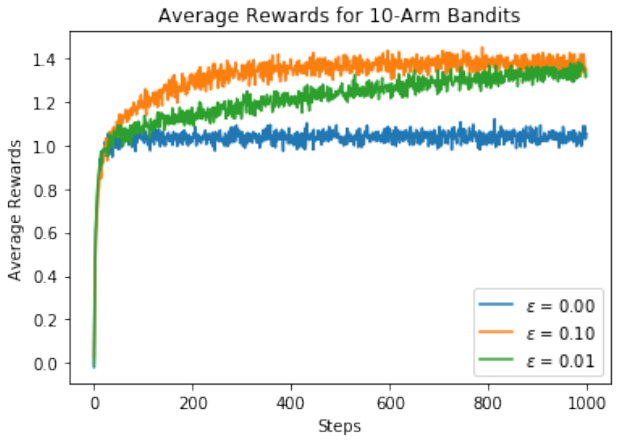
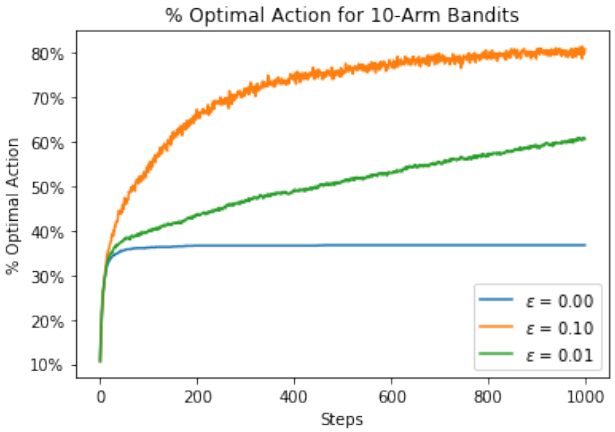
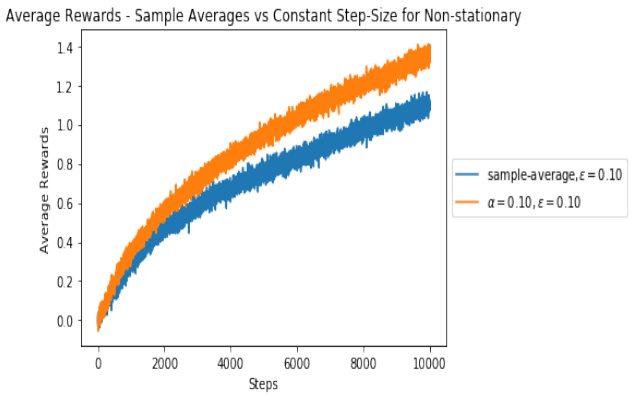
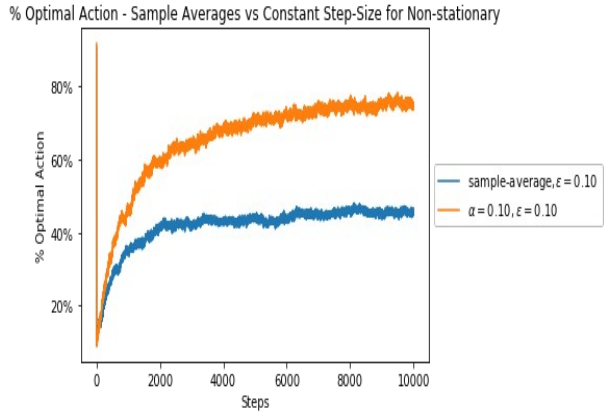
ASSIGNMENT - 1

Additional Part:
10arm bandit implementation:
Assumptions : stationary case, 2000 tasks, 1000 time steps

Avearge Rewards	Optimal Count
	

Part 1 : Ex 2.5
Assumptions : non-stationary case, 2000 tasks, 10000 time steps

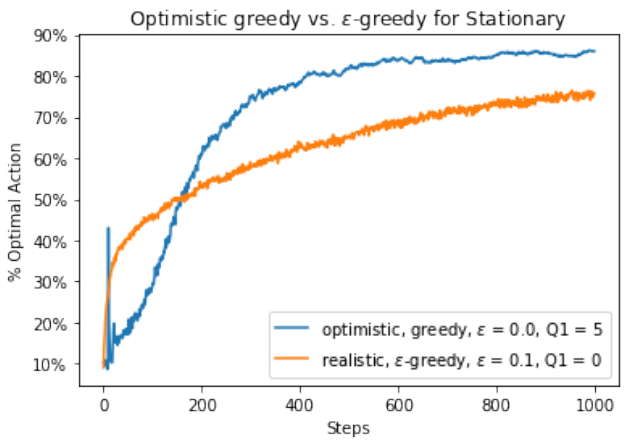
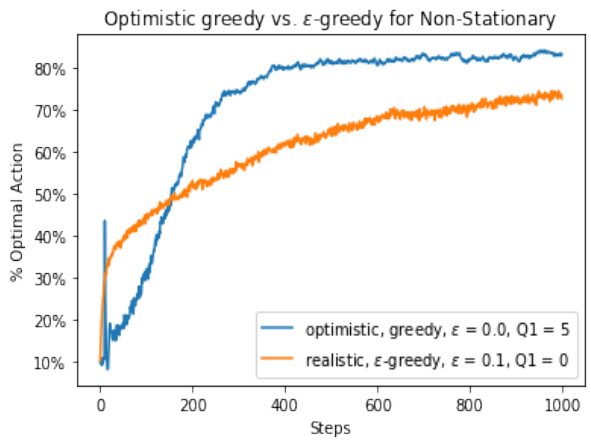
Avearge Rewards	Optimal Count
	

Analysis:

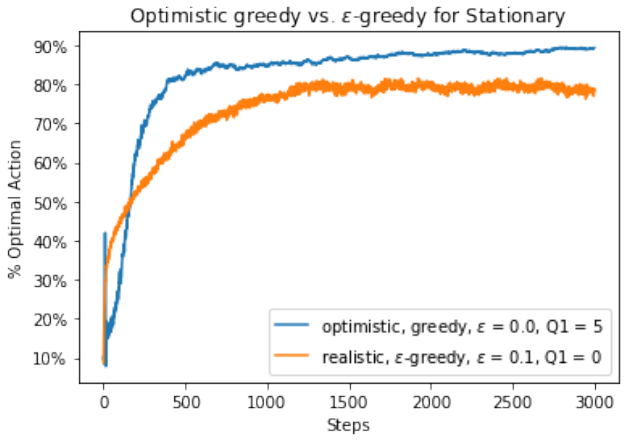
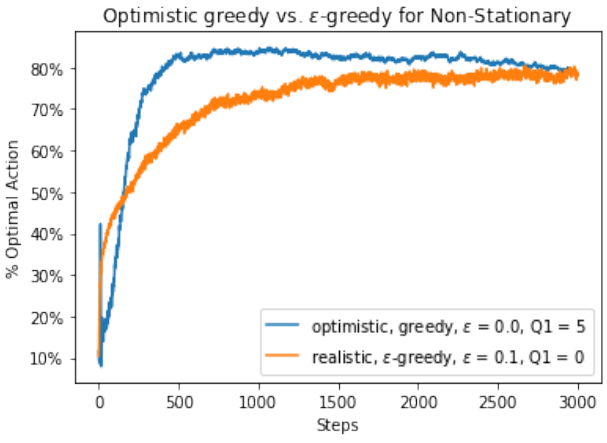
- For both optimal count as well as avergae reward, constant step size method performs better than sample avergae method for estmating action-values.
- It can be clearly seen that constant step size method achieves higher average reward as much as 1.4 and also choses optimal action more frequently than sample-average method which has reached the highest reward in between 1.0 -1.2 only and in case of optimal action, the curve plateaus nearly after 2000-4000 steps.
- This clearly indicates that the sample average methods for action estimation does not do well in case of non-stationary environment and hence the constant step size method is much appropriate for this case.

Part 2: a) Fig 2.3

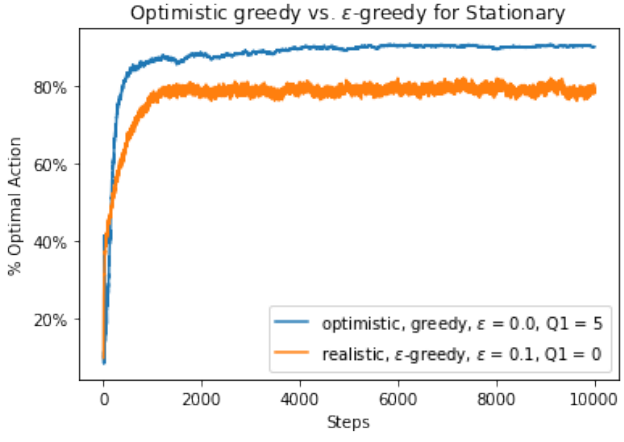
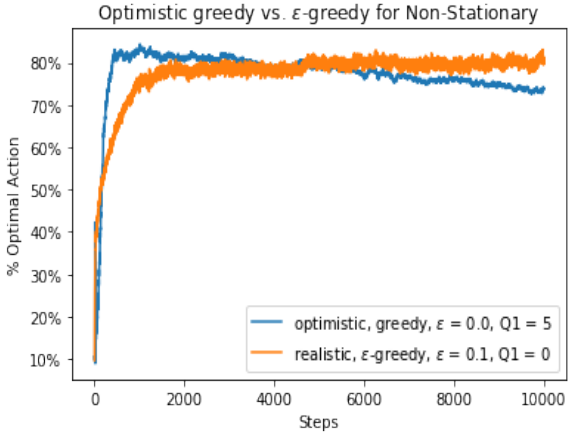
Assumptions : non-stationary case, 2000 tasks, 1000 time steps

Stationary	Non-Stationary
	

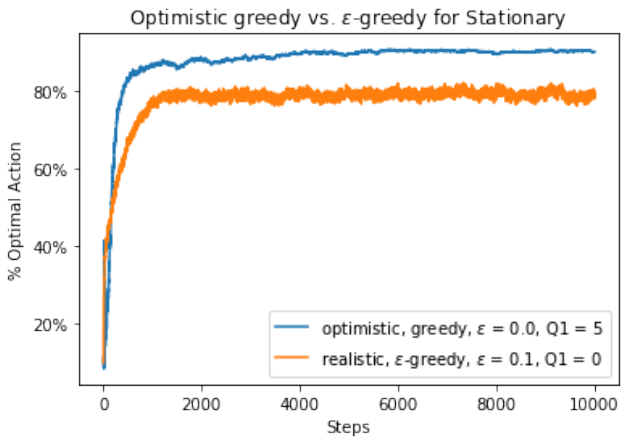
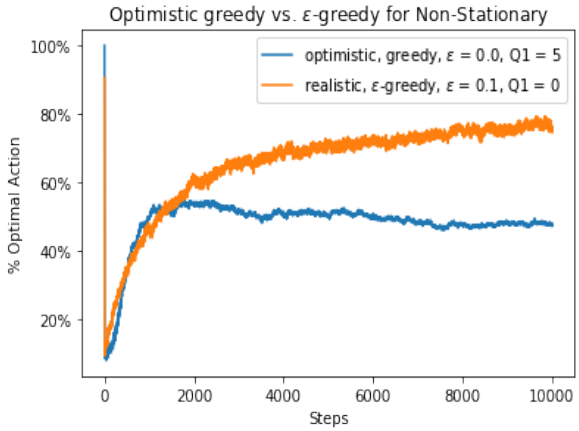
Assumptions : non-stationary case, 2000 tasks, 3000 time steps

Stationary	Non-Stationary
	

Assumptions : non-stationary case, 2000 tasks, 10000 time steps

Stationary	Non-Stationary
	

Assumptions : 2000 tasks, 10000 time steps and all q*(a) starts out equal in non-stationary case

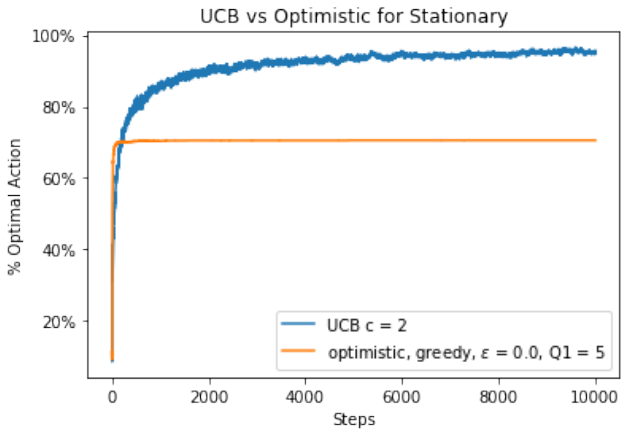
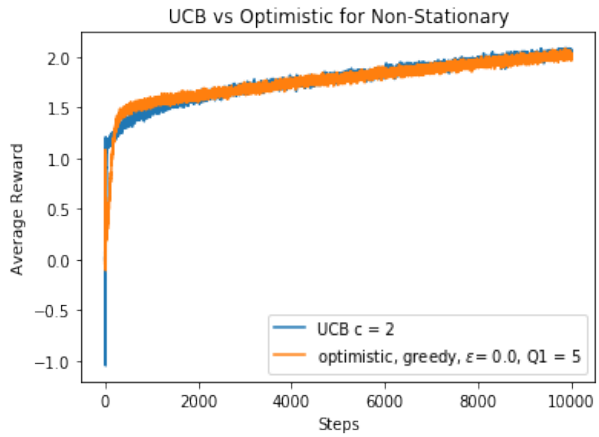
Stationary	Non-Stationary
	

Analysis:

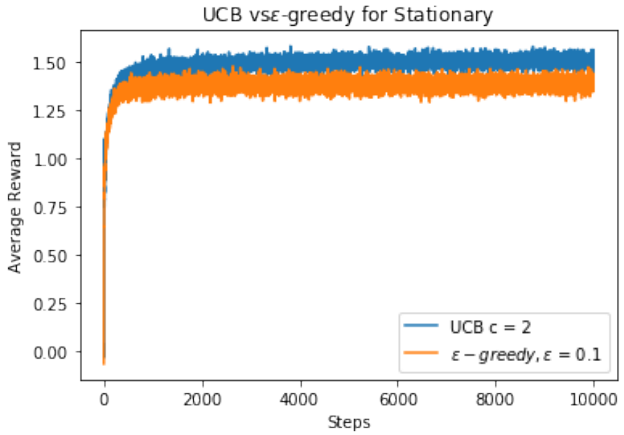
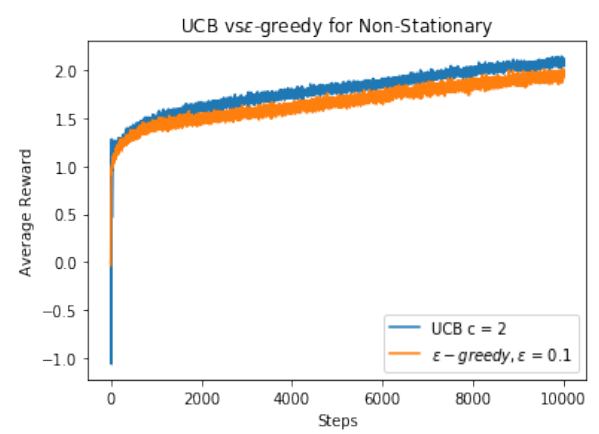
- **Stationary Case :**
  - optimistic-method for stationary case explores more and performs poorly initially but eventually performs better with more runs and hence turns out better method than the epsilon-greedy.
- **Non-Stationary Case :**
  - The initial spike in the graph is due to the fact that we start out with equal q*(a) and estimates due to which the algorithm choses the same action as the optimal action but with time, as the q*(a) values for all actions change, the algorithm starts chosing different actions as estimates also begin to change.
  - Hence, slowly both the algorithms begin to learn but optimistic-method for non-stationary case performs better initially, however, with more runs or over time, its performance deteriorates.
  - This shows that optimistic-method is not suited for non-stationary problems because its drive for explorattion is inherently temporary. If the task changes creating a renewed need for exploration, this method cannot help.

Part 2: b) Ex 2.6 done separately
Part 3: Ex 2.7 done separately

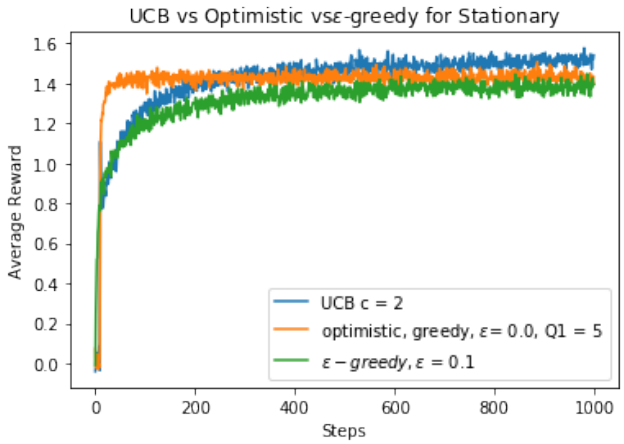
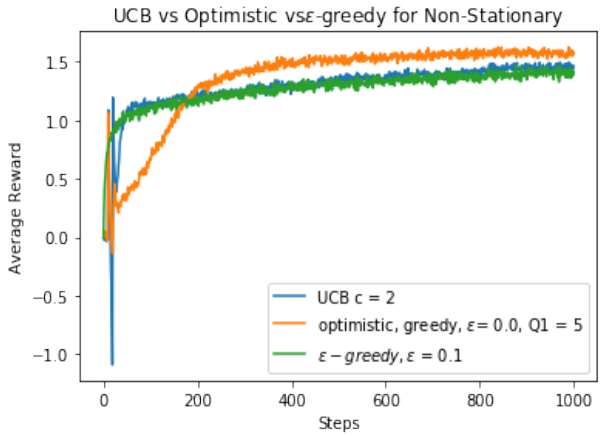
Part 4. UCB vs Optimistic Value
Assumptions : 2000 tasks, 10000 time steps

Stationary	Non-Stationary
	

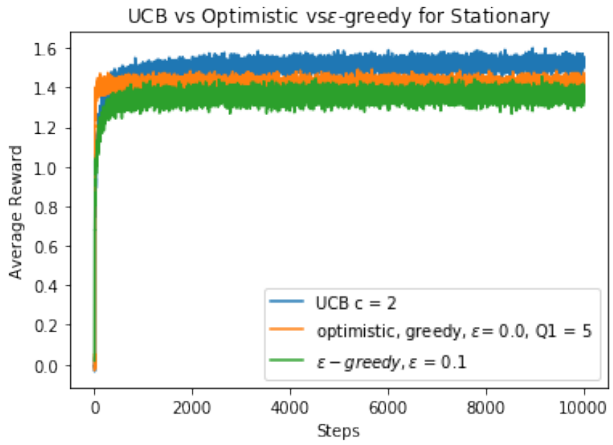
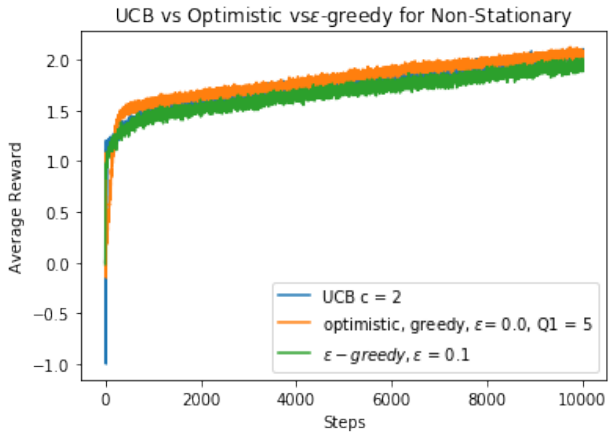
UCB vs E-greedy
Assumptions : 2000 tasks, 10000 time steps

Stationary	Non-Stationary
	

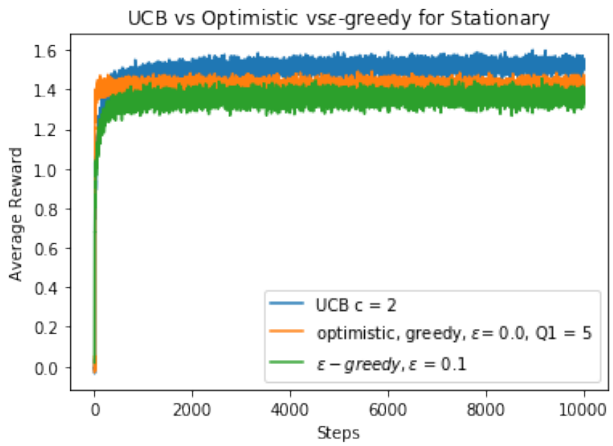
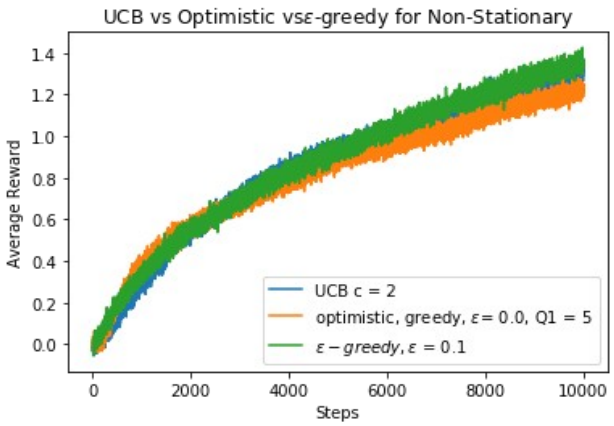
UCB vs Optimistic vs E-greedy
Assumptions : 2000 tasks, 1000 time steps

Stationary	Non-Stationary
	

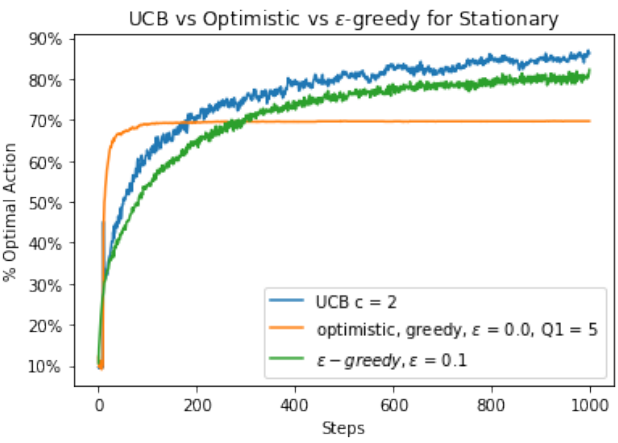
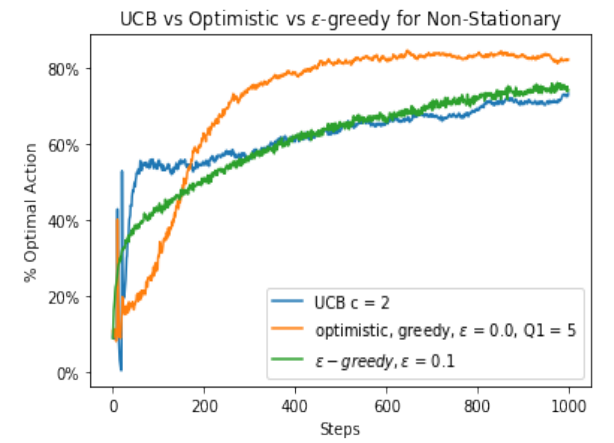
Assumptions : 2000 tasks, 10000 time steps

Stationary	Non-Stationary
	

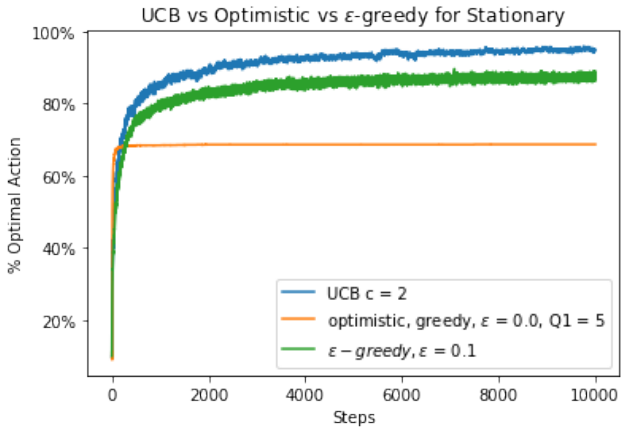
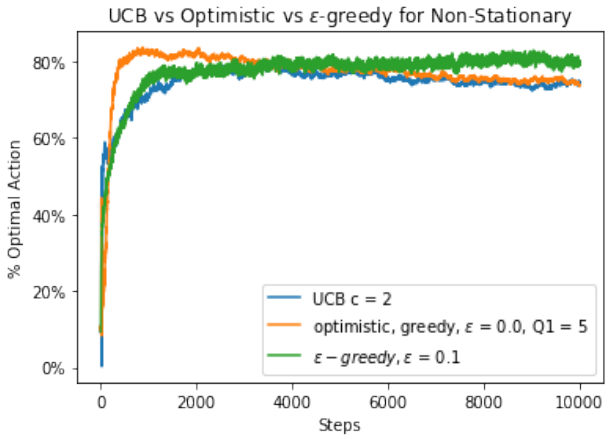
Assumptions : 2000 tasks, 10000 time steps and all q*(a) start out equal initially in non-stationary case

Stationary	Non-Stationary
	

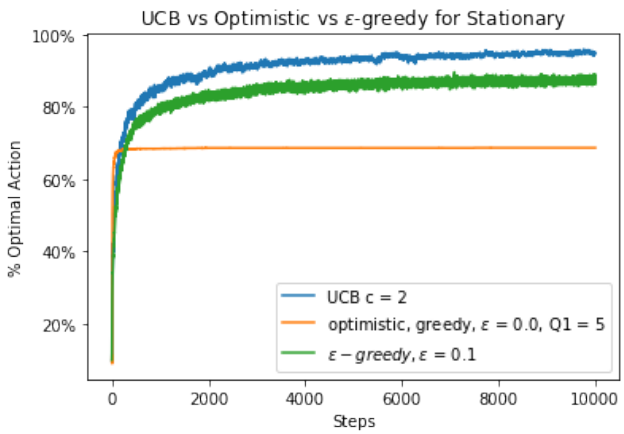
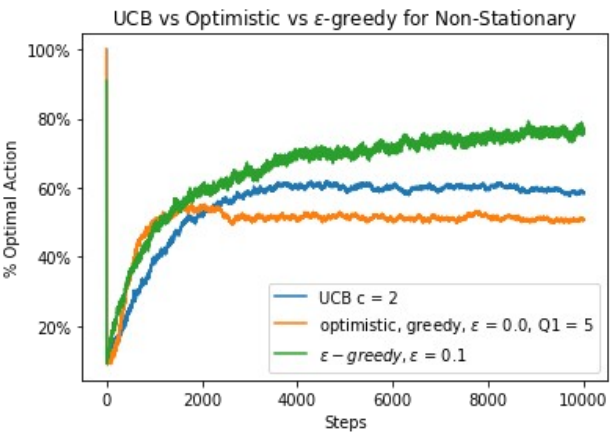
UCB vs Optimistic vs E-greedy (Optimal Actions)
Assumptions : 2000 tasks, 1000 time steps

Stationary	Non-Stationary
	

Assumptions : 2000 tasks, 10000 time steps

Stationary	Non-Stationary
	

Assumptions : 2000 tasks, 10000 time steps and all q*(a) are equal initially in non-stationary case

Stationary	Non-Stationary
	

Analysis:

Stationary Case :

- Of all three methods, UCB performs best. Although optimistic-greedy improved slightly faster than the other two methods but performs worst in long runs. Although epsilon-greedy starts converging ealry but UCB outperforms in the long runs. Same is the case for optimal action count.

- **Non-Stationary Case:**
  - Initially, UCB and optimistic greedy have more oscillations and spikes as they explore more but eventually they become better with more runs. UCB performs slightly better than epsilon-greedy and optimistic where optimistic is slow to converge or learn initially as compared to other two but paces up later.
  - However, for the case of optimal action count, UCB does not perform well compared to the other two whereby it starts to learn or converge early but later performance deteriorates. Also, it oscillates more throughut. This indicates that although it performs good in stationary case but is not a suitable choice for non-stationary case. Overall (i.e from optimal action plot as well), UCB does not perform well for non-stationary case when compared to the other two models. Epsilon-greedy balances exploitation and exploration most appropriately in case of non-stationary case and performs best among the other two.