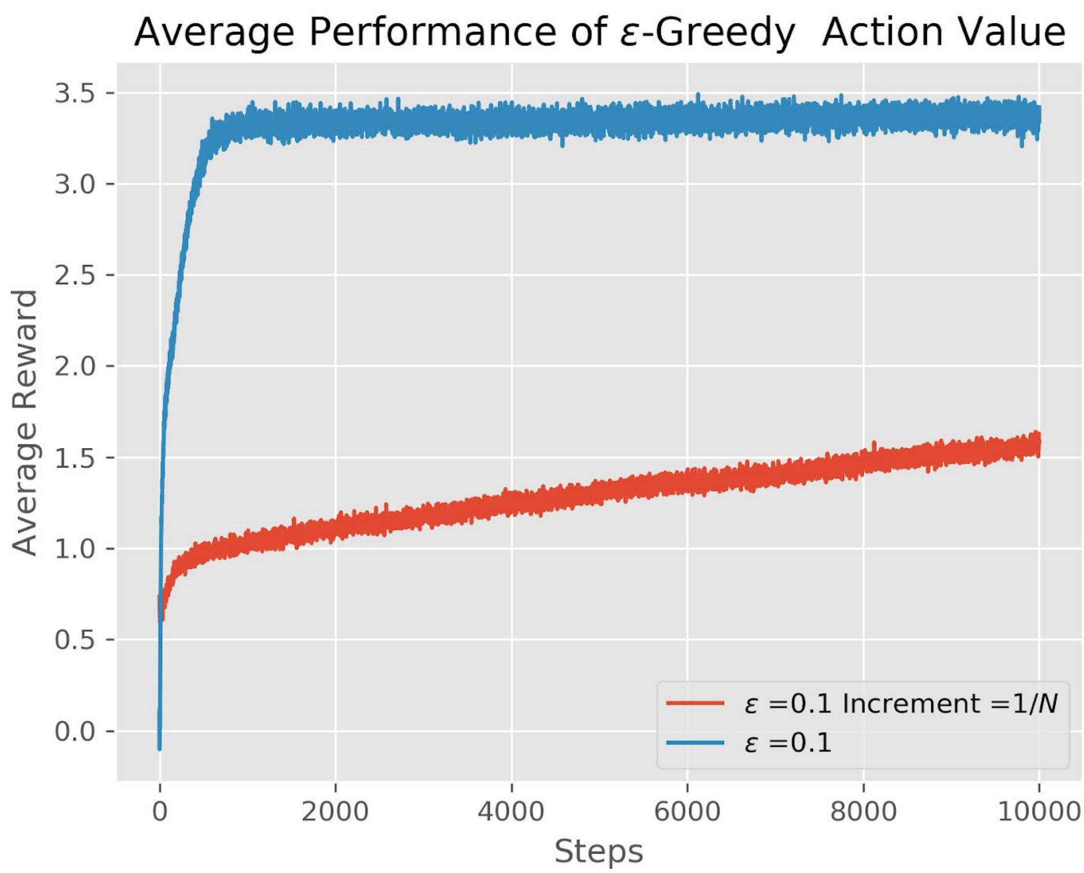
RL - Assignment -1

(Ashish Sethi-MT18024)

Table of Contents

Exercise - 2.5

Average Performance of ε-greedy action value on the ten-armed testbed



Optimal Actions of ε-greedy action value on the ten-armed testbed

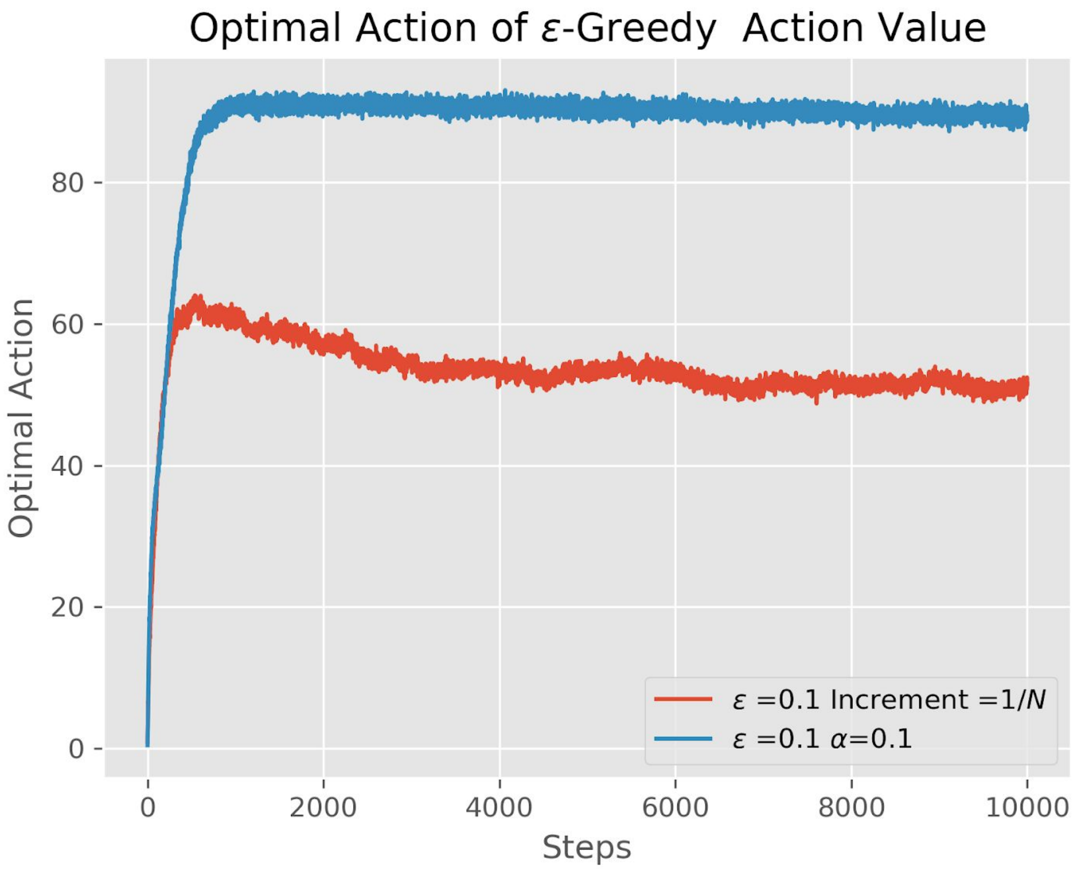
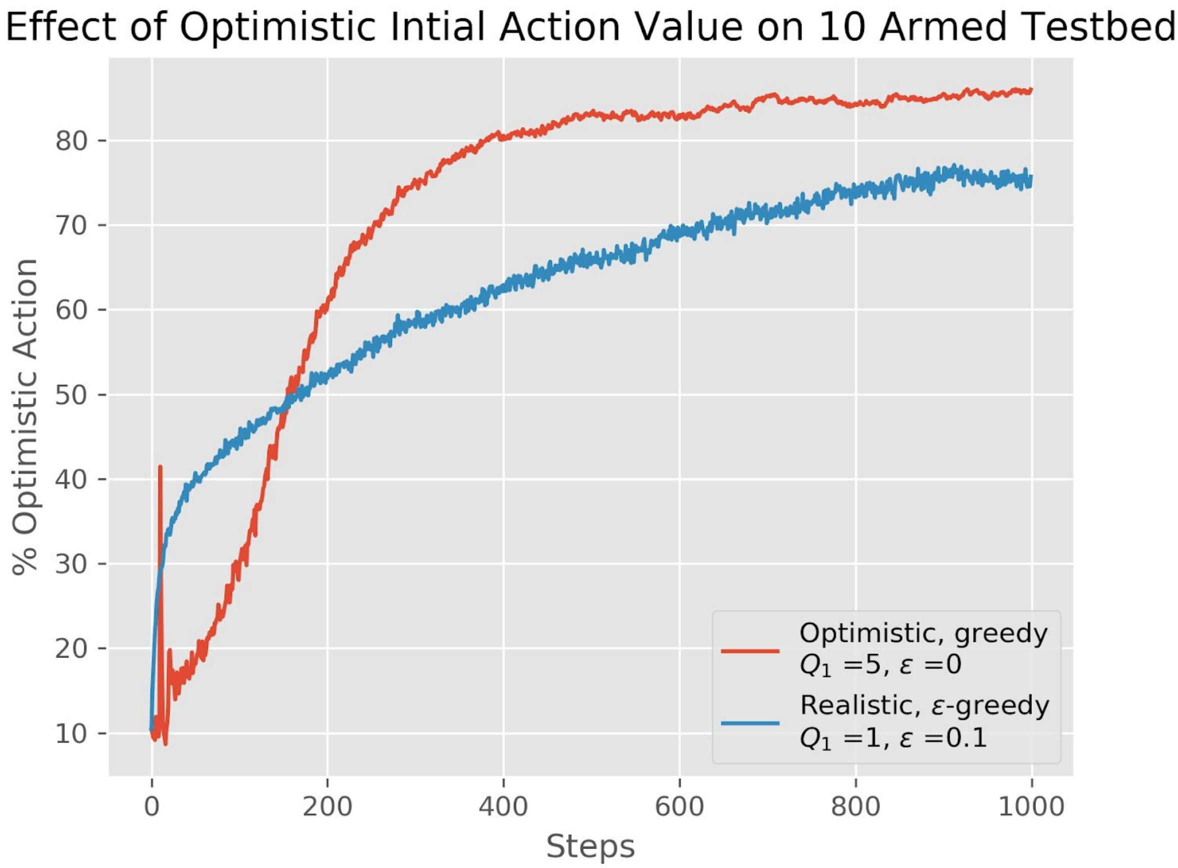


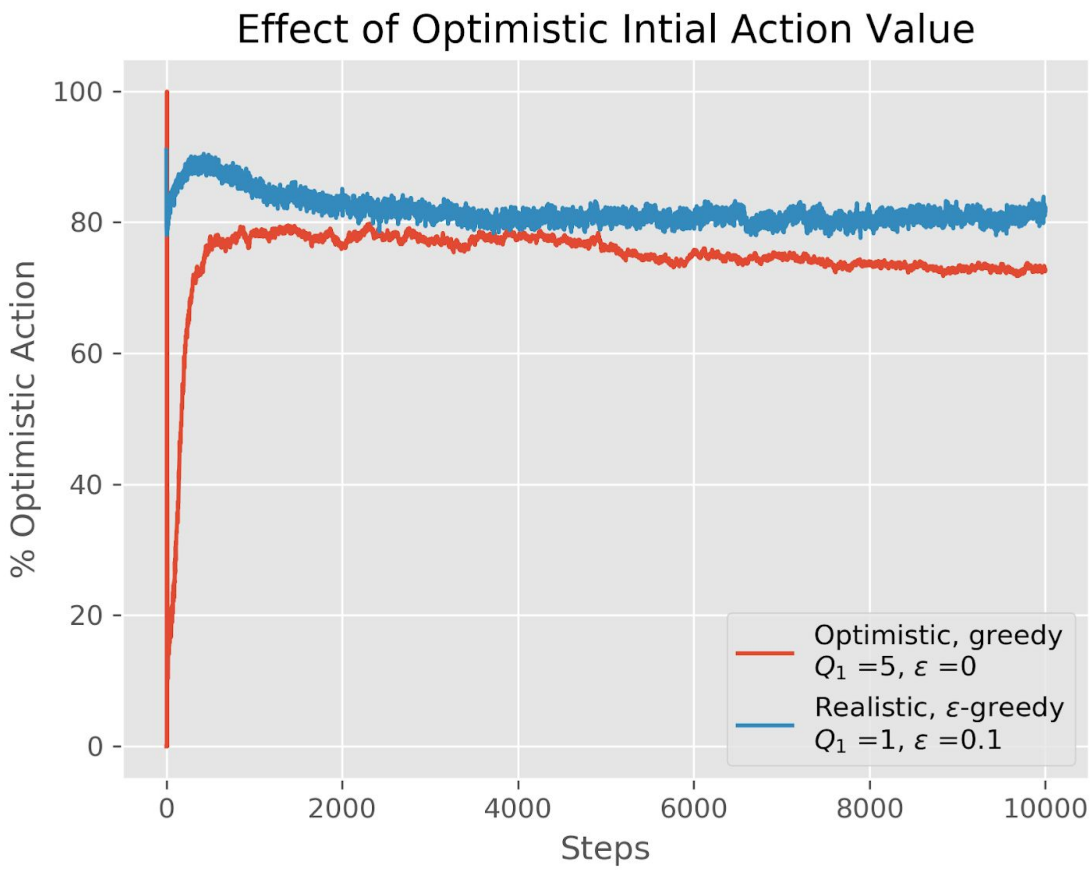
Figure 2.3

For Stationary



- Initially when our agent the exploring it selects all arms once and gets mostly disappointed until it tries out all the arms. But after trying out all the arms there is a high probability to select the optimal action that is why we see the spike at the $t = 10$.
- Another thing which we observed prior to the probability of selecting optimal actions is about 10% .

For Non-Stationary



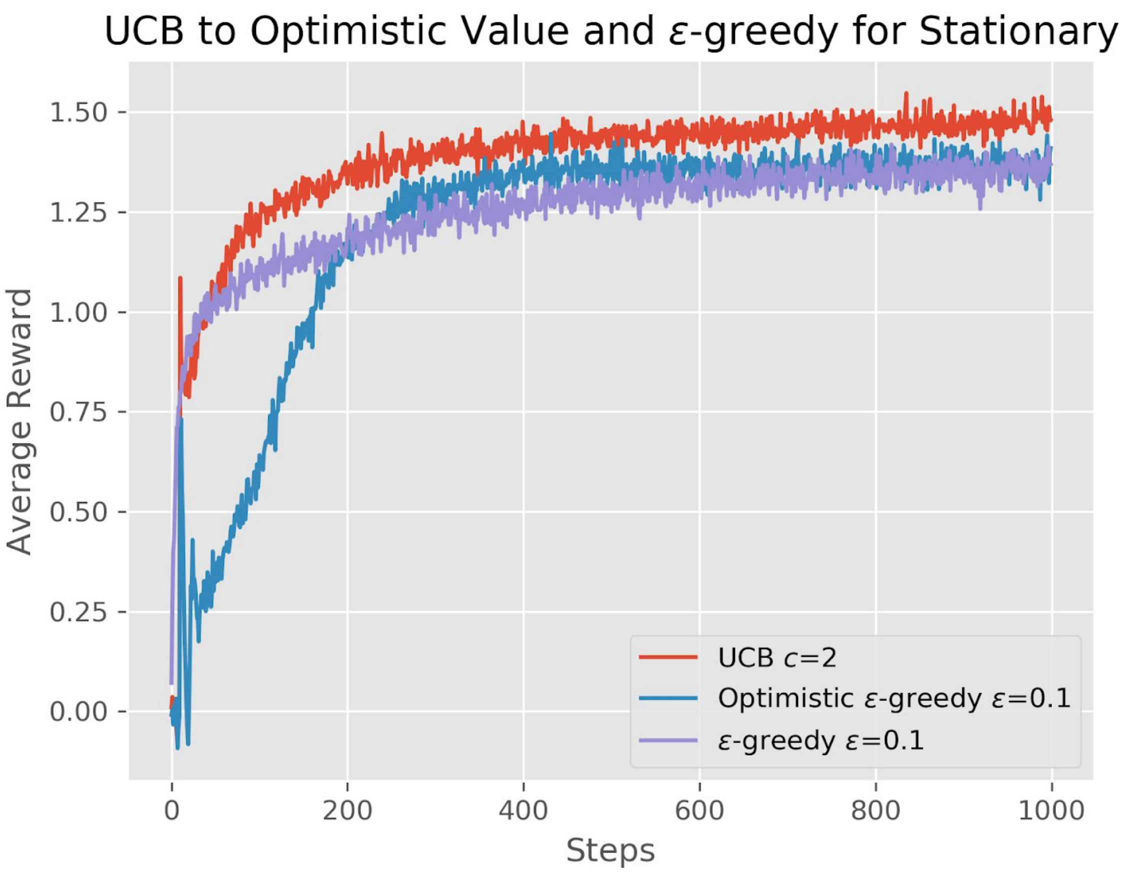
In non-stationary case observe similar observation as compared to stationary case and some more observations which are more related to the non-stationary property of bandits.

- Initially when our agent the exploring it selects all arms once and gets mostly disappointed until it tries out all the arms. But after trying out all the arms there is a high probability to select the optimal action that is why we see the spike at the $t = 10$.
- Another thing which we observed prior to the probability of selecting optimal actions is about 10% .
- In the case of the non-stationary condition, we can see that $Q = 0$ is performing better than the $Q = 5$.

Comparison of UCB to Optimistic Value and ε-greedy

For Stationary

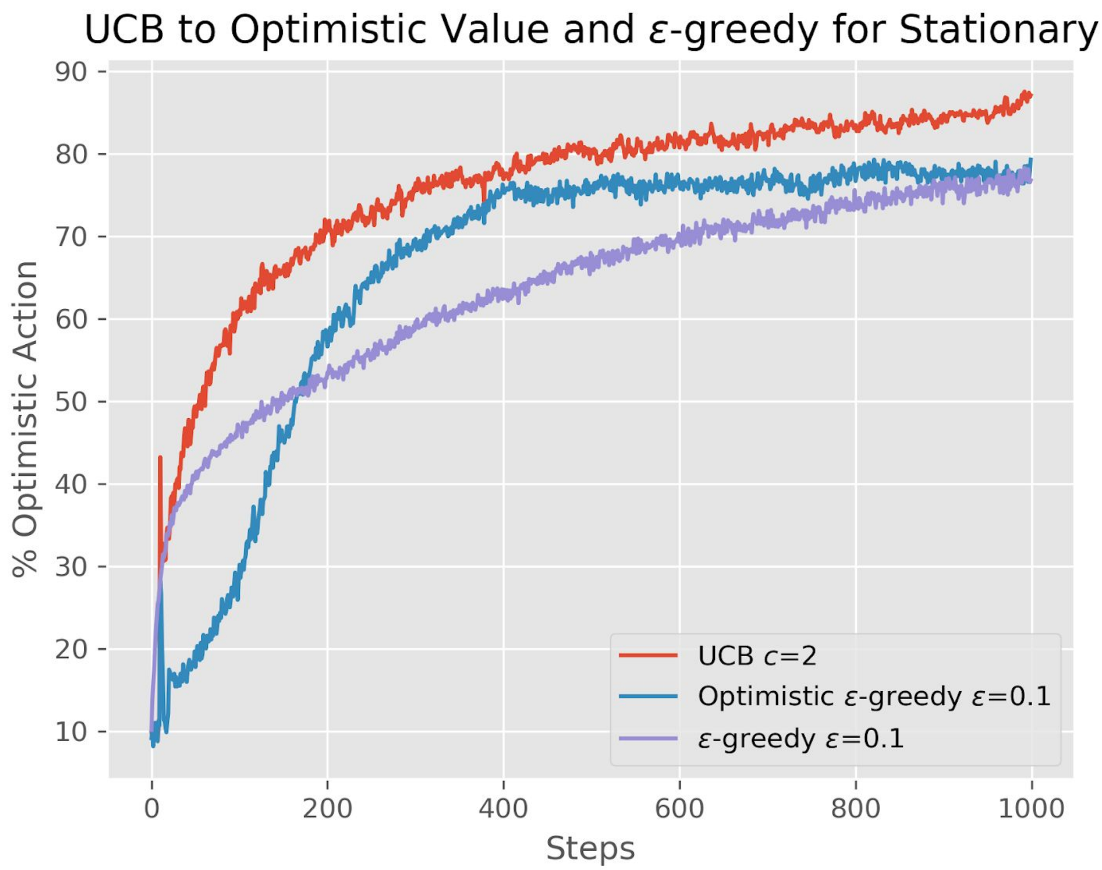
Average rewards Vs Number of total steps



Observations :

- In the case of stationary bandits, we can see that UCB is performing best as compared to ε-greedy method. After UCB optimistic ε- greedy method is performing which ideally should be.

% Optimal Action Vs Number of total steps

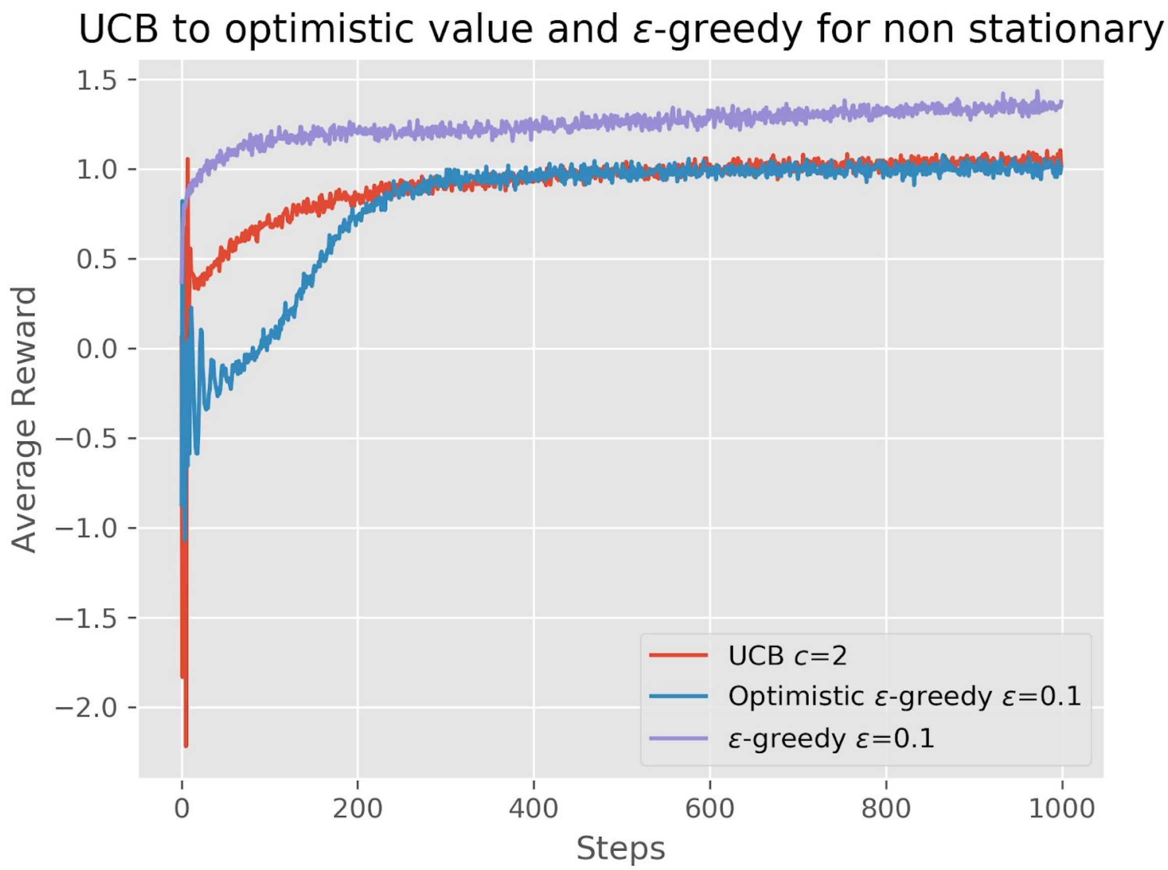


Observations :

Similar trends of average rewards we can see in the %optimal Actions graphs. Which theoretical also proved to be right.

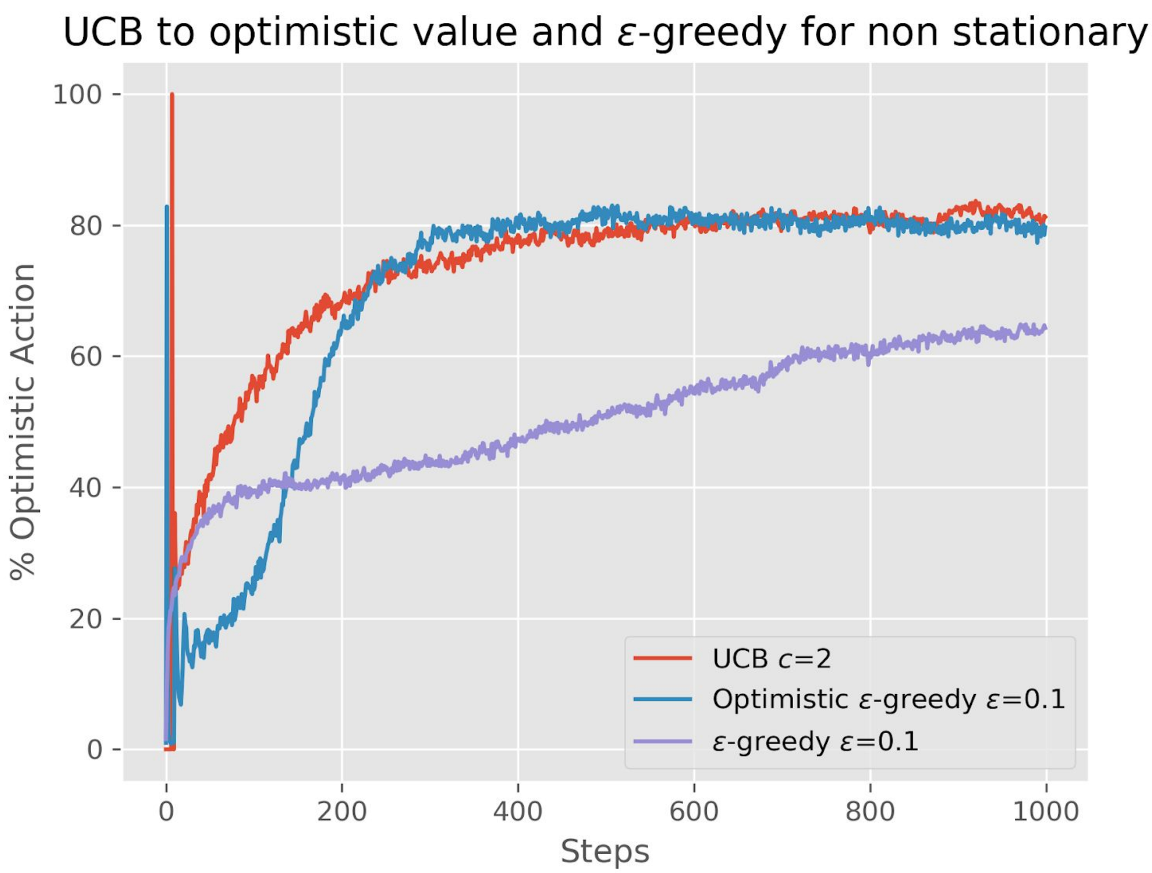
For Non-Stationary

Average rewards Vs Number of total steps



Observations:
In the case of non-stationary bandits ε- greedy is performing best among all the methods, Major reason for this is UCB often performs well, as shown here, but is more di!cult than "-greedy to extend beyond bandits to the more general reinforcement learning settings considered in the rest of this book.
One di!culty is in dealing with nonstationary problems.

% Optimal Action Vs Number of total steps



Date :
Page No. :

Q3

As we know that

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

$$= \alpha R_n + (1-\alpha) Q_n \quad - (2)$$

Now we change this to β which is a step size constant.

and $\beta = \frac{\alpha}{O_n}$.

putting values of β in eqn (2)

$$Q_{n-1} = \frac{\alpha}{\bar{O}_n} R_n + \left(1 - \frac{\alpha}{\bar{O}_n}\right) Q_n$$

$$\frac{\alpha}{\bar{O}_n} R_n + \left(\frac{\bar{O}_n - \alpha}{\bar{O}_n}\right) Q_n$$

By substituting $\bar{O}_n = \bar{O}_{n-1} + \alpha(1 - \bar{O}_{n-1})$

$$\frac{\alpha}{\bar{O}_n} R_n + \left(\frac{\bar{O}_{n-1} + \alpha(1 - \bar{O}_{n-1}) - \alpha}{\bar{O}_n}\right) Q_n$$

Date :
Page No. :

$$\frac{\alpha R_n}{\bar{o}_n} + \left[\frac{\bar{o}_{n-1} + \cancel{\alpha} - \alpha \bar{o}_{n-1} - \cancel{\alpha}}{\bar{o}_n}\right] \varphi_n.$$

$$\frac{\alpha R_n}{\bar{o}_n} + \frac{\bar{o}_{n-1}}{\bar{o}_n}(1-\alpha)\varphi_n \quad — (3)$$

Now from eq^n (1) we again put the value of φ_n.

$$\varphi_{n+1} = \frac{\alpha R_n}{\bar{o}_n} + \frac{\bar{o}_{n-1}}{\bar{o}_n}(1-\alpha)\left[\frac{\alpha}{\bar{o}_{n-1}} R_{n-1} + \left(1 - \frac{\alpha}{\bar{o}_{n-1}}\right) o_{n-1}\right]$$

$$= \frac{\alpha R_n}{\bar{o}_n} + \frac{\alpha(1-\alpha)}{\bar{o}_n} R_{n-1} + \frac{\bar{o}_{n-1}}{\bar{o}_n}(1-\alpha)(1-\alpha)\frac{\bar{o}_{n-2}}{\bar{o}_{n-1}}\varphi_{n-1}$$

$$= \alpha\frac{R_n}{\bar{o}_n} + \frac{\alpha(1-\alpha)}{\bar{o}_n} R_{n-1} + (1-\alpha)^2\frac{\bar{o}_{n-2}}{\bar{o}_n}\varphi_{n-1}.$$

$$= (1-\alpha)^n \frac{\bar{o}_0}{o_n}\varphi_1 + \sum_{i=1}^{n} \alpha(1-\alpha)^{n-i}\frac{R_i}{\bar{o}_i} \quad — (4)$$

Date :
Page No. :

from equation (4)

$$\varphi_{n+1} = (1-\alpha)^n \frac{\bar{O}_0}{O_2} \cdot \varphi_1 + \sum_{i=1}^{n} \alpha(1-\alpha)^{n-i} \frac{R_i}{O_i}$$

since $\bar{O}_0 = 0$ (given)

equation (4) becomes.

$$\varphi_{n+1} = \sum_{i=1}^{n} \alpha(1-\alpha)^{n-i} \frac{R_i}{O_i}$$

Method II

In this method we solve our equation from starting point zero whereas in the last method we solved from the back side that is n.

$$O_1 = \bar{O}_0 + \alpha(1 - O_0) \quad \text{(given)}$$

$$= \bar{O}_0 + \alpha - \alpha\bar{O}_0$$

$$\bar{O}_0 = 0$$

Date :
Page No. :

$\bar{O}_1 = \alpha$

similarly for $\bar{O}_2$

$$\bar{O}_2 = \bar{O}_1 + \alpha(1 - \bar{O}_1)$$

$$= \bar{O}_1 + \alpha(1 - \alpha)$$

$$= \alpha + \alpha - \alpha^2$$

$$\bar{O}_2 = 2\alpha - \alpha^2.$$

And from constant stepsize equation
We know

$$Q_{n+1} = \alpha R_n + (1-\alpha) Q_n$$

Substituting α with $\beta_n = \frac{\alpha}{\bar{O}_n}$

In this case Q_1 doesn not

because $\bar{O}_0$ is zero

Date:
Page No.:

so the Next term.

$$\phi_2 = \frac{\alpha}{\sigma_1} R_1 + \left(1 - \frac{\alpha}{\sigma_1}\right)\phi_1$$

$$= \frac{\cancel{\alpha}}{\cancel{\alpha}} R_1 \left(1 - \frac{\alpha}{\alpha}\right)\phi_1$$

$$= R_1 + 0.$$

$$\boxed{\phi_2 = R_1} \quad -(1)$$

for ϕ_3.

$$\phi_3 = \frac{\alpha}{\sigma_2} R_2 + \left(1 - \frac{\alpha}{\sigma_2}\right)\phi_2$$

$$= \frac{\cancel{\alpha}}{\cancel{\alpha}(2-\alpha)} R_2 + \left(1 - \frac{\alpha}{2(1-\alpha)}\right)\phi_2$$

$$= \frac{R_2}{2-\alpha} + \left(\frac{2-2\alpha-\alpha}{2(1-\alpha)}\right)\phi_2$$

and $\phi_2 = R_2$

Date :
Page No. :

$$\frac{R_2}{2-\alpha}\left[R_2 + (3-\alpha)R_1\right] \quad -(2)$$

if we proceed further in similar fashion we can see that there will no term for Q_n involved. which was involved in the case of constant step size.

Scanned by CamScanner