Homework -1

Ans 2 Explanation for spikes:

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

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

since Q_{t-1} is larger in beginning, Q_t would always decrease. Therefore the agent chooses the action with maximum Q_t which would be an action it would'nt have tried earlier.

after K turns, the agent on an average would have chosen all the K possible actions. Therefore in the $(K+1)^{th}$ turn agent would choose an action that would yield maximum reward. On an average, this action would be the optimal action, i.e., argmax(q^*). Therefore in most runs of the algorithm, an optimal action is chosen in the $(K+1)^{th}$ step, resulting in a spike.

Ans 3) $\beta_n = \alpha / \bar{o}_n$

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

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

substituting β_n

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

substituting $\bar{o}_n$

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

for $n = 1$

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

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

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

$= Q_0 + \dfrac{\alpha}{\alpha}(R_1 - Q_0)$

$= Q_0 + R_1 - Q_0 = R_1$

$\therefore$ Q_n does not have an initial bias i.e. does not depend on Q_0

Ans 4 looking at the graphs generated for optimal actions for stationary and non-stationary settings, we can see that both optimistic and UCB plots peek earlier as compared to ϵ-greedy method. This happens because both these algorithms do exploration inherently using Q_{t+1}.
Moreover, the UCB algorithm is the ~~fastest~~ quickest to reach the average rewards because it mostly uses exploration in earlier phases. This happens because, the action selected is governed by

$$A_t = \underset{a}{\operatorname{argmax}} \left[Q_t(a) + c\sqrt{\frac{\ln t}{N_t(a)}} \right]$$

in the initial phases $N_t(a)$ is very small making the ~~term~~ the second term larger which promotes exploration as little weight is given to $Q_t(a)$.