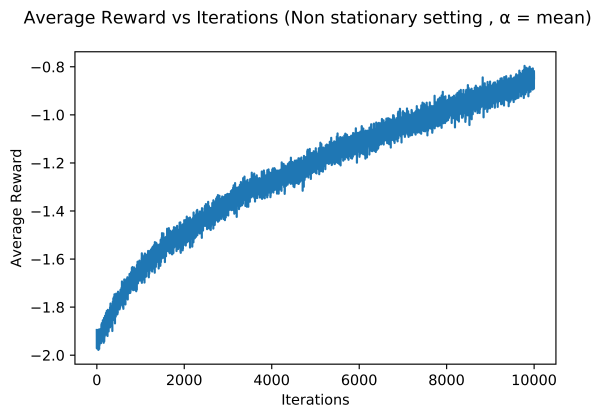
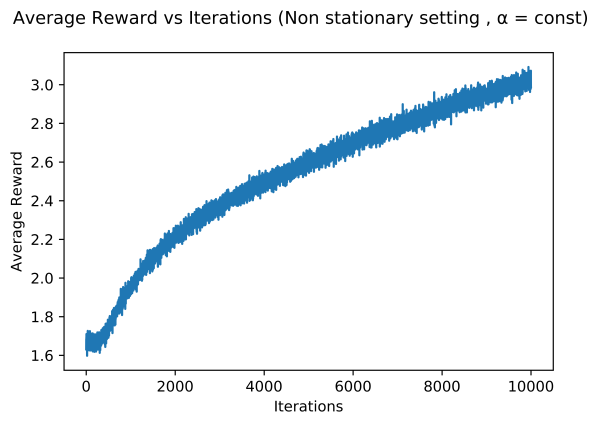
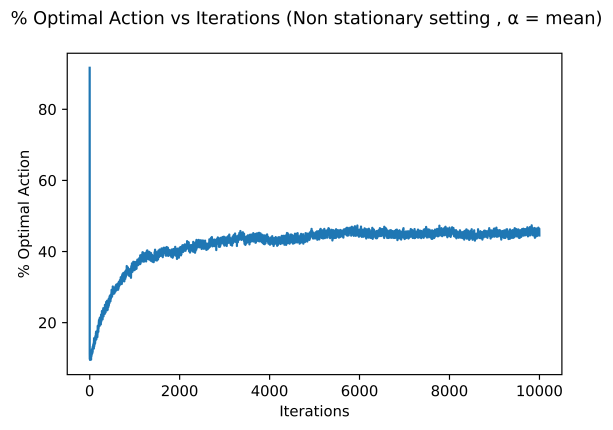
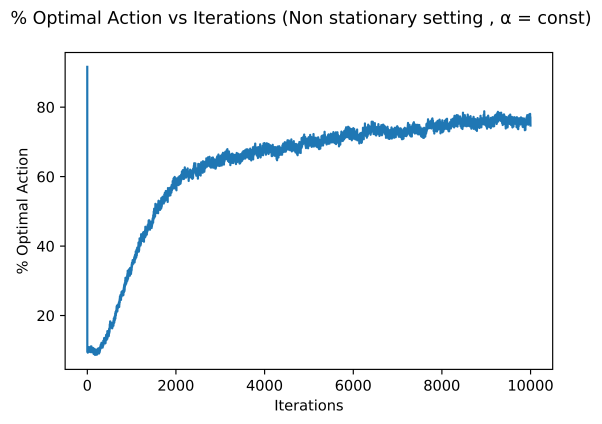
**Assignment Q1.**

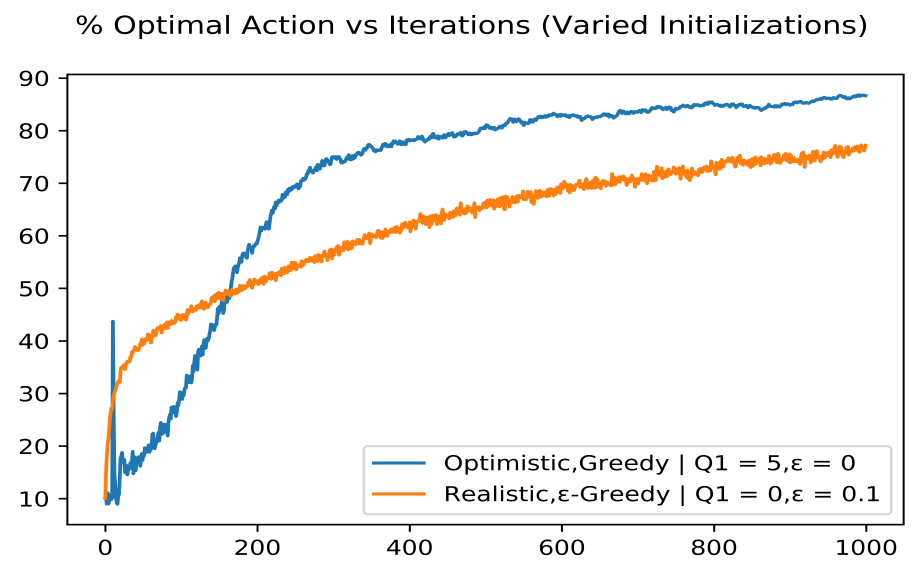
**Q 2.5** -



****

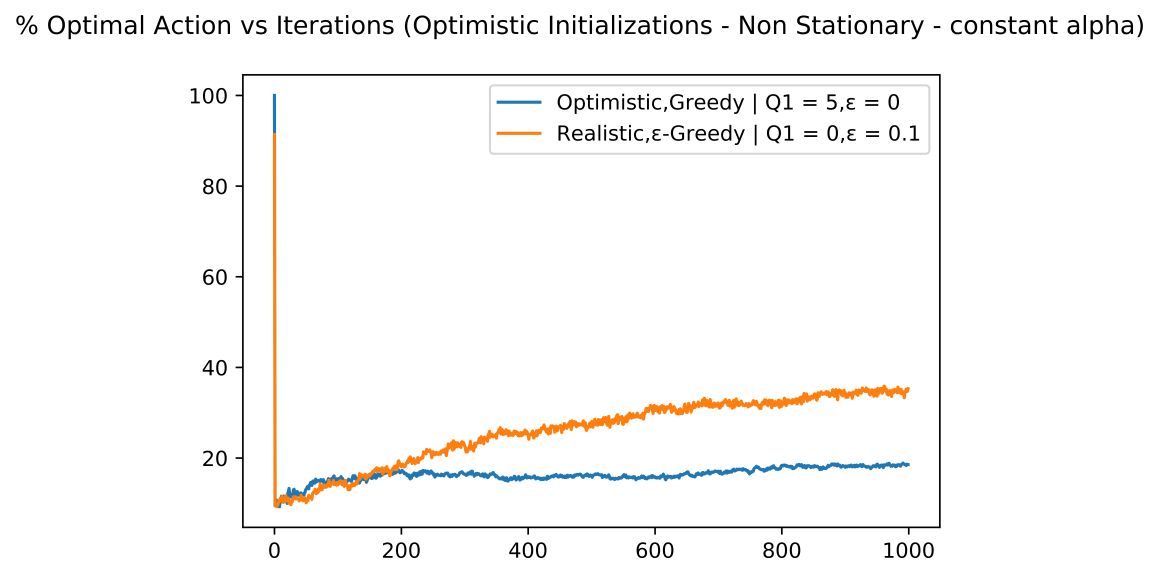
**Assignment Q2.**

**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 Qt+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** -

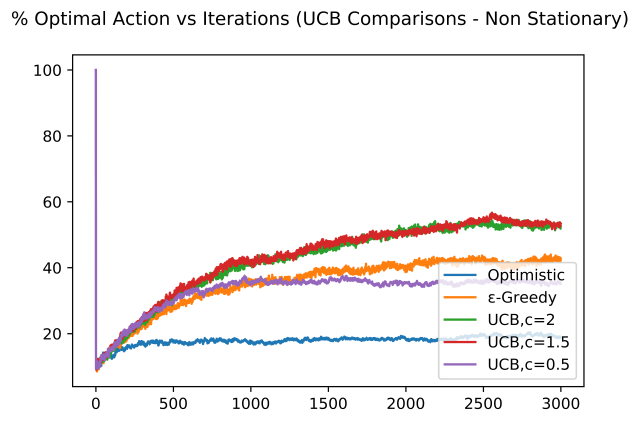
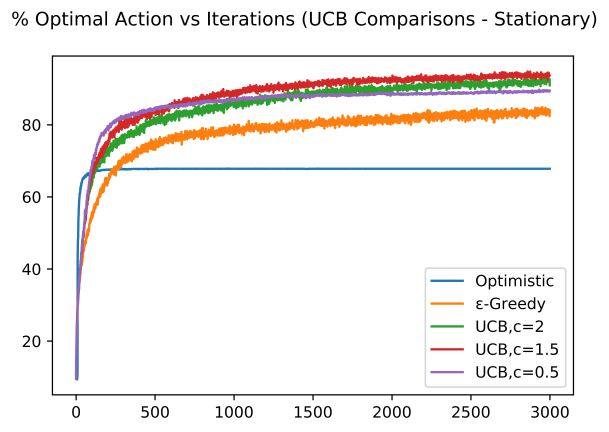


**Explanation of Fig 2.3 for non stationary case** - Optimistic initialization promotes exploration only in the initial iterations, i.e. it is only a temporary impetus. This works well for the stationary case due to the greater exploration in less time, but this does not work well for the non stationary 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.**

|  |  |
| --- | --- |
| RL ex2.7_1 | RL ex2.7_2 |

**Assignment Q4.**

****

**UCB comparison with other methods** - UCB performs better compared to both optimistic initialization as well as ε-Greedy methods. This affect is seen both in the stationary as well as the non-stationary setting, validating the belief that the UCB methods promotes exploration by selecting actions with potentially better rewards. The figures also demonstrate the effect of using different constants in the UCB method.