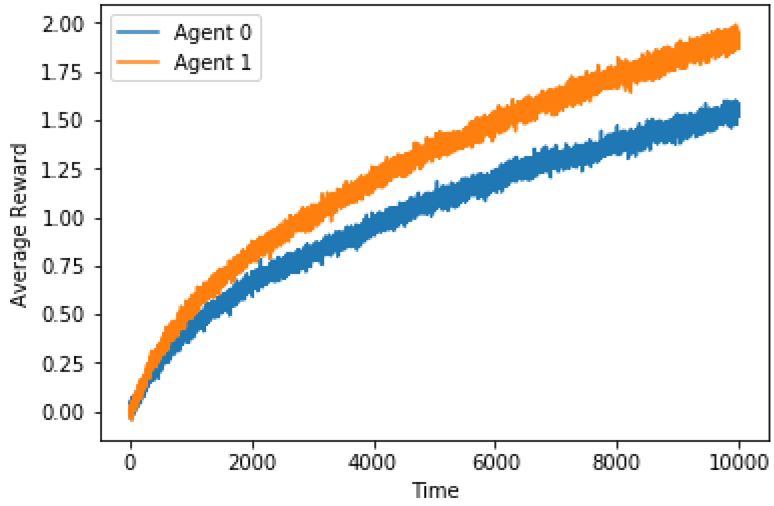
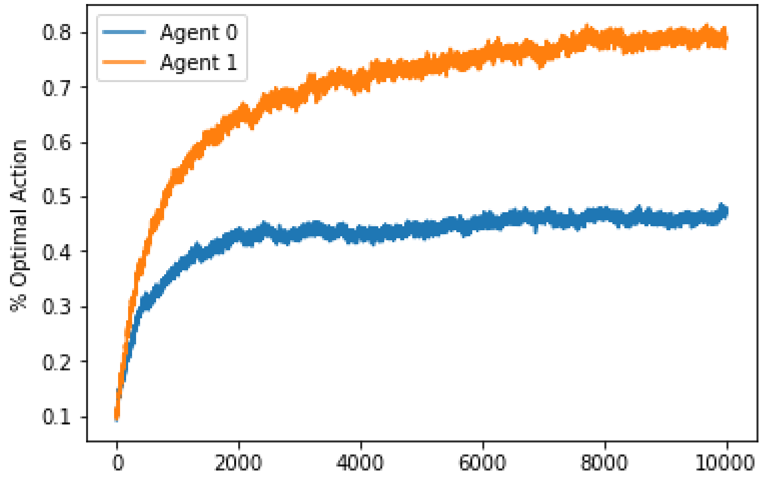
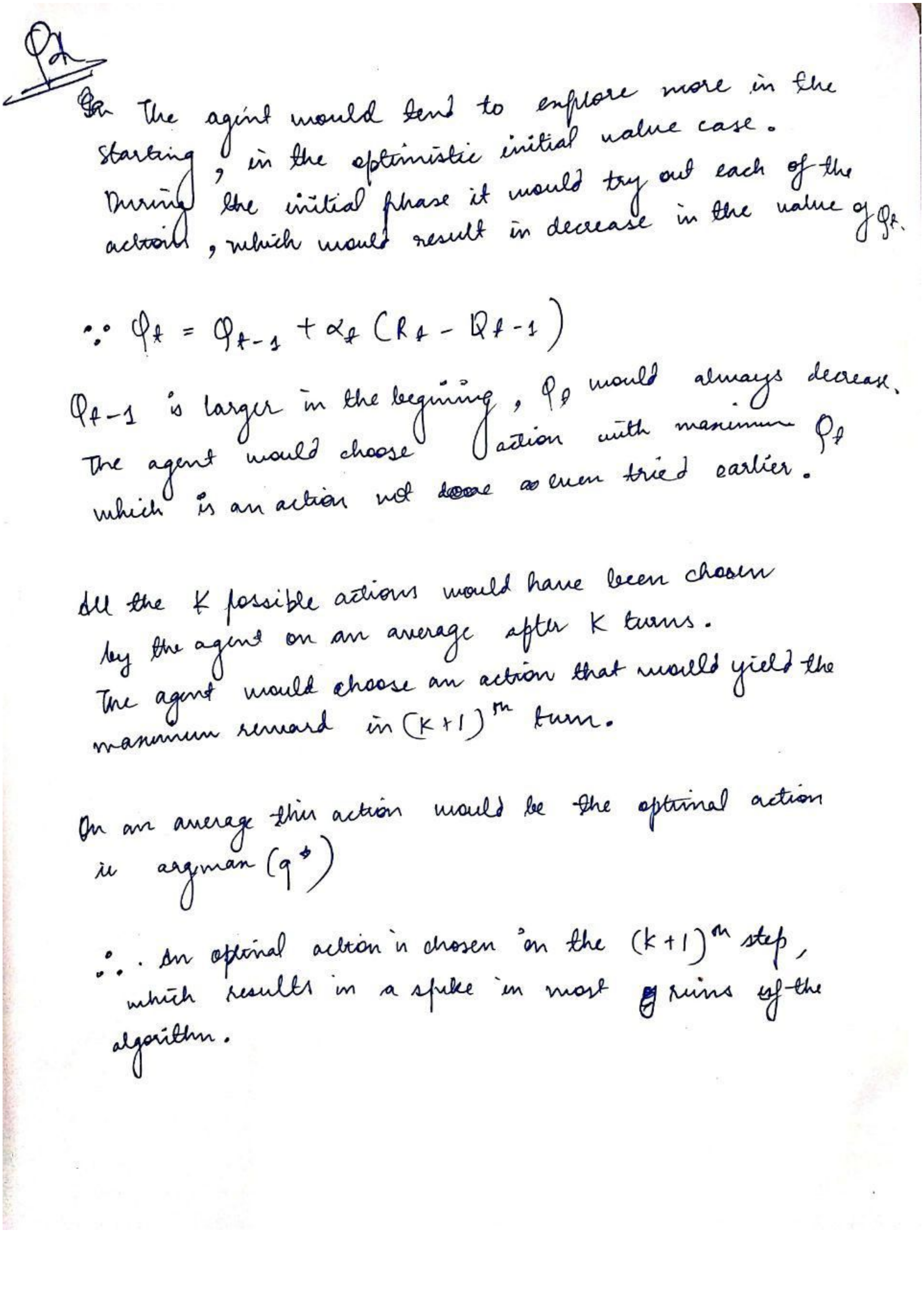
Anushka Bhandari
2016134
RL HomeWork 1
Q1.





Agent 0 : Sample average.
Agent 1 : Constant alpha alpha= 0.1
Since more weight is given to the recent rewards in constant alpha than to the past rewards.
The constant alpha performs much better than the sample average.

Q2i



Q2

~~Q2~~ The agent would tend to explore more in the starting, in the optimistic initial value case. During the initial phase it would try out each of the actions, which would result in decrease in the value of Q_t.

$$\because\ Q_t = Q_{t-1} + \alpha_t (R_t - Q_{t-1})$$

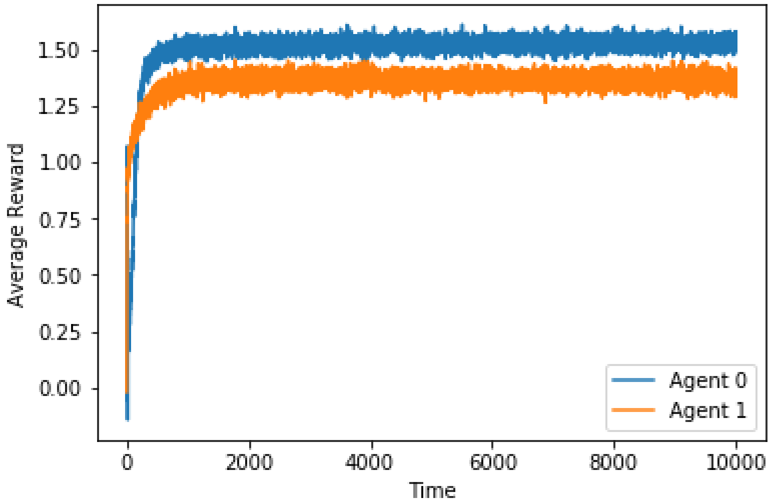
Q_{t-1} is larger in the begining, Q_t would always decrease. The agent would choose action with maximum Q_t which is an action not ~~done~~ as even tried earlier.

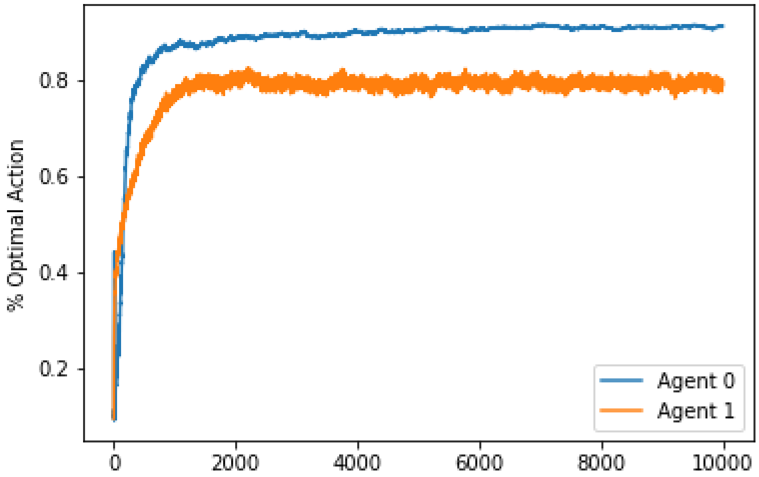
All the K possible actions would have been chosen by the agent on an average after K turns. The agent would choose an action that would yield the maximum reward in $(K+1)^{th}$ turn.

On an average this action would be the optimal action ie argmax(q^*)

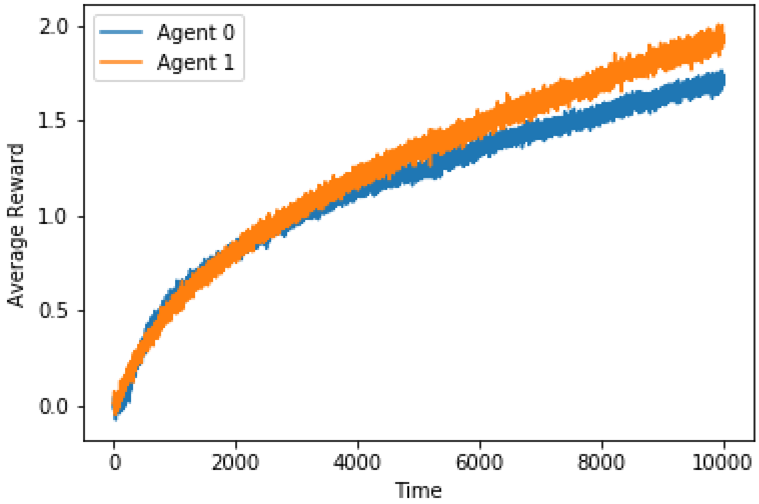
$\therefore$ An optimal action is chosen on the $(k+1)^{th}$ step, which results in a spike in most ~~of~~ runs of the algorithm.

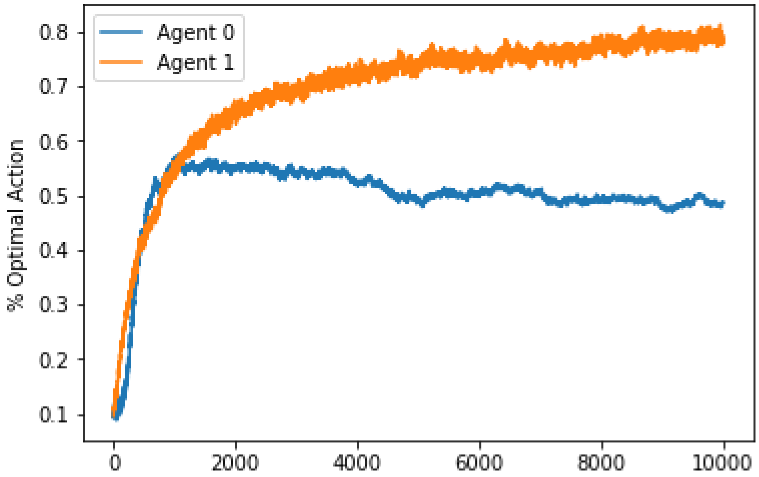
Stationary Environment:





Non Stationary Environment:





Ans 3

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

$$\bar{o}_n = \bar{o}_{n-1} + \alpha (1 - \bar{o}_{n-1}) \quad \text{for } n \geq 0 \text{ with } \bar{o}_0 = 0$$

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

After substituting β_n

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

After substituting $\bar{o}_n$

$$Q_n = Q_{n-1} + \frac{\alpha (R_n - Q_{n-1})}{\bar{o}_{n-1} + \alpha (1 - \bar{o}_{n-1})}$$

For $n = 1$

$$Q_1 = Q_0 + \frac{\alpha (R_1 - Q_0)}{\bar{o}_0 + \alpha (1 - \bar{o}_0)}$$

$\bar{o}_0 = 0$ (given)

$$Q_1 = Q_0 + \frac{\alpha (R_1 - Q_0)}{0 + \alpha (1 - 0)} = Q_0 + \frac{\alpha}{\alpha} (R_1 - Q_0)$$

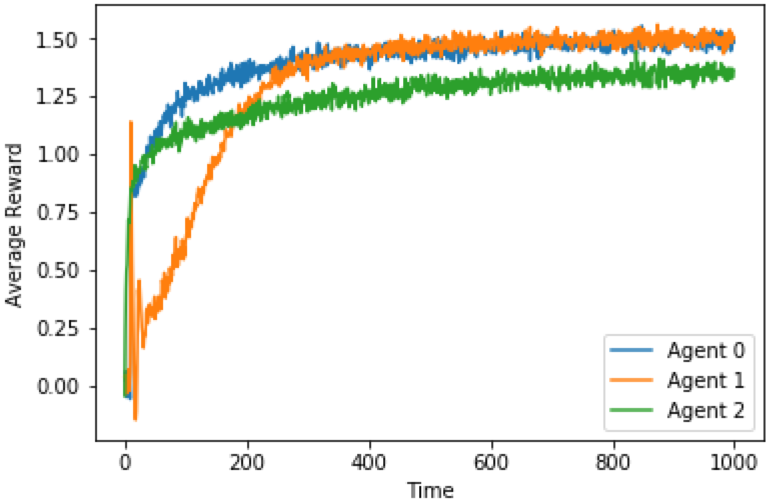
$$= Q_0 + R_1 - Q_0 = R_1$$

∴ Q_n does not have an initial bias

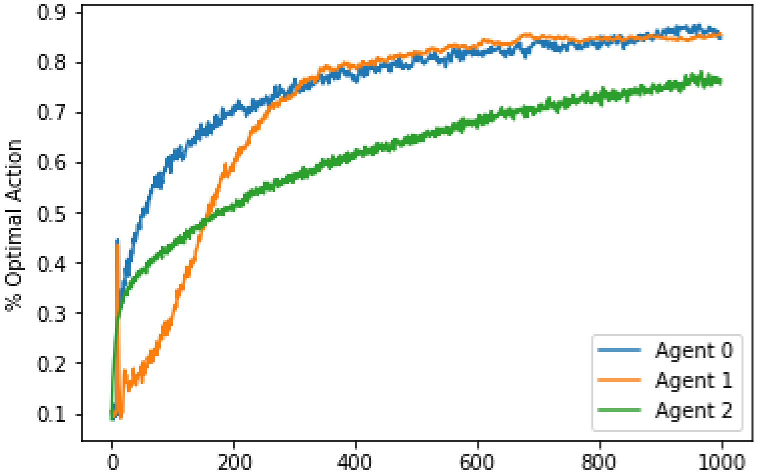
Therefore does not depend on Q_0.

Q4.

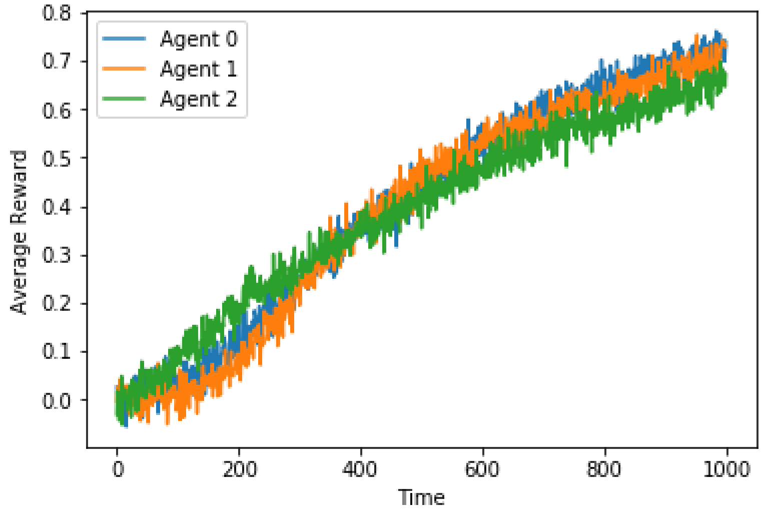
Stationary Environment

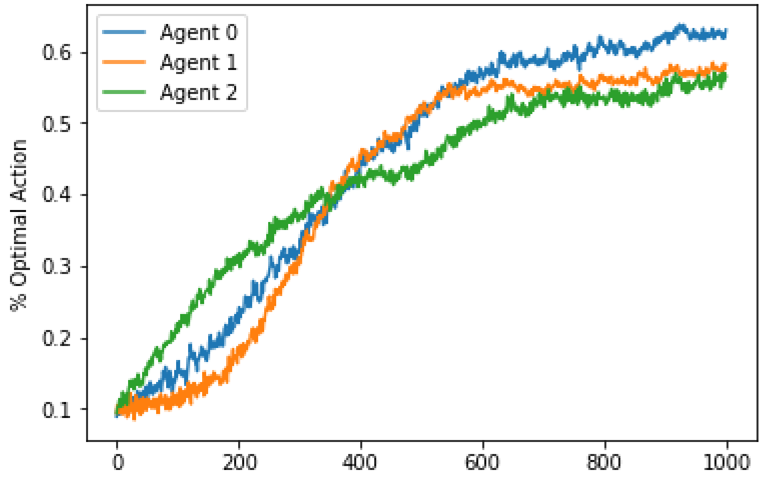


.



Non Stationary





Agent 0 : UCB
Agent 1 : Optimistic Value
Agent 2: E-Greedy
UCB selects among non greedy actions as per their potential of being optimal, while taking into account two points. 1) How close their estimates are to being maximal 2) the uncertainties in those estimates. Hence ensuring that every action is selected once in a while leading to better result. Thats why UCB outperforms Optimistic Value and E Greedy for the stationary case.

UCB performs worse in the early steps in the case of non stationary environment. But in the later time it starts performing better than the rest. In the early steps the uncertainty is more since Nt(a) is close to 0 but after some steps of times, some certainty is achieved. The term of uncertainty that is the uncertainty factor also assumes stationarity.