Q4) The equation of UCB is given by,

$$A_t = \arg\max_a \left[Q_t(a) + c\sqrt{\frac{\ln t}{N_t(a)}} \right],$$

the term $c\sqrt{\frac{\ln(t)}{N_t(a)}}$ helps exploring and exploring wisely rather than the case of ϵ-greedy where actions are chosen randomly. In the beginning time steps $N_t(a)$ is close to 0 indicating that all actions are to be explored. As $N_t(a)$ increases, 'a' action is chosen less. The significance of t is that if $N_t(a)$ does not increase and t increases, we should explore action a because it has not been explored for some time.

After certain amt of time t, the ln t factor saturates leading to more exploitation.

As we can see from the graph, at initial time steps UCB has better optimal action % because of exploration (spikes can also be seen). But in the

later time steps optimistic approach surpasses ucb because after certain amount of exploration ~~ucb~~ optimistic action chooses greedy action (epsilon is 0) unless the factor $N_t(a)$ in ucb does not allow greedy actions to be chosen that frquently.

In fact in case of stationary, ϵ-greedy performs better after a certain number of time steps due to the same reason. ($N_t(a)$ in ucb doesn't allow greedy choices).