 = 0$

* Now to show independence from $Q_1(a)$:

Now for Q_1,

$$\alpha_{new} = \frac{\alpha_{old}}{\bar{\theta}_1}$$

and $\bar{\theta}_1 = 0 + \alpha_{old}(1-0)$ $\boxed{\bar{\theta}_0 = 0}$

$\bar{\theta}_1 = \alpha_{old}$

$$\boxed{\alpha_{new} = 1}$$

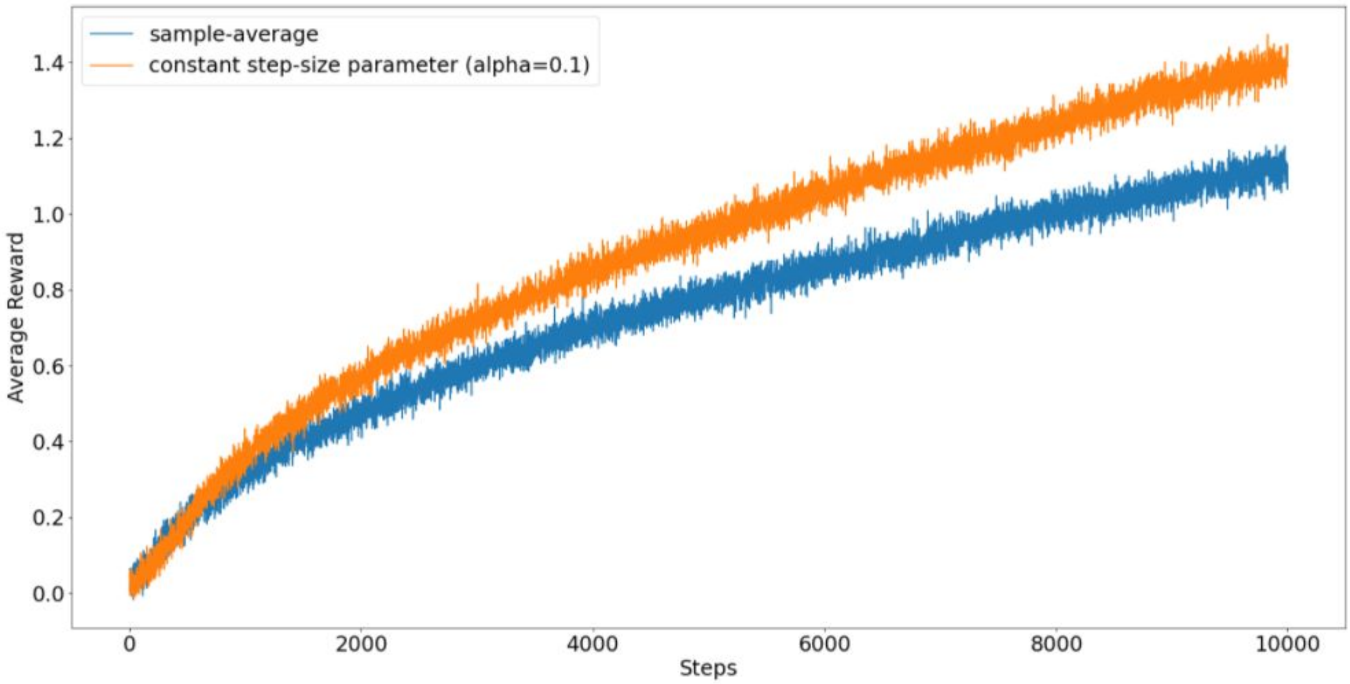
Page No.
Date

and:

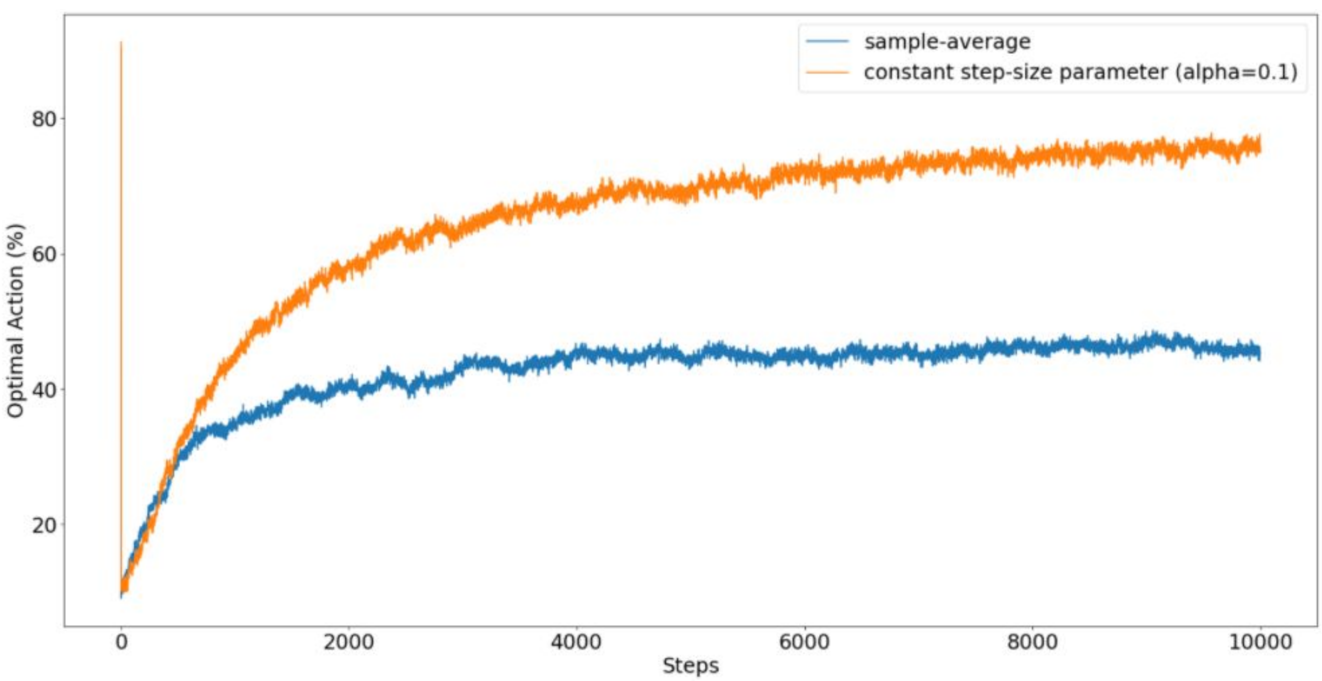
$$\boxed{Q_{n+1} = \cancel{(1-1)^n Q_1}^{\,0} + \sum_{i=1}^{n} \left(\frac{\alpha_{old}}{\bar{O}_n}\right)\left(1-\left(\frac{\alpha_{old}}{\bar{O}_n}\right)\right)^{n-i} R_i}$$

Thus it can be clearly seen that for $\alpha_{new} = \frac{\alpha_{old}}{\bar{O}_n}$, Q_{n+1} does not depend on $Q_1(a)$.

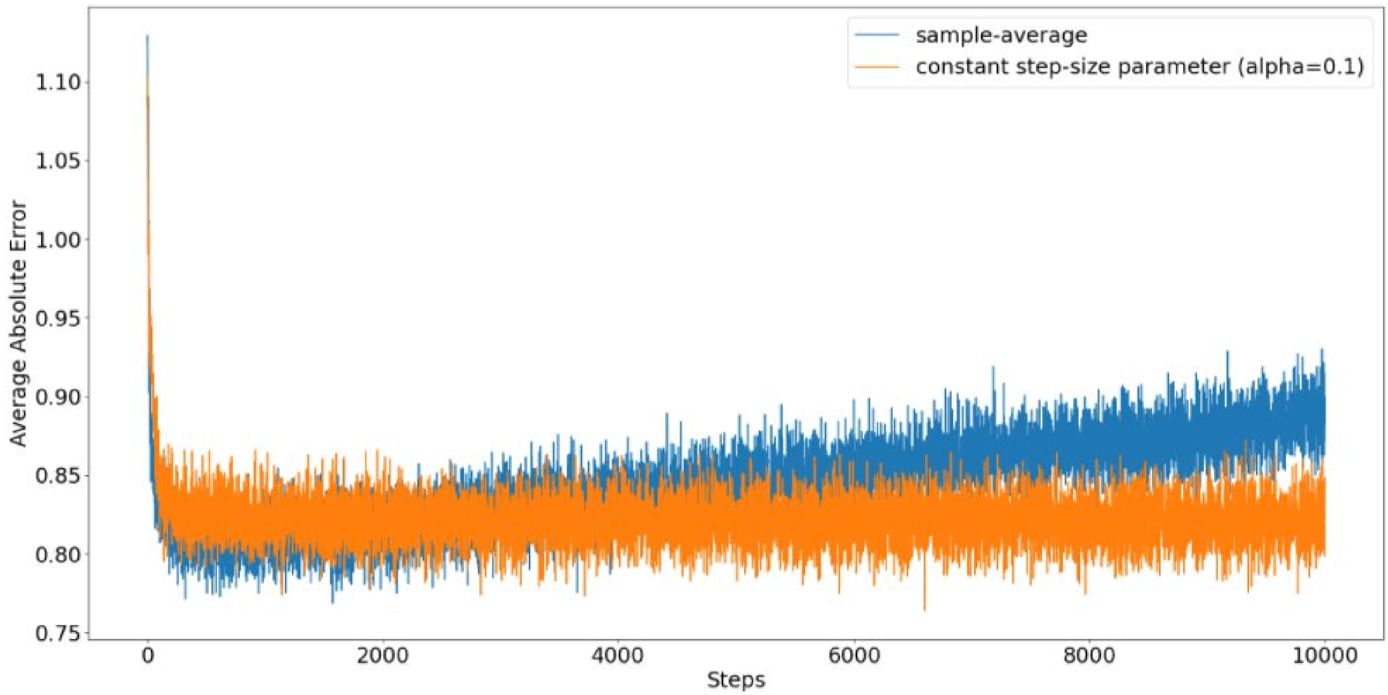
Question 5



Average Reward



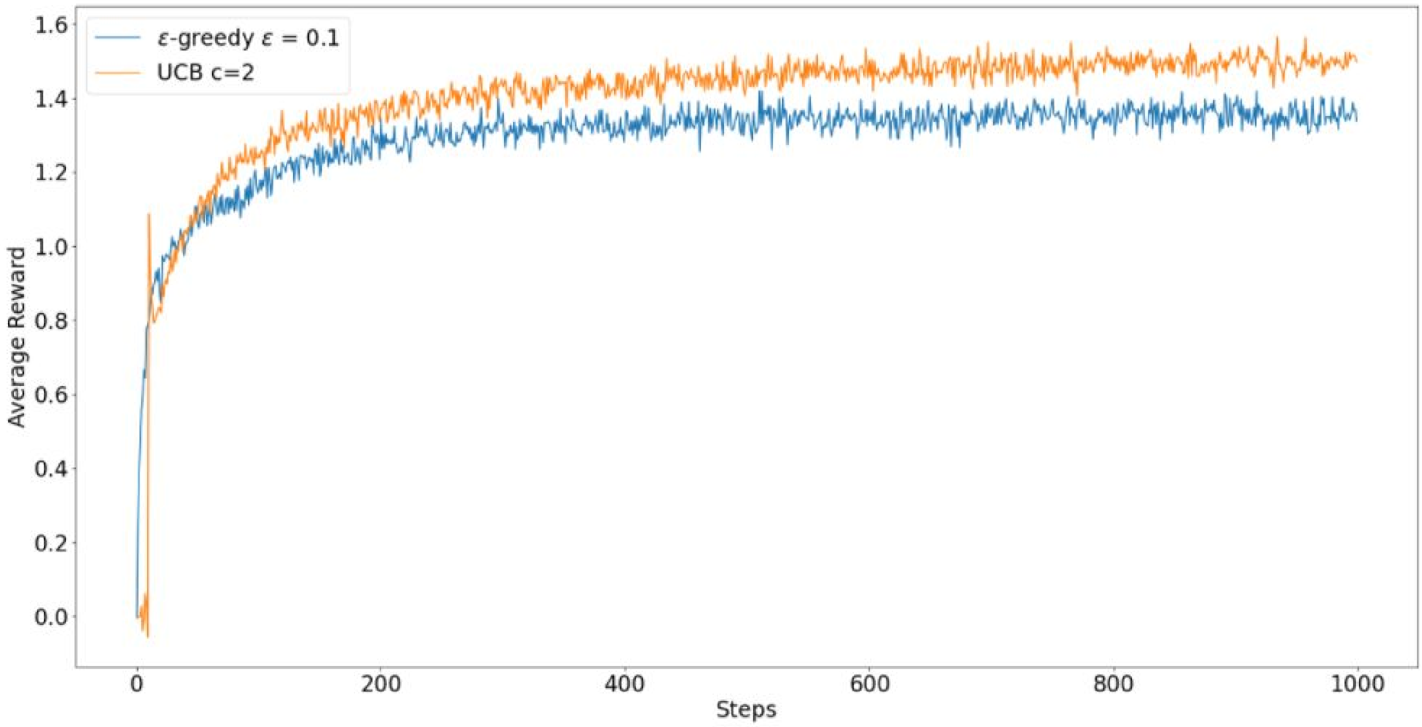
Optimal Action (%)



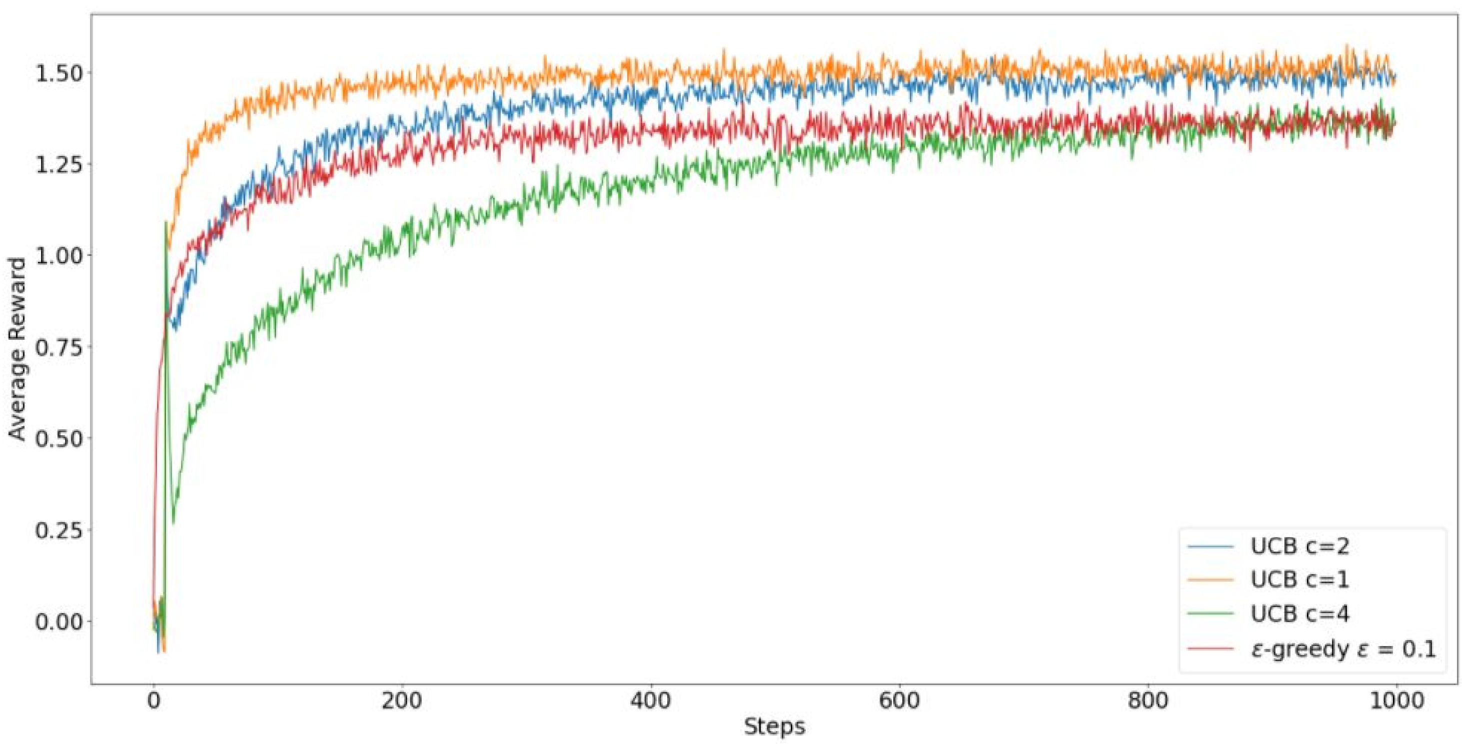
Average Absolute Error

Result: For non-stationary setting, constant step-size parameter tends to give better results as they, unlike sample-average, do not converge and thus readjust themselves according to any change in the distribution of the arms.

Question 6:



ε-greedy ε = 0.1
UCB c=2
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Average Reward
0
200
400
600
800
1000
Steps

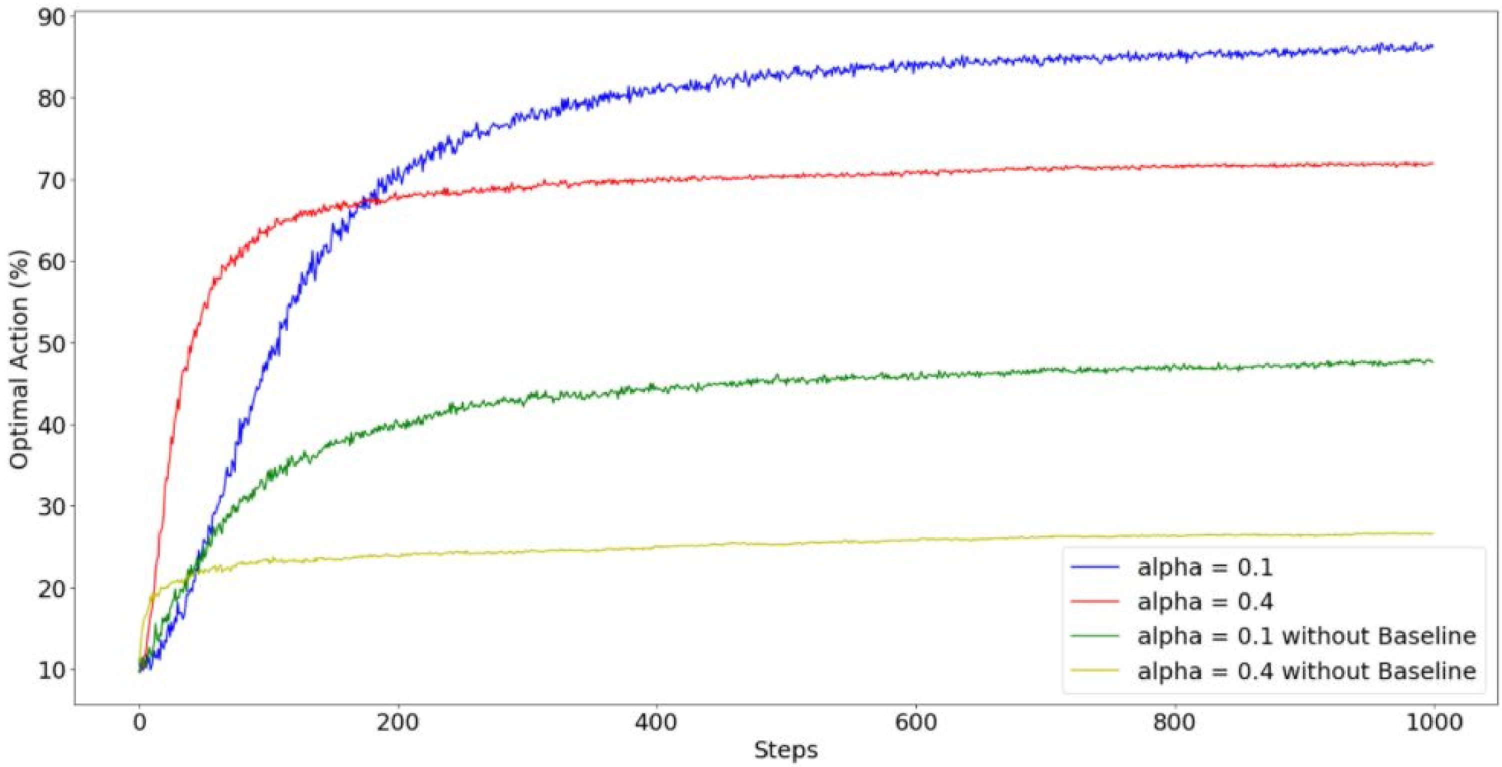


1.50
1.25
1.00
0.75
0.50
0.25
0.00
Average Reward
0
200
400
600
800
1000
Steps
UCB c=2
UCB c=1
UCB c=4
ε-greedy ε = 0.1

Result: According to UCB, when $N_t(a)=0$ for any arms A_t, then that arm is considered to be the maximizing action.
Now, this occurs for all the 10 arms initially in random order. Thus, we see the low average reward for the first 10 steps as then we are exploring and picking a different arm each step.

Now for the 11th step, the term $c * \sqrt{lnt/Nt(a)}$ is the same irrespective of the picked arm. Thus, now the discriminatory factor will only be the value of Q_t which is dependent on the first round of exploration (affected by the reward of each arm and Q_1). Now, that action is picked whose Q_t is greater and this would be seen across all 2000 simulations, thus a spike in average reward can be seen for the 11th step. Further onwards $c * \sqrt{lnt/Nt(a)}$ is different for each arm and thus average rewards progress steadily as the model acquires more certainty about the system.

Question 7:



Result: The optimal action is better for α = 0.1 and with Baseline rather than for without baseline. It is because the Baseline is used to capture the deviation from the mean rather than the absolute value.

Both lower α and Baseline lead to lower variance changes thus performing better.