

Regret

$$R_p(T) = T \mu_{a^*} - E \left[ \sum_{t=1}^T R_t \right]$$

$$R_v(T) = \sum_{a=1}^K \Delta_a E[N_a(T)] \rightarrow \text{Regret decomposition.}$$

$$\Delta_a = \mu^* - \mu_a$$

↳ best mean reward amount allocation.

$N_a(T)$ : no. of times arm  $a$  is selected in  $T$  rounds.

$\epsilon$ -greedy strategy:  $\epsilon$ -fraction of  $T$  rounds are exploratory.

During exploitation, Suboptimal arms are chosen only when their mean appears better than the optimal arm's mean.

For exploration:  $E[N_a(T)] = \frac{\epsilon T}{K}$

In  $(1-\epsilon)T$  rounds, the agent exploits.

Suboptimal arms are rarely chosen during exploitation, so

$$E[N_a(T)] \approx 0$$

$$\begin{aligned} \therefore E[N_a(T)] &= E[N_a^{\text{explore}}(T)] + E[N_a^{\text{exploit}}(T)] \\ &\geq E[N_a^{\text{explore}}(T)] = \frac{\epsilon T}{K} \end{aligned}$$

Summing over all  $K-1$  suboptimal arms:

$$\sum_{a \neq a^*} E[N_a(T)] \geq \frac{\epsilon T}{K} (K-1)$$

$\hookrightarrow$  best arm.

The regret is:  $R_V(T) = \sum_{a \neq a^*} \Delta_a E[N_a(T)]$

$\Delta_a \geq \Delta_{\min}$  &  $a \neq a^*$ , we have:

$$\begin{aligned} R_V(T) &\geq \Delta_{\min} \sum_{a \neq a^*} E[N_a(T)] \\ &\geq \Delta_{\min} \frac{\epsilon \cdot T}{K} (K-1) \\ &= \dots \end{aligned}$$

$\text{Q-2})$  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 chosen  $m$ -times during the ETC exploration phase.

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

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

$$(\hat{a} = \arg \max_K \bar{\mu}_K)$$

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)

Hoeffding Inequality :

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

↳ true mean of arm  $a$   
↳ empirical mean of arm  $a$ .

To select a wrong arm, ie,  $a \neq a^*$ , we need

$$\hat{\mu}_a \geq \hat{\mu}_{a^*}$$

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

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

$$\begin{aligned} \Rightarrow P(\hat{\mu}_a \geq \hat{\mu}_{a^*}) &= P((\hat{\mu}_a - \mu_a) - \Delta_a + (\mu_{a^*} - \hat{\mu}_{a^*}) \geq 0) \\ &= P((\hat{\mu}_a - \mu_a) + (\mu_{a^*} - \hat{\mu}_{a^*}) \geq \Delta_a) \\ &\leq P(\hat{\mu}_a - \mu_a \geq \frac{\Delta_a}{2}) + P(\mu_{a^*} - \hat{\mu}_{a^*} \geq \frac{\Delta_a}{2}) \\ &\quad (\text{Boole's inequality}) \end{aligned}$$

Using the Hoeffding Ineq. for each term :

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

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

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

c)  $\text{Min } \text{Zexp}\left(-\frac{m\Delta a^2}{z}\right)$  when  $m \rightarrow \infty$ .  
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,  $m = \frac{T}{K}$ .