Q-2) For a suboptimal arm $a \in [K]$, we define

$$E[N_a(T)] = E\left[\sum_{t=1}^{T} \mathbb{1}_{A_t = a}\right] = \sum_{t=1}^{T} P(A_t = a)$$

a) Each arm $a \in [K]$ is choosen $m$-times during the ETC exploration phase.

For the remaining $(T - mK)$ rounds, the arm $(\hat{a})$ with the highest empirical average is choosen.

$$\therefore E[N_a(T)] = m + P(\hat{a} = a) \cdot (T - mK)$$

$$\left(\hat{a} = \underset{K}{\operatorname{argmax}}\, \mu_K\right)$$

b) $$P(\hat{a} = a) \leq P(\hat{\mu}_a \geq \hat{\mu}_{a^*})$$

$$\hat{\mu}_a = \frac{1}{m} \sum_{i=1}^{m} r_{a,i}$$

$r_{a,i}$ : reward when using arm $a$ in exploration round $i$. (total $m$ rounds for each arm)

$r_{a,i}$ are drawn independently from reward distribution of arm $a$.

Hoeffding inequality :

$$P(|\hat{\mu}_a - \mu_a| \geq \varepsilon) \leq 2\exp(-2m\varepsilon^2)$$

$\hat{\mu}_a$ → empirical mean of arm $a$.

$\mu_a$ → true mean of arm $a$

To select a wrong arm, ie, $a \neq a^*$, we need $\hat{\mu}_a \geq \hat{\mu}_{a^*}$

$$P(\hat{\mu}_a \geq \hat{\mu}_{a^*}) = P\left((\hat{\mu}_a - \mu_a) + (\mu_a - \mu_{a^*}) + (\mu_{a^*} - \hat{\mu}_{a^*}) \geq 0\right)$$

$$\mu_a - \mu_{a^*} = -\Delta_a$$

$$\Rightarrow P(\hat{\mu}_a \geq \hat{\mu}_{a^*}) = P\left((\hat{\mu}_a - \mu_a) - \Delta_a + (\mu_{a^*} - \hat{\mu}_{a^*}) \geq 0\right)$$

$$= P\left((\hat{\mu}_a - \mu_a) + (\mu_{a^*} - \hat{\mu}_{a^*}) \geq \Delta_a\right)$$

$$\leq P\left(\hat{\mu}_a - \mu_a \geq \frac{\Delta_a}{2}\right) + P\left(\mu_{a^*} - \hat{\mu}_{a^*} \geq \frac{\Delta_a}{2}\right)$$

(Boole's inequality)

Using the Hoeffding ineq. for each term:

$$P\left(\hat{\mu}_a - \mu_a \geq \frac{\Delta_a}{2}\right) \leq \exp\left(-2m\left(\frac{\Delta_a}{2}\right)^2\right)$$

and $P\left(\mu_{a^*} - \hat{\mu}_{a^*} \geq \frac{\Delta_a}{2}\right) \leq \exp\left(-2m\left(\frac{\Delta_a}{2}\right)^2\right)$

$$\therefore P(\hat{\mu}_a \geq \hat{\mu}_{a^*}) \leq \exp\left(-\frac{m\Delta_a^2}{2}\right) + \exp\left(-\frac{m\Delta_a^2}{2}\right)$$

$$= 2\exp\left(-\frac{m\Delta_a^2}{2}\right)$$

c)

Min $2\exp\left(-\frac{m\Delta_a^2}{2}\right)$ when $m \to \infty$.

ie, when we do infinitely many exploration rounds.

But this not possible as we have $T$ rounds in total. and $0 < m \leq \frac{T}{K}$. So to minimize the upper bound, $m = \frac{T}{K}$.