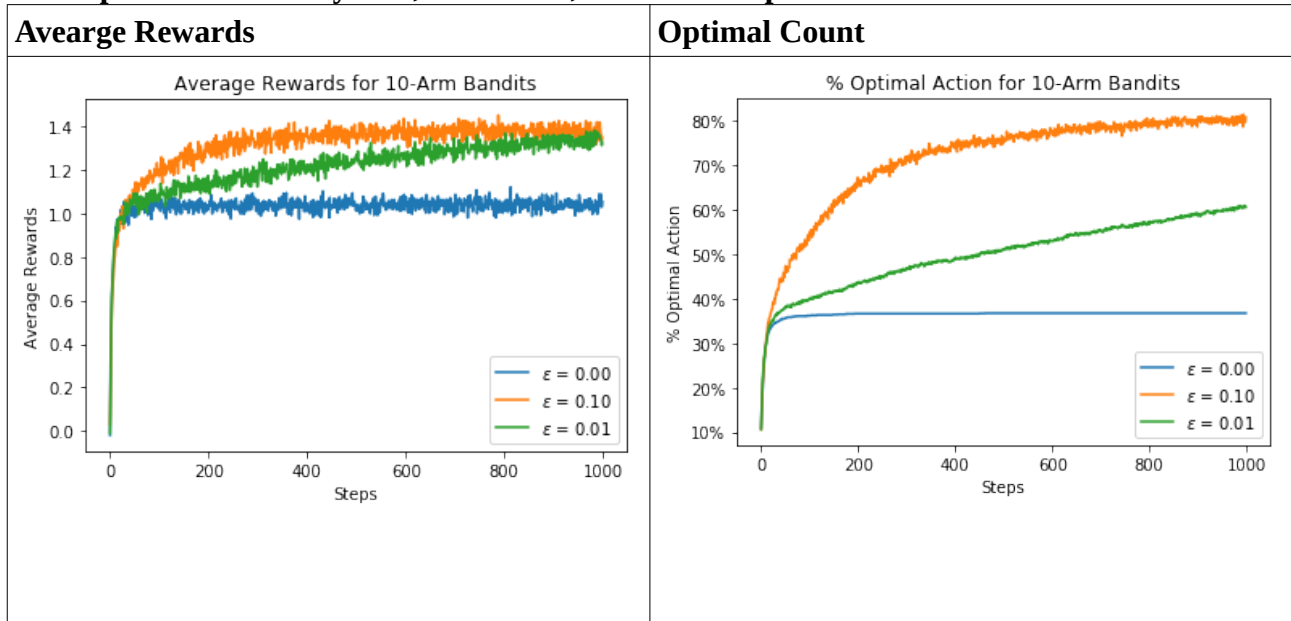


ASSIGNMENT - 1

Additional Part:

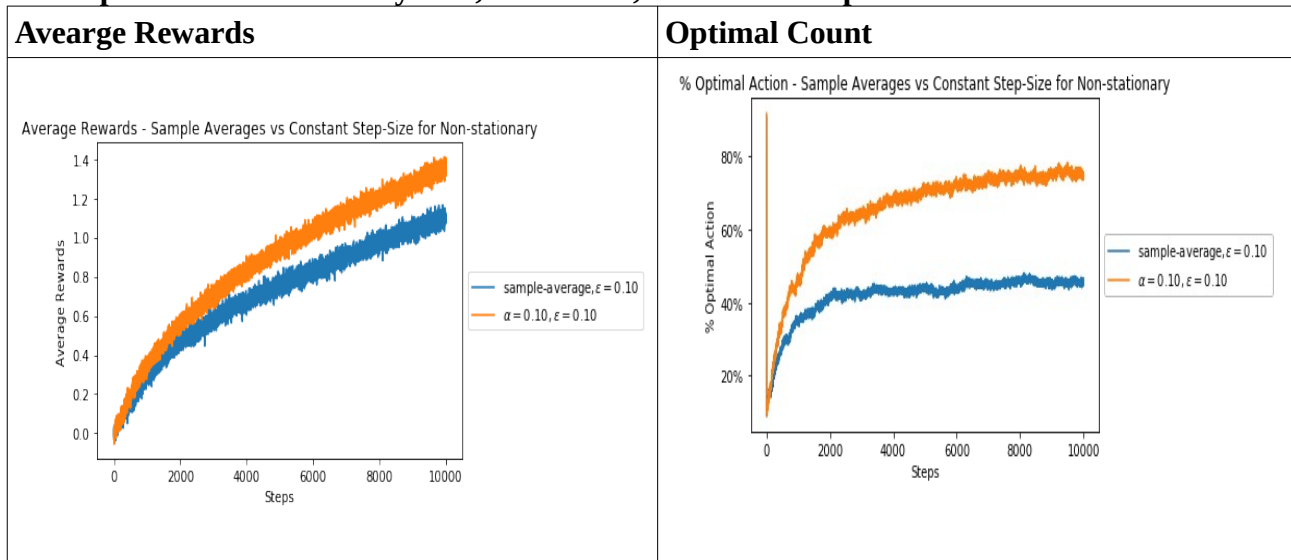
10arm bandit implementation:

Assumptions : stationary case, 2000 tasks, 1000 time steps



Part 1 : Ex 2.5

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

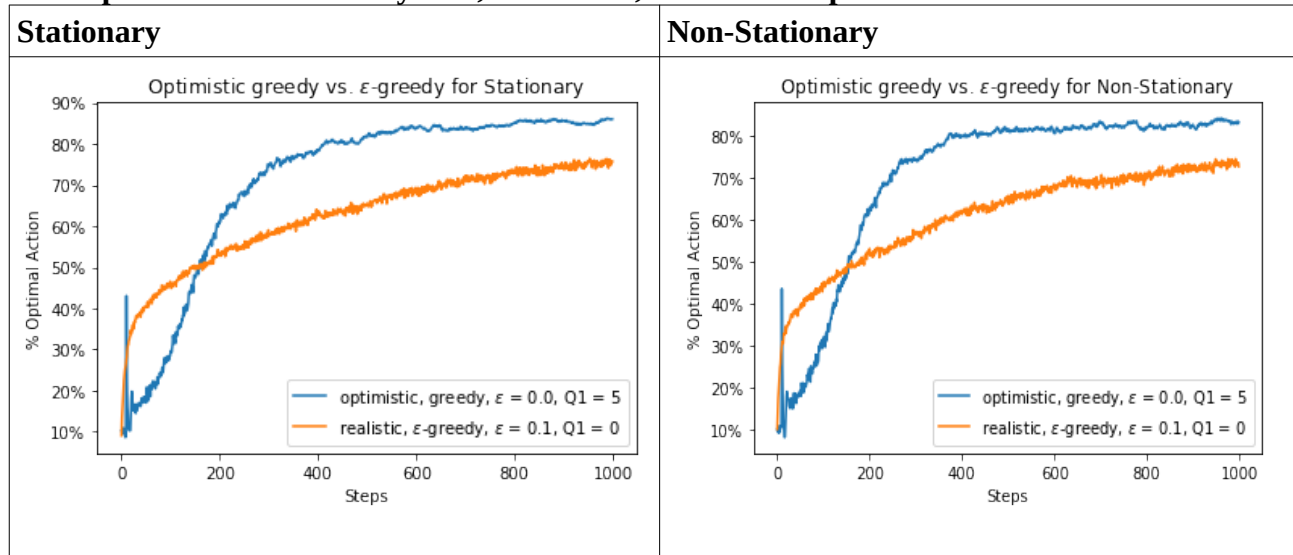


Analysis:

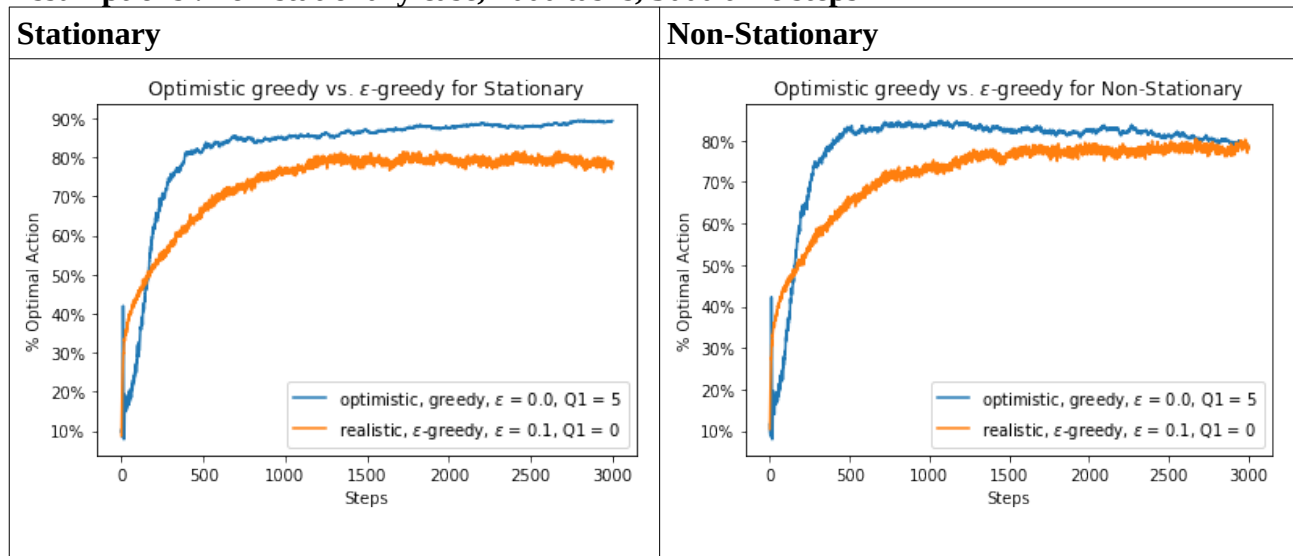
- For both optimal count as well as average reward, constant step size method performs better than sample average method for estimating action-values.
- It can be clearly seen that constant step size method achieves higher average reward as much as 1.4 and also chooses optimal action more frequently 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

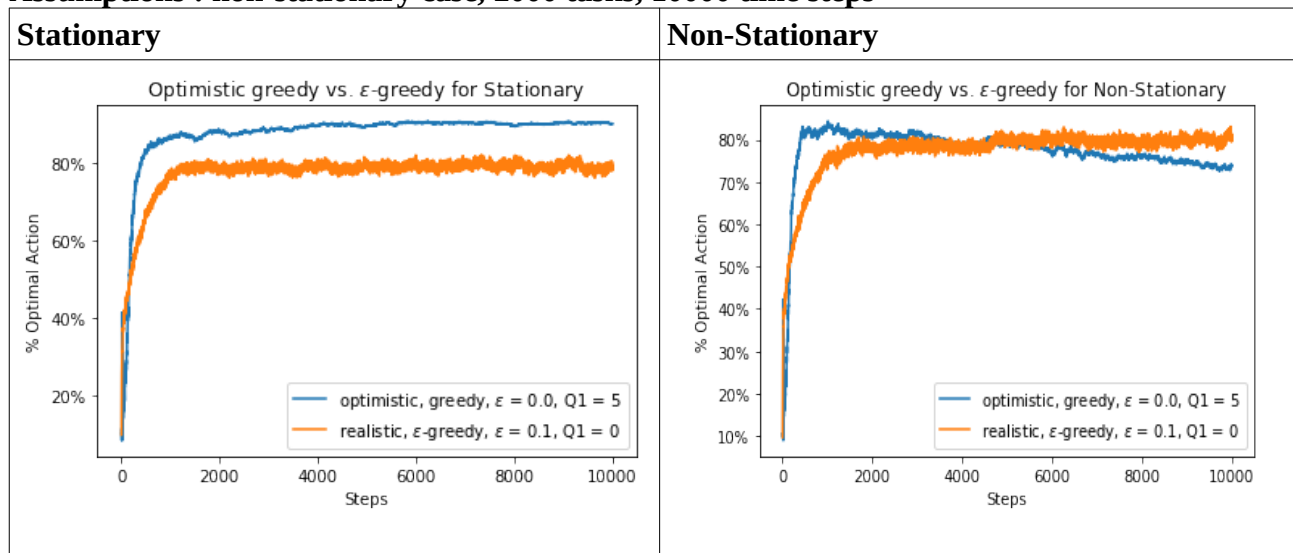
Assumptions : non-stationary case, 2000 tasks, 1000 time steps



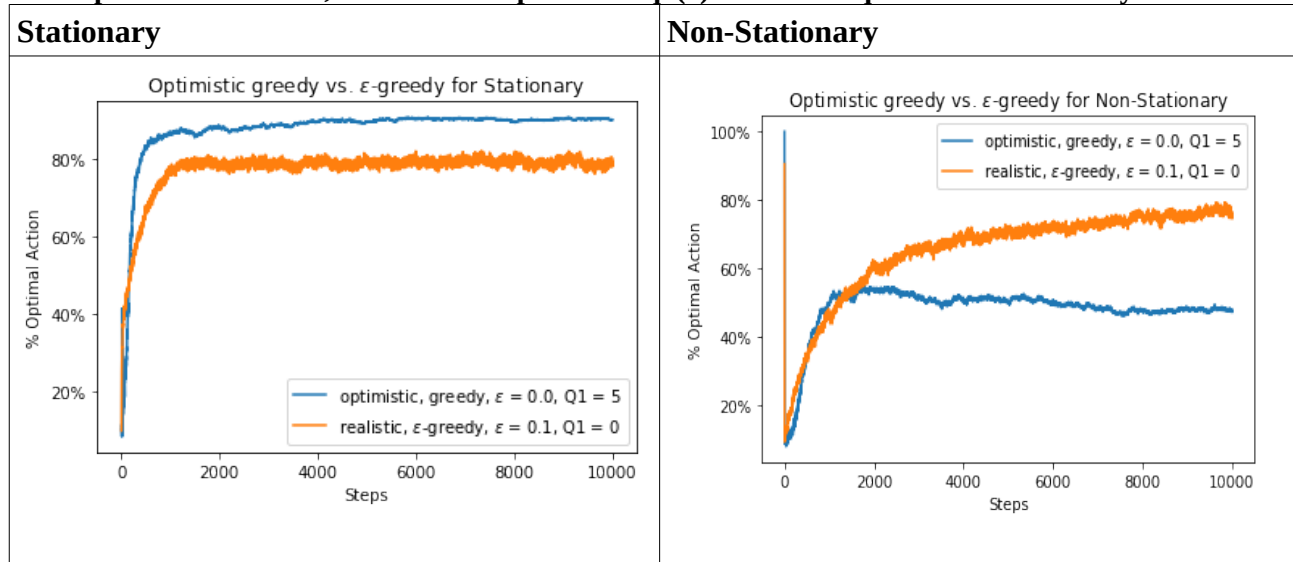
Assumptions : non-stationary case, 2000 tasks, 3000 time steps



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



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



Analysis:

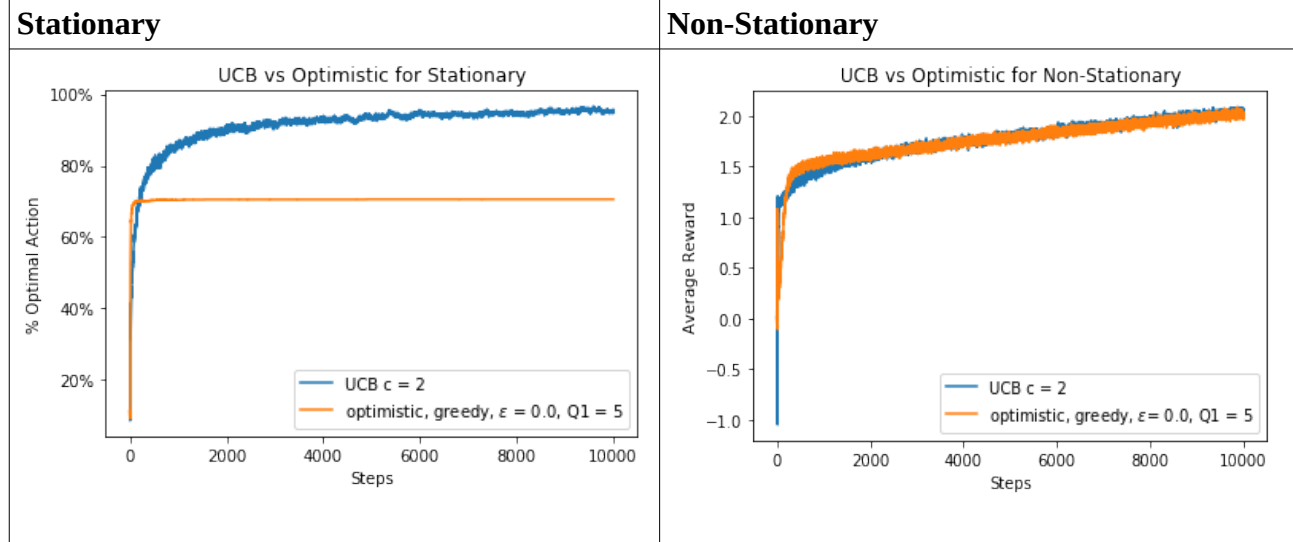
- **Stationary Case :**
 - optimistic-method for stationary case explores more and hence performs poorly initially performs better eventually method than the epsilon-greedy.
- **Non-Stationary Case :**
 - optimistic-method for non-stationary case performs better initially but with more runs or over time, its performance deteriorates.
 - This shows that optimistic-method is not suited for non-stationary problems because its drive for exploration 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

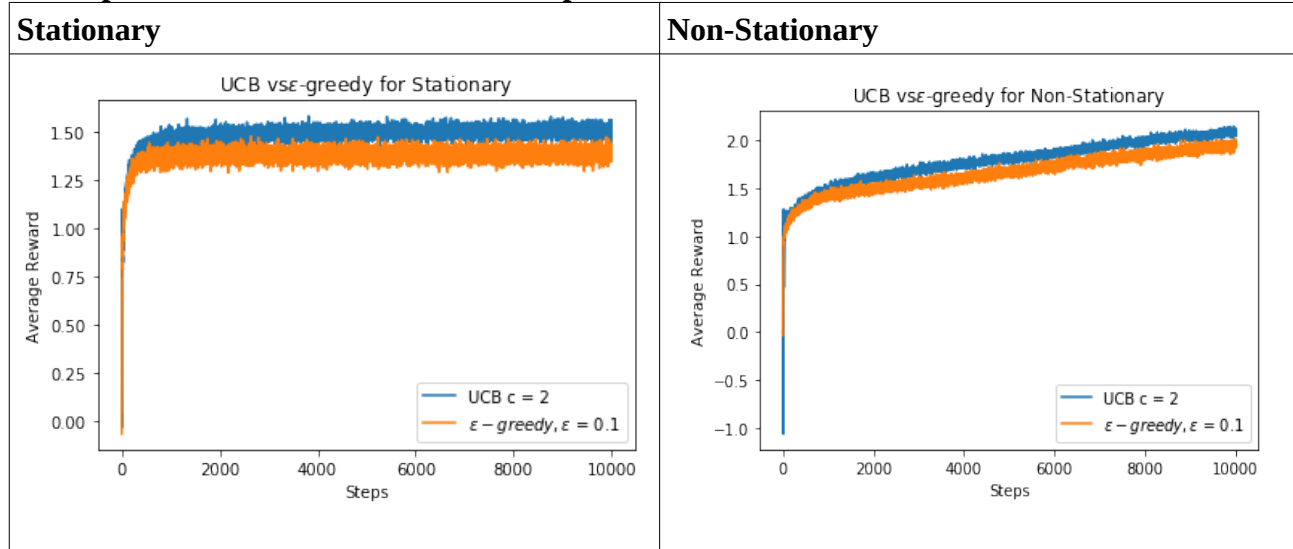
Part 4. UCB vs Optimistic Value

Assumptions : 2000 tasks, 10000 time steps



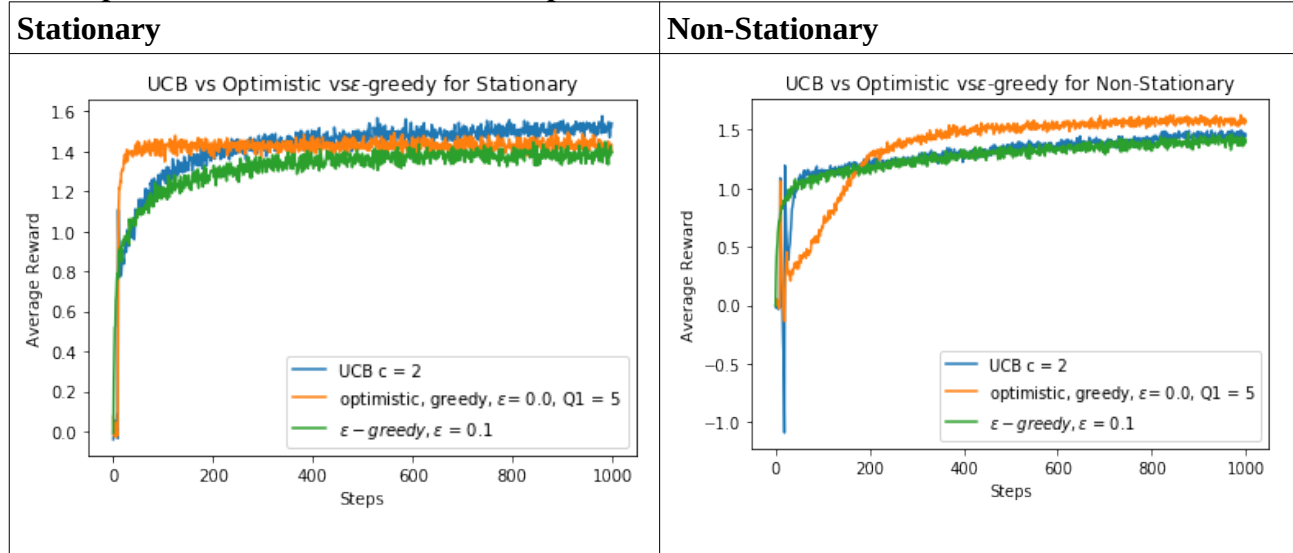
UCB vs E-greedy

Assumptions : 2000 tasks, 10000 time steps

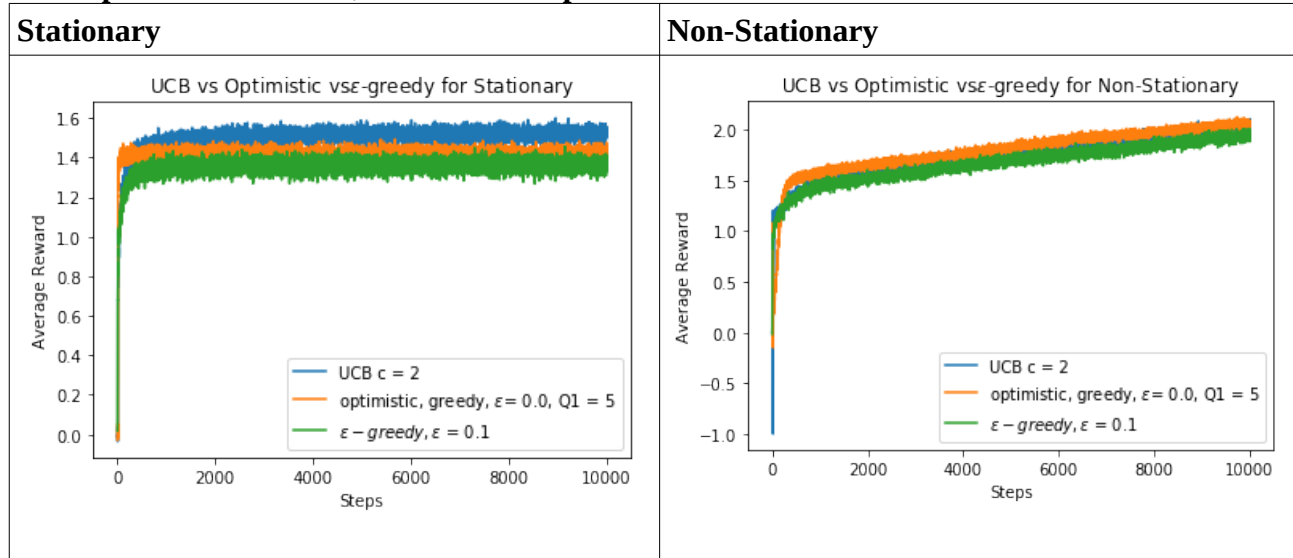


UCB vs Optimistic vs E-greedy

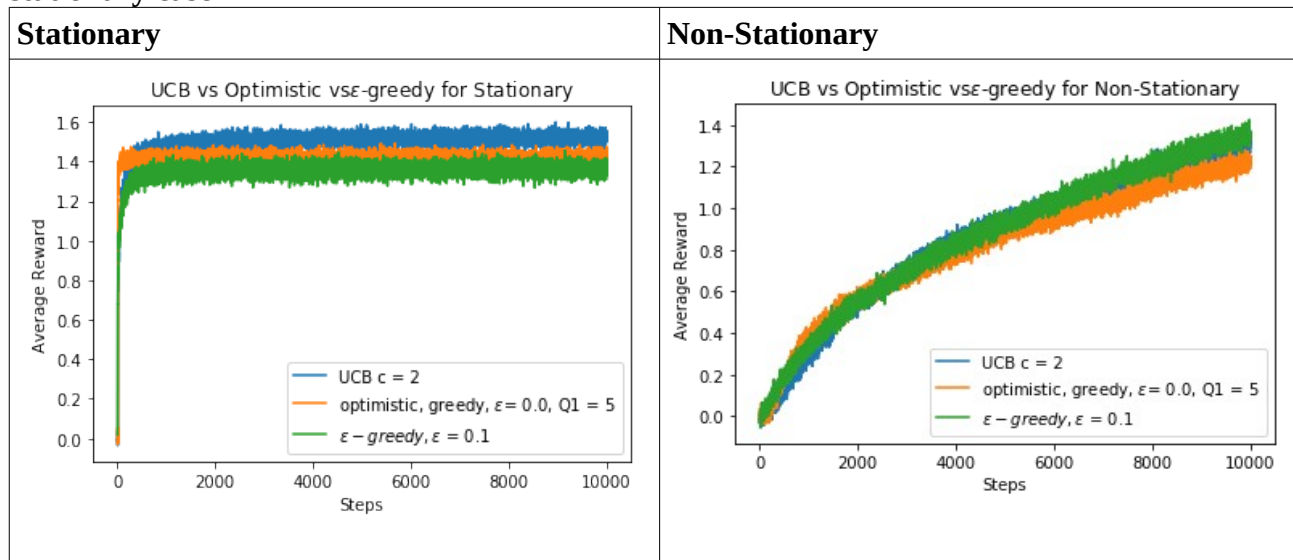
Assumptions : 2000 tasks, 1000 time steps



Assumptions : 2000 tasks, 10000 time steps

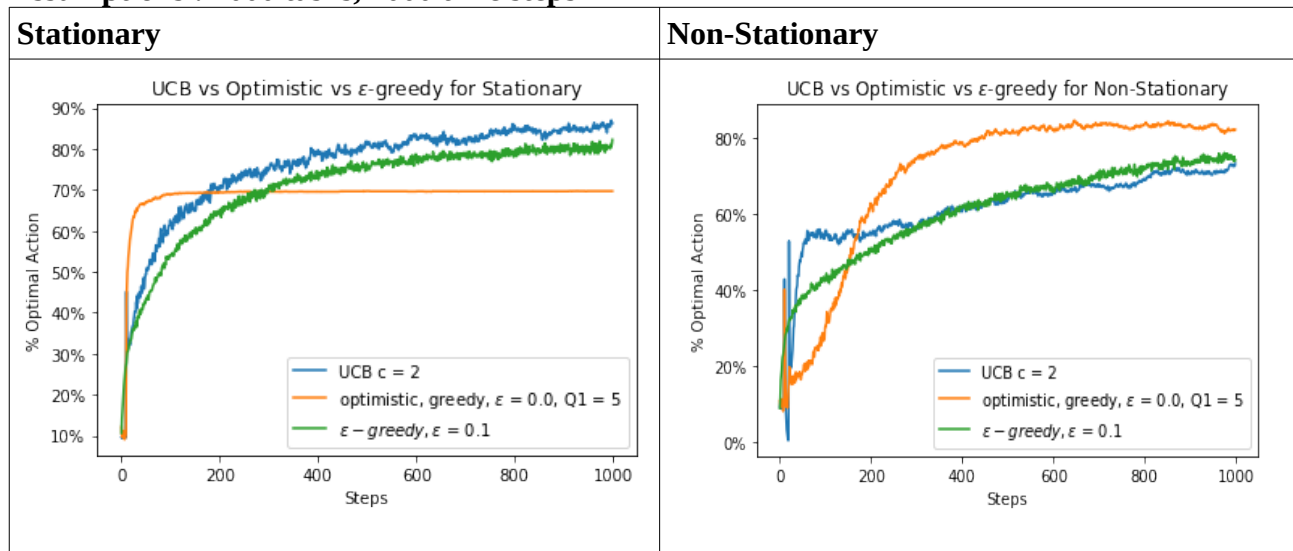


Assumptions : 2000 tasks, 10000 time steps and all $q^*(a)$ start out equal initially in non-stationary case

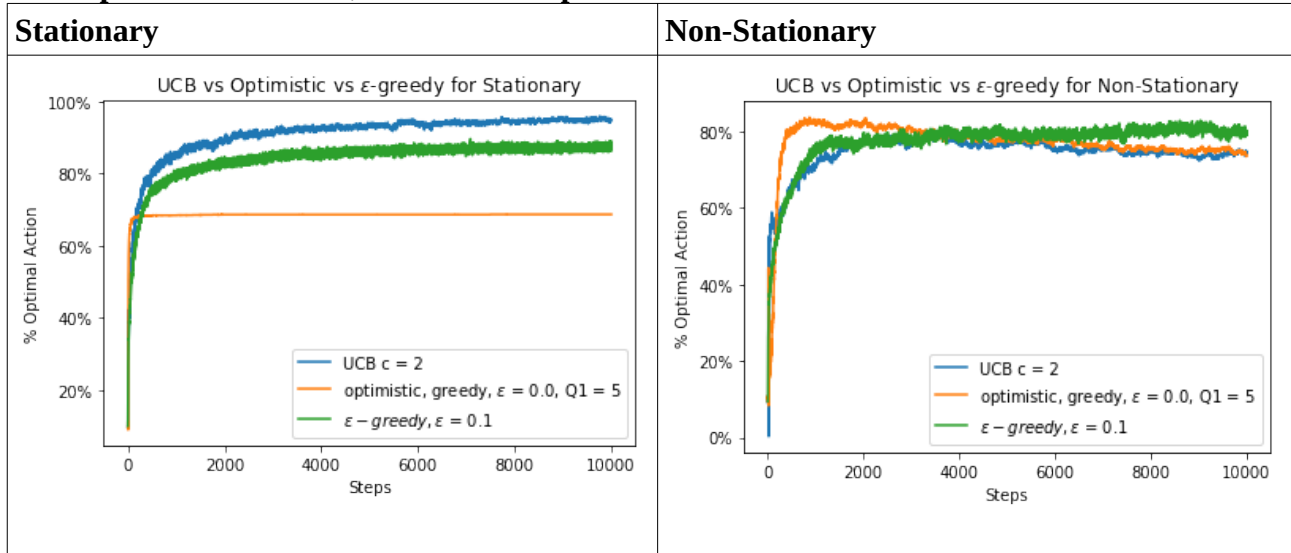


UCB vs Optimistic vs E-greedy (Optimal Actions)

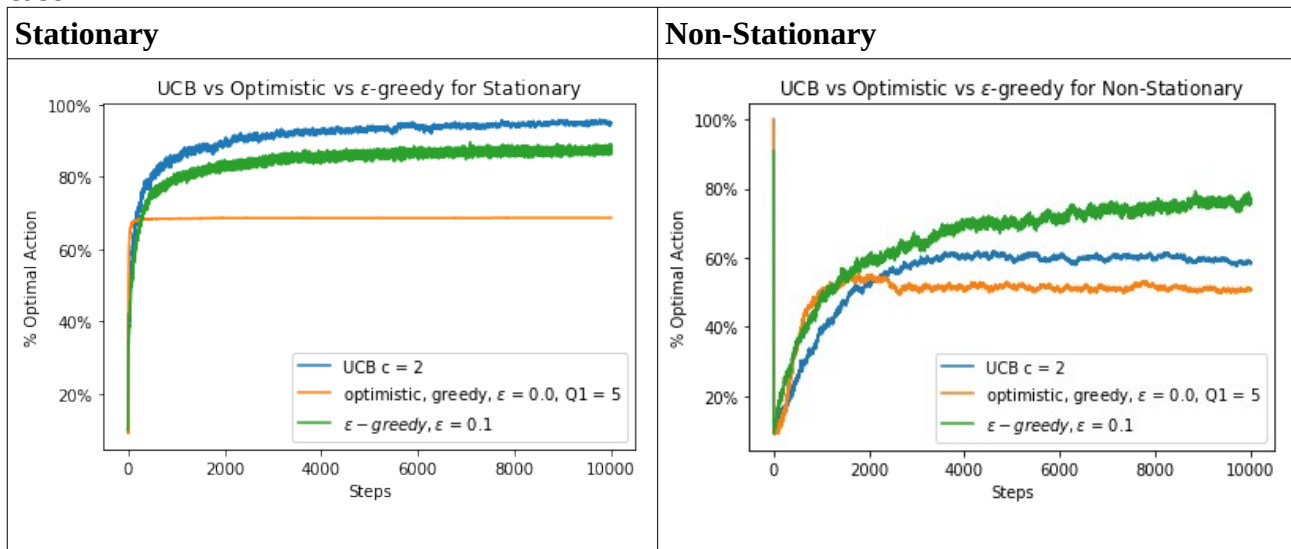
Assumptions : 2000 tasks, 1000 time steps



Assumptions : 2000 tasks, 10000 time steps



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



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 early but UCB outperforms in the long runs. Same is the case for optimal action count.
- Non-Stationary Case:**
 - initially, UCB and optimistic greedy has 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 throughout. This indicates that although it performs good in stationary case but is not a suitable choice for non-stationary case. Overall, UCB does not perform well for non-stationary case when compared to the other two models. Epsilon-greedy balances exploitation and exploration rightly in case of non-stationary case and performs best among all other.