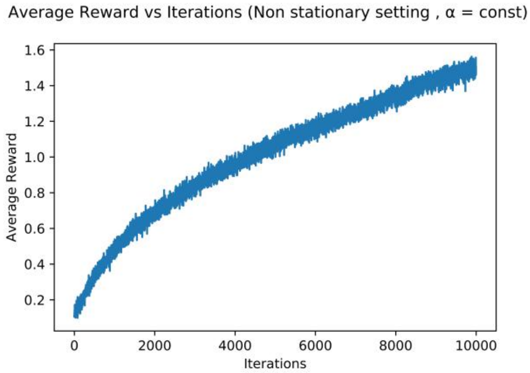
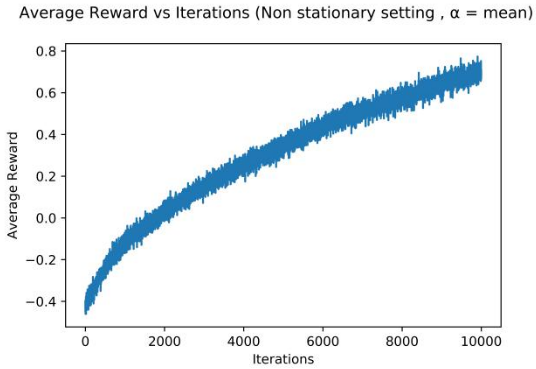
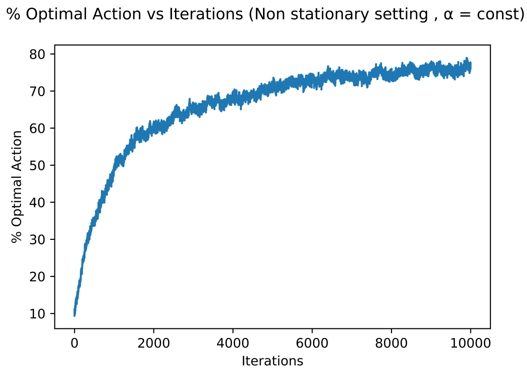
Name: Aman Mehra Roll No: 2017017

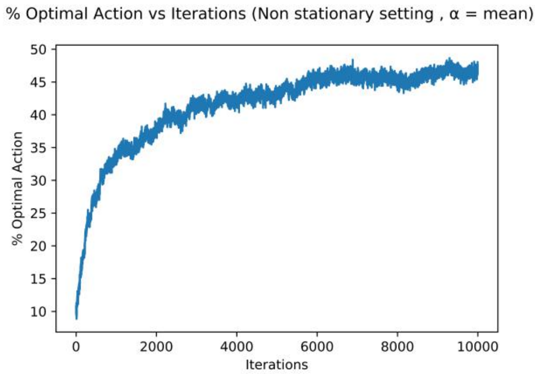
Assignment Q1.

Q 2.5 -



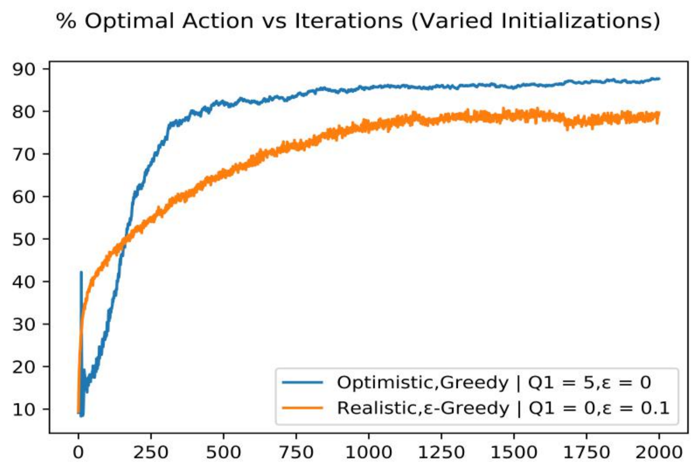






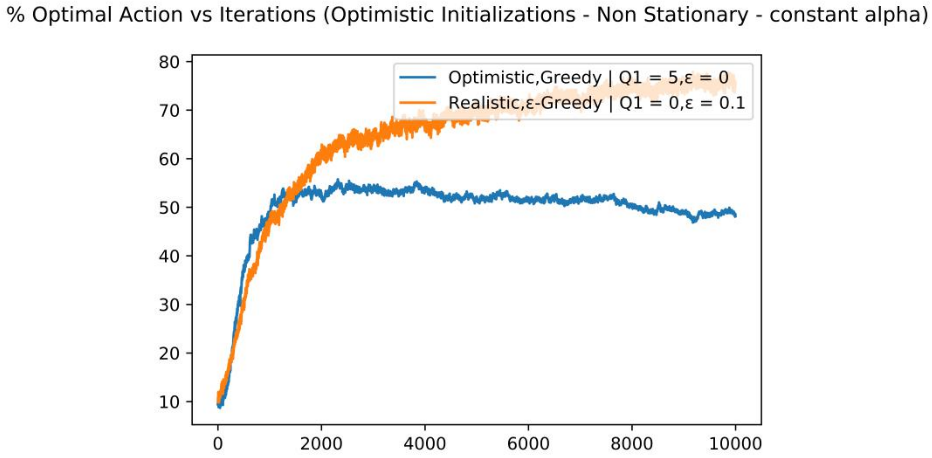
Assignment Q2.

Fig 2.3 -



Q 2.6 - When we do an optimistic initialization, the agent expects greater than much higher than actual estimated rewards. But on selecting an action the agent receives a reward lower than its expectation. Thus $Q_{t+1}(a)$ becomes smaller implicitly enforcing exploration. Due to the experiment using a small finite number of possible bandit arms (=10) we see that in the initial epochs a large fraction of the agents end up choosing the optimal arm on a given iteration leading to the perceived spike. Subsequently as the reward received is lower than expected agents continue exploration and move to other arms. Such high fraction of optimal arms do not happen all of sudden again after this, leading to no further spikes.

Fig 2.3 for Non Stationary Case -



Explanation of Fig 2.3 for non stationary case - Optimistic initialization promotes exploration only in the initial iterations, i.e. it is only a temporary impetus. This works well for the stationary case due to the greater exploration in less time, but this does not work well for the non stationary case as the exploration impetus is soon lost. Since the distribution of the arms are changing the lack of exploration in later iterations reduces the overall reward received.

Assignment Q3.

Ex 2.7 $Q_{n+1} = Q_n + \beta_n [R_n - Q_n]$

$$\beta_n = \frac{\alpha}{\bar{o}_n}$$

$$\bar{o}_{n+1} = \bar{o}_n + \alpha(1-\bar{o}_n) \quad ; \quad \bar{o}_0 = 0$$

Now to show that β_n avoids initial bias by getting rid of the Q_1 term in all future Q_n

$$\bar{o}_1 = \alpha \qquad \beta_1 = 1$$

$$\bar{o}_2 = \alpha(2-\alpha) \qquad \beta_2 = \frac{1}{2-\alpha}$$

$$Q_2 = Q_1 + \frac{\alpha}{\alpha}[R_1 - Q_1] = R_1$$

$$Q_3 = Q_2 + \frac{1}{2-\alpha}[R_2 - Q_2]$$

$$= R_1 + \frac{1}{2-\alpha}(R_2 - R_1) = \frac{1-\alpha}{2-\alpha} R_1 + \frac{1}{2-\alpha} R_2$$

Generalizing

$$Q_2 = \beta_1 R_1$$

$$Q_3 = \beta_2 R_2 + \beta_1 (1-\beta_2) R_1$$

$$\vdots$$

$$Q_{n+1} = \beta_n R_n + \beta_{n-1}(1-\beta_n) R_{n-1} + \cdots + \beta_1 (1-\beta_2) \cdots (1-\beta_n) R_1 \quad \text{—(1)}$$

Thus as there is no correlation between Q_{n+1} and Q_1, β_n is unbiased.

Now $\bar{o}_n = \bar{o}_{n-1} + \alpha(1-\bar{o}_{n-1})$

$$\Rightarrow \bar{o}_n = (1-\alpha)\bar{o}_{n-1} + \alpha$$

$$= (1-\alpha)^2 \bar{o}_{n-2} + \alpha(1-\alpha) + \alpha$$

$$\vdots$$

$$(1-\alpha)^n \bar{o}_0 + \alpha \sum_{i=0}^{n-1} (1-\alpha)^i$$

Now $\bar{o}_0 = 0$

$$\Rightarrow \bar{o}_n = \alpha \sum_{i=0}^{n-1} (1-\alpha)^i$$

$$= \alpha \left[\frac{(1-\alpha)^n - 1}{1-\alpha-1} \right]$$

$$= \cancel{\alpha}\left(\frac{1-(1-\alpha)^n}{\cancel{\alpha}}\right)$$

$$\Rightarrow \beta_n = \frac{\alpha}{1-(1-\alpha)^n}$$

For Q_n's equation (Eqn (1))

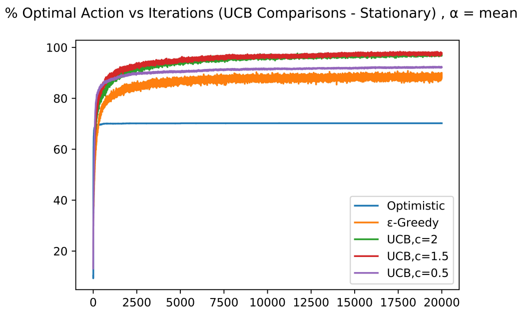
$$\Rightarrow Q_n = \frac{\alpha}{1-(1-\alpha)^{n-1}} R_{n-1} + \cdots \left(\frac{1-\alpha}{1-(1-\alpha)^{n-1}}\right) \cdots \left(\frac{1-\alpha}{1-(1-\alpha)^i}\right) R_i$$

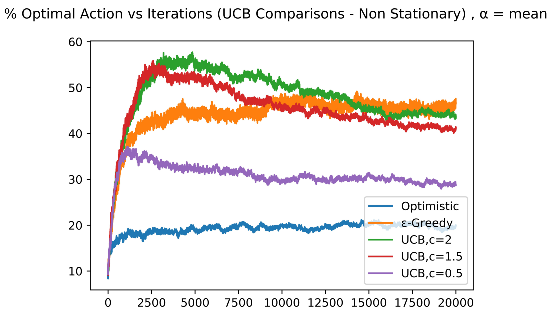
Clearly this is exponentially decreasing, the older the reward gets.

~~thi~~ Thus this is an exponential recency weighted average without initial bias.

Name: Aman Mehra Roll No: 2017017

Assignment Q4.





UCB comparison with other methods - UCB (using sample mean as step size) performs better compared to both optimistic initialization as well as ε-Greedy methods in the stationary setting. In the non-stationary setting, we see that UCB performs significantly better for a fairly long period initially, but gradually performance reduces. After a long period of time ε-Greedy becomes the best performer. This can be attribute to the log t factor in UCB which as time progresses leads to a lower exploration impetus and after a long time the impetus for exploration falls below that of ε-Greedy. The figures also demonstrate the effect of using different constants in the UCB method. Higher the constant the longer the exploration impetus remains.