Q.2.2

'Consider a k-armed bandit with k=4 actions, denoted as 1, 2, 3 & 4. Consider applying to this problem a bandit algorithm using ε-greedy selection, sample-average action-value estimates, and initial estimates of $Q_1(a) = 0$, for all a. Suppose the initial sequence of actions and rewards is $A_1 = 1$, $R_1 = -1$, $A_2 = 2$, $R_2 = 1$, $A_3 = 2$, $R_3 = -2$, $A_4 = 2$, $R_4 = 2$, $A_5 = 3$, $R_5 = 0$. On some of these timesteps the ε case may have occured, causing an action to be selected at random. On which steps did this definitely occur? On which timesteps could it possibly ~~occur~~ have occured?

A

A_2 & A_5 are definitely exploratory

Any of the other actions could have been exploratory.

Q.2.3

$\varepsilon = 0.01$ will perform better because in both cases as $t \to \infty$ we have $Q_t \to q_*$

The total reward & probability of choosing the optimal action will therefore be 10 times larger in the case of $\varepsilon = 0.1$.

Q.2.1

0.75

Q.2.4

$$Q_{n+1} = (1-\alpha)^n Q_1 + \sum_{i=1}^{n} \alpha (1-\alpha)^{n-i} R_i$$

α_n is not constant

Let $\alpha_0 = 1$, then

$$Q_{n+1} = \prod_{i=1}^{n} (1-\alpha_i) + \sum_{i=1}^{n} \alpha_i R_i \prod_{k=i+1}^{n} (1-\alpha_k)$$

where $\prod_{i=x}^{y} f(i) = 1$ if $x > y$

Q.2.5

Initially, the action value is larger than the mean. Thus, the agent might select a good arm by chance. Then the estimate updates. The estimate will decrease with large probability. Since, it is the beginning phase, some ~~arms~~ worse arms might haven't played and thus their action value haven't been updated therefore might have values larger than good arms. Under the greedy frame, the agent will play & pick the worse arms & update

- Ways to improve - If possible, assign large value to better arms & smaller to worse arms

- Ways to worsen - Action values assigned to worse arms ~~have~~ are larger than arms with higher expected rewards

Q.2-7

There is no dependance of Q_{k+1} on Q_1 for $k>1$ since $\beta_1 = 1$

Now it remains to show that weights in the remaining sum decrease as we look further into the past. That is

$$w_i = \beta_i \prod_{k=i+1}^{n} (1-\beta_k)$$

increases with i for fixed n. For this we observe that

$$\frac{w_{i+1}}{w_i} = \frac{\beta_{i+1}}{\beta_i(1-\beta_{i+1})} = \frac{1}{1-\alpha} > 1 \quad (\alpha < 1)$$

If $\alpha = 1$ then $\beta_t = 1 \ \forall t$

Q-2-8

In the first 10 steps the agent cycles through all the actions because $N_t(a) = 0$, then a is considered maximal.
On the 11th step, the agent will choose most greedily.
It will continue to choose greedily until $\ln(t)$ overtakes $N_t(a)$ for one of the actions, in which case the agents begins to explore again hence reducing rewards

In the long run, $N_t = O(t)$ & $\frac{\ln(t)}{t} \to 1$

So the agent is asymptotically greedy.

2.2-4

Let the two actions be 0 & 1.

$$P(A_t = 1) = \frac{e^{H_t(1)}}{e^{H_t(1)} + e^{H_t(0)}} = \frac{1}{1 + e^{-x}}$$

where $x = H_t(1) - H_t(0)$ is relative preference of 1 over 0

2.2-10

Assume rewards are stationary.
One should always chose the action with the highest reward

In the first case, both action 1 & 2 have expected 0.5. It doesn't matter which is picked.

In the second case, one should run a normal bandit method seperately on each color. The expected reward from identifying the optimal actions in each case is 0.55.