Anytime optimal algorithms in stochastic multi-armed bandits - Supplementary Material

Rémy Degenne REMY.DEGENNE@MATH.UNIV-PARIS-DIDEROT.FR
LPMA, Université Paris Diderot

Vianney Perchet VIANNEY.PERCHET@NORMALESUP.ORG
CREST, ENSAE

Table 1. Recall of Notations.

K	Number of arms
$\nu^{(k)}$	Distribution of arm k
$\mu^{(k)}$	Expectation of arm k
$\mu^* = \mu^{(1)}$	Expectation of the best arm
$X_r^{(k)}$	Reward of arm k when observed for the r^{th} time
$\overline{X}_r^{(k)}$	Empirical mean of arm k after r observations. $\overline{X}_r^{(k)} = \frac{1}{r}\sum_{u=1}^{r} X_u^{(k)}$
Δ_k	gap between the means of arm k and of the best arm. $\Delta_k = \mu^* - \mu^{(k)}$
$\Delta_{\min}/\Delta_{\max}$	smallest/largest of the positive gaps
$T^{(k)}(t)$	Number of pulls of arm k during the t first stages
$\mathbb{E}R_t$	Expected regret after t stages
$\overline{\log}(x)$	$\max(1, \log(x))$
$\widehat{\log}(x)$	$\max(0, \log(x))$

1. Single-pull UCB2

UCB2 (Auer et al., 2002), reproduced here as Algorithm 1, was introduced with the unusual particularity for UCB-like algorithms that it requires a chosen arm to be pulled not once but an exponentially increasing number of times. The advantage of this algorithm over UCB1 is that it enjoys a bound for the expected regret at time t with a leading factor $\sum_{k,\Delta_k>0} \frac{\log(t\Delta_k^2)}{\Delta_k}$, which leads to a distribution independent bound proportional to $\sqrt{Kt\log K}$. This is an improvement over the $\sqrt{Kt\log t}$ bound of UCB since $K \ll t$.

Lemma 1 (Theorem 2 of (Auer et al., 2002))**.** *An upper bound for the expected regret of UCB2 is*

$$\mathbb{E}R_t \leq \sum_{k,\Delta_k>0} \left(\frac{(1+\alpha)(1+4\alpha)\overline{\log}(2et\Delta_k^2)}{2\Delta_k} + \frac{c_\alpha}{\Delta_k} \right),$$

where c_α varies only with α and goes to infinity when α goes to 0.

Algorithm 1 UCB2.

Input $\alpha > 0$.
For all $k \in \{1, \dots, K\}$, initialize $r_k = 0, s_k = 1$.
Pull each arm once.
for $t \geq 1$ **do**
  Select k that maximizes $\overline{X}_{s_k}^{(k)} + \sqrt{\frac{(1+\alpha)\log(\frac{et}{s_k})}{2s_k}}$.
  Pull the arm k for $\tau(r_k+1) - \tau(r_k)$ stages, with $\tau(x) = \lceil (1+\alpha)^x \rceil$.
  $s_k \leftarrow \tau(r_k+1)$.
  $r_k \leftarrow r_k + 1$.
end for

We show that the block structure of UCB2 is only a convenience for the proof. We introduce a single-pull variant of UCB2 (Algorithm 2) that removes the block structure and we prove a similar upper bound.

Algorithm 2 single-pull UCB2.

Input $\alpha > 0$.
For all $k \in \{1, \dots, K\}$, initialize $s_k = 1$.
Pull each arm once.
for $t \geq 1$ **do**
  Pull arm k that maximizes $\overline{X}_{s_k}^{(k)} + \sqrt{\frac{(1+\alpha)\overline{\log}(\frac{et}{s_k})}{2s_k}}$.
  $s_k \leftarrow s_k + 1$.
end for

Theorem 1. *The expected regret of single-pull UCB2 satisfies*

$$\mathbb{E}R_t \leq \sum_{k,\Delta_k>0} \left(\frac{(1+80\alpha)\overline{\log}(2et\Delta_k^2)}{2\Delta_k} + \frac{C_\alpha}{\Delta_k} \right),$$

where C_α is a function of α.

Proof. This proof follows broadly the one of UCB2 (Auer et al., 2002).

Let $\hat{r}_k$ be the largest integer such that $\hat{r}_k \leq \frac{(1+80\alpha)\overline{\log}(2et\Delta_k^2)}{2\Delta_k^2} + 1$.

$$T^{(k)}(t) \leq 1 + \sum_{r\geq 1} \mathbb{I}_{\{k \text{ pulled more than } r \text{ times}\}}$$
$$\leq \hat{r}_k + \sum_{r\geq \hat{r}_k} \mathbb{I}_{\{k \text{ pulled more than } r \text{ times}\}} .$$

We define $\delta > 0$ and $\epsilon_{u,s} = \sqrt{\frac{(1+\alpha)\overline{\log}(\frac{eu}{s})}{2s}}$. Now consider the following implications, where we use again that when a suboptimal arm is played the optimal arm is underestimated or a suboptimal arm is overestimated:

k has been pulled more than r times
$\Rightarrow$ k was pulled once when it was pulled r times,
at a time t_k, when * was pulled s times ,
$\Rightarrow \exists s \geq 1, \exists t_k \geq r+s, \overline{X}_r^{(k)} + \epsilon_{t_k,r} \geq \overline{X}_s^* + \epsilon_{t_k,s}$,
$\Rightarrow \exists s \geq 1, \exists t' \geq r+s, \overline{X}_s^* + \epsilon_{t',s} \leq \mu^* - \delta\frac{\Delta_k}{2}$
or $\exists t_k \geq r, \overline{X}_r^{(k)} + \epsilon_{t_k,r} \geq \mu^* - \delta\frac{\Delta_k}{2}$,
$\Rightarrow \exists s \geq 1, \overline{X}_s^* + \epsilon_{r+s,s} \leq \mu^* - \delta\frac{\Delta_k}{2}$
or $\overline{X}_r^{(k)} + \epsilon_{t,r} \geq \mu^* - \delta\frac{\Delta_k}{2}$,

where we used in the last inequality that for a given s the function $f : u \mapsto \epsilon_{u+s,s} = \sqrt{\frac{(1+\alpha)\overline{\log}(\frac{eu}{s}+e)}{2s}}$ is increasing.

We get an upper bound for $\mathbb{E}T^{(k)}(t)$,

$$\mathbb{E}T^{(k)}(t) \leq \hat{r}_k + \sum_{r\geq \hat{r}_k} \mathbb{P}\{\overline{X}_r^{(k)} + \epsilon_{t,r} \geq \mu^* - \delta\frac{\Delta_k}{2}\}$$
$$+ \sum_{r\geq \hat{r}_k} \mathbb{P}\{\exists s \geq 1, \overline{X}_s^* + \epsilon_{r+s,s} \leq \mu^* - \delta\frac{\Delta_k}{2}\} .$$

Since $\hat{r}_k \leq \frac{(1+80\alpha)\overline{\log}(2et\Delta_k^2)}{2\Delta_k^2} + 1$, with lemmas 2 and 3 with $\delta = \alpha$ and $\eta = \sqrt{\frac{1+\alpha}{2}} - 1$, we get

$$\mathbb{E}T^{(k)}(t) \leq \frac{(1+80\alpha)\overline{\log}(2et\Delta_k^2)}{2\Delta_k^2} + 1 + \frac{32}{\alpha^2\Delta_k^2}$$
$$+ \frac{2(1+\alpha)^{3/2}}{\alpha^2\Delta_k^2 \log(\frac{1+\alpha}{2})} .$$

Hence the result.

□

1.1. Experiments

The reward variables used are all Gaussian with variance $\sigma^2 = 1/2$. While the unique best arm will always have mean 0, the gaps between this arm and the 9 suboptimal amrs are the main parameters influencing the behaviour of the algorithms and depend on the experiment.

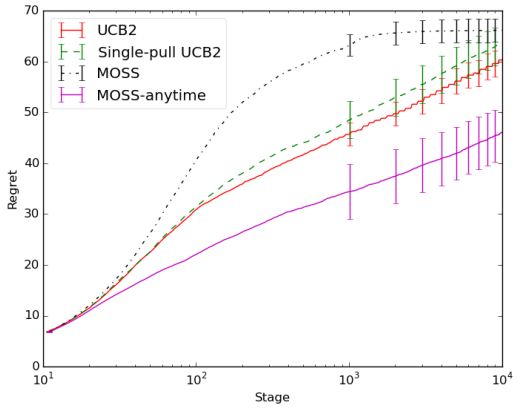

Figure 1. Regrets of the algorithms in the equal gaps case, averaged over 100 runs.

In the first case, reported in Figure 1, all 9 suboptimal arms have the same gap $\Delta = \sigma$.

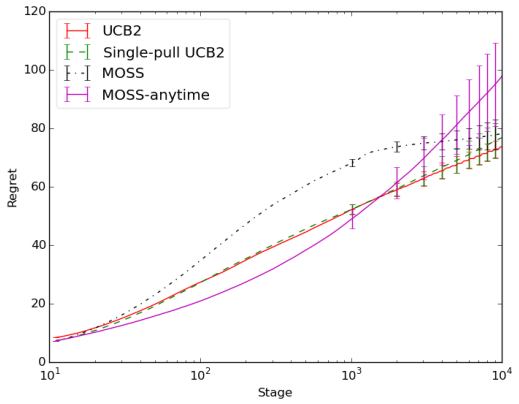

Figure 2. Regrets of the algorithms in the increasing gaps case, averaged over 800 runs.

Figure 2 shows the results of an experiment with increasing gaps: the 9 suboptimal arms have gaps increasing linearly between σ and 3σ.

2. FULL INFORMATION

Theorem 2. *The expected regret of FTL in the full information setting with K arms with equal gaps verifies for $t \geq 1$,*

$$\mathbb{E}R_t \leq \frac{2}{\Delta}(2 + \log(K-1)) \,.$$

Proof. Let $\delta \in (0, \Delta)$. The value of δ will be chosen later. The expected regret can be bounded as

$$\begin{aligned}\mathbb{E}R_t &= \Delta \mathbb{E}[\sum_{s=1}^{t} \mathbb{I}_{\{I_s \neq 1\}}] \\ &\leq \Delta \mathbb{E}[\sum_{s=1}^{\infty} \mathbb{I}_{\{I_s \neq 1\}}] \\ &= \Delta \sum_{s=1}^{\infty} \mathbb{P}\{\exists k \in [2,K], \overline{X}_s^{(k)} > \overline{X}_s^{(1)}\} \,.\end{aligned}$$

The next inequality uses the following argument: if the algorithm pulls a suboptimal arm then either the optimal arm was underestimated or one of the suboptimal arms was overestimated.

$$\begin{aligned}\mathbb{E}R_t \leq \Delta \sum_{s=1}^{\infty} \Big(&\mathbb{P}\{\overline{X}_s^{(1)} \leq \mu^{(1)} - \delta\} + \\ &\mathbb{P}\{\exists k \in [2,K], \overline{X}_s^{(k)} > \mu^{(1)} - \delta\} \Big) \,.\end{aligned}$$

By Hoeffding's inequality,

$$\begin{aligned}\Delta \sum_{s=1}^{\infty} \mathbb{P}\{\overline{X}_s^{(1)} \leq \mu^{(1)} - \delta\} &\leq \Delta \sum_{s=1}^{\infty} \exp(-2s\delta^2) \\ &\leq \frac{\Delta}{2\delta^2} \,,\end{aligned}$$

and

$$\begin{aligned}&\mathbb{P}\{\exists k \in [2,K], \overline{X}_s^{(k)} > \mu^{(1)} - \delta\} \\ &\quad = 1 - \mathbb{P}\{\forall k \in [2,K], \overline{X}_s^{(k)} \leq \mu^{(1)} - \delta\} \\ &\quad = 1 - \prod_{k=2}^{K} \mathbb{P}\{\overline{X}_s^{(k)} \leq \mu^{(1)} - \delta\} \\ &\quad = 1 - \prod_{k=2}^{K} (1 - \mathbb{P}\{\overline{X}_s^{(k)} > \mu^{(1)} - \delta\}) \\ &\quad = 1 - \prod_{k=2}^{K} (1 - \mathbb{P}\{\overline{X}_s^{(k)} - \mu^{(1)} + \Delta > \Delta - \delta\}) \\ &\quad \leq 1 - (1 - \exp(-2s(\Delta-\delta)^2))^{K-1} \,.\end{aligned}$$

For $n \in \mathbb{N}$, let $f_n(s) = 1 - (1 - \exp(-2s(\Delta-\delta)^2))^n$. The bound on the regret can be written

$$\mathbb{E}R_t \leq \frac{\Delta}{2\delta^2} + \Delta \sum_{s=1}^{\infty} f_{K-1}(s) \,.$$

For $n \geq 1$, $f_n'(s) = 2n(\Delta-\delta)^2(-f_n(s) + f_{n-1}(s))$. We use it to compute the integral of f_n,

$$\begin{aligned}\int_0^{+\infty} f_n(s)ds &= -\frac{1}{2n(\Delta-\delta)^2} \int_0^{+\infty} f_n'(s)ds \\ &\quad + \int_0^{+\infty} f_{n-1}(s)ds \\ &= \frac{1}{2n(\Delta-\delta)^2} + \int_0^{+\infty} f_{n-1}(s)ds \\ &= \sum_{k=1}^{n} \frac{1}{2k(\Delta-\delta)^2} \,.\end{aligned}$$

Finally we can bound the regret as

$$\begin{aligned}\mathbb{E}R_t &\leq \frac{\Delta}{2\delta^2} + \Delta \sum_{s=1}^{\infty} f_{K-1}(s) \\ &\leq \frac{\Delta}{2\delta^2} + \Delta \int_0^{\infty} f_{K-1}(s)ds \\ &= \frac{\Delta}{2\delta^2} + \frac{\Delta}{2(\Delta-\delta)^2} \sum_{k=1}^{K-1} \frac{1}{k} \,.\end{aligned}$$

Let $S_K = \sum_{k=1}^{K-1} \frac{1}{k}$. The last expression is minimal for $\delta = \frac{\Delta}{1+S_K^{1/3}}$ and gives the inequality

$$\mathbb{E}R_t \leq \frac{1}{2\Delta}(1 + S_K^{1/3})^3 \,.$$

With $\delta = \frac{\Delta}{2}$, we get

$$\begin{aligned}\mathbb{E}R_t &\leq \frac{2}{\Delta}(1 + \sum_{k=1}^{K-1} \frac{1}{k}) \\ &\leq \frac{2}{\Delta}(2 + \log(K-1)) \,.\end{aligned}$$

□

This upper bound in $\frac{\log K}{\Delta}$ naturally poses the question of a possible matching lower bound, with the same dependency in K. The question remains open.

Theorem 3. *FTL in the full information setting with equal gaps verifies verifies for $t \geq 1$,*

$$\sup_{\Delta} \mathbb{E}R_t \leq \sqrt{2t(2 + \log(K-1))} \,.$$

Proof. First remark that the regret of any algorithm up to time t with gaps Δ is bounded by Δt. Then for FTL, for any $\Delta > 0$,

$$\begin{aligned}\mathbb{E}R_t &\leq \min\{\frac{2}{\Delta}(2 + \log(K-1)), \Delta t\} \\ &\leq \sqrt{2t(2 + \log(K-1))} \,.\end{aligned}$$

□

3. MOSS-anytime

Algorithm 3 MOSS-anytime.

Input $\alpha > 0$.
Pull each arm once.
For $1 \leq k \leq K$, set $s_k = 1$.
for $t \geq 1$ **do**
  Pull arm k that maximizes
    $\overline{X}_{s_k}^{(k)} + \sqrt{\frac{(1+\alpha)}{2} \frac{\max(0, \log(\frac{t}{K s_k}))}{s_k}}$.
  $s_k \leftarrow s_k + 1$.
end for

Theorem 4 (Upper bounds for MOSS-anytime)**.** *In the K arms bandit setting, for $\alpha = 1.35$, the expected regret of MOSS-anytime verifies*

$$\mathbb{E}R_t \leq 75 \frac{K}{\Delta_{\min}} \left(\overline{\log}(\frac{2t\Delta_{\min}^2}{K}) + 1 \right) + \Delta_{\max}$$

and

$$\mathbb{E}R_t \leq 113\sqrt{Kt} + \Delta_{\max} .$$

Proof of Theorem 4. The beginning of this proof uses a decoupling of the arms inspired from the proof of the upper bounds of MOSS (Audibert & Bubeck, 2010) but then departs from it to control the probabilities of the suboptimal pulls in an anytime fashion. In this second part, the critical arguments are well chosen relative weights for the different sources of regret, the use of Hoeffding's maximal inequality and a peeling technique.

Let k_0 be an integer in $[1, K]$ that will be chosen later. Let $\epsilon_{t,s} = \sqrt{\frac{(1+\alpha)}{2} \frac{\max(0, \log(\frac{t}{Ks}))}{s}}$ be the exploration term of the algorithm and $\delta > 0$ a constant to be chosen later. For $k \in \{k_0 + 1, \dots, K\}$, we define $z_k = \mu^* - \delta \frac{\Delta_k}{2}$, $z_{k_0} = +\infty$ and $z_{K+1} = 0$. We will consider the smallest value possibly taken by the index of the optimal arm after a time t,

$$A_t^* = \min_{s \geq 1} \min_{u \geq t} \overline{X}_s^* + \epsilon_{u,s} ,$$

and after r pulls of suboptimal arms,

$$B_r^* = \min_{s \geq 1} \min_{u \geq r+s} \overline{X}_s^* + \epsilon_{u,s} .$$

Step 1: separating the events that the optimal arm is underestimated or that a suboptimal arm is overestimated. We allow a regret of Δ_{k_0} at each stage,

$$\mathbb{E}R_t \leq t\Delta_{k_0} + \mathbb{E}[\sum_{k=k_0+1}^{K} (\Delta_k - \Delta_{k_0}) T^{(k)}(t)] .$$

We will bound the regret incurred for $k > k_0$. We note π_s the arm pulled at time s.

$$\begin{aligned}
&\mathbb{E}R_t - t\Delta_{k_0} \\
&\quad \leq \mathbb{E}[\sum_{k=k_0+1}^{K} (\Delta_k - \Delta_{k_0}) \sum_{s \geq 0} \mathbb{I}_{\{k \text{ pulled at time } s\}}] \\
&\quad \leq \mathbb{E}[\sum_{k=k_0+1}^{K} \sum_{j=k_0}^{K} (\Delta_k - \Delta_{k_0}) \sum_{s \geq 0} \mathbb{I}_{\{\pi_s = k, A_s^* \in [z_{j+1}, z_j)\}}] \\
&\quad \leq \sum_{s \geq 0} \mathbb{E}[\sum_{j=k_0}^{K} \sum_{k=k_0+1}^{j} (\Delta_k - \Delta_{k_0}) \mathbb{I}_{\{\pi_s = k, A_s^* \in [z_{j+1}, z_j)\}}] \\
&\quad + \sum_{s \geq 0} \mathbb{E}[\sum_{j=k_0}^{K} \sum_{k=j+1}^{K} (\Delta_k - \Delta_{k_0}) \mathbb{I}_{\{\pi_s = k, A_s^* \in [z_{j+1}, z_j)\}}]
\end{aligned}$$

Here we get two sums: one quantifying the event that the optimal arm is underestimated (against values depending on the arms) and a second one quantifying the event that one of the suboptimal arms is pulled even if the optimal arm is not underestimated.

Step 2: bounding the probability that the optimal arm is underestimated.

$$\begin{aligned}
&\sum_{j=k_0}^{K} \sum_{k=k_0+1}^{j} (\Delta_k - \Delta_{k_0}) \mathbb{I}_{\{\pi_s = k, A_s^* \in [z_{j+1}, z_j)\}} \\
&\leq \sum_{j=k_0}^{K} \sum_{k=k_0+1}^{j} (\Delta_j - \Delta_{k_0}) \mathbb{I}_{\{\pi_s = k, A_s^* \in [z_{j+1}, z_j)\}} \\
&= \sum_{j=k_0}^{K} (\Delta_j - \Delta_{k_0}) \mathbb{I}_{\{A_s^* \in [z_{j+1}, z_j)\}} \sum_{k=k_0+1}^{j} \mathbb{I}_{\{\pi_s = k\}} \\
&\leq \sum_{j=k_0}^{K} (\Delta_j - \Delta_{k_0}) \mathbb{I}_{\{A_s^* \in [z_{j+1}, z_j), \pi_s \in [k_0+1, K]\}} .
\end{aligned}$$

We now use $\mathbb{I}_{\{A_s^* \in [z_{j+1}, z_j)\}} = \mathbb{I}_{\{A_s^* < z_j\}} - \mathbb{I}_{\{A_s^* < z_{j+1}\}}$ and

reorder the sum,

$$\begin{aligned}
&\sum_{j=k_0}^{K}(\Delta_j-\Delta_{k_0})\mathbb{I}_{\{A_s^*\in[z_{j+1},z_j),\pi_s\in[k_0+1,K]\}}\\
&=\sum_{j=k_0}^{K}(\Delta_j-\Delta_{k_0})(\mathbb{I}_{\{A_s^*<z_j\}}-\mathbb{I}_{\{A_s^*<z_{j+1}\}})\mathbb{I}_{\{\pi_s\in[k_0+1,K]\}}\\
&=\left(\sum_{j=k_0}^{K}(\Delta_j-\Delta_{k_0})\mathbb{I}_{\{A_s^*<z_j\}}\right.\\
&\qquad\left.-\sum_{j=k_0+1}^{K+1}(\Delta_{j-1}-\Delta_{k_0})\mathbb{I}_{\{A_s^*<z_j\}}\right)\mathbb{I}_{\{\pi_s\in[k_0+1,K]\}}\\
&=\mathbb{I}_{\{\pi_s\in[k_0+1,K]\}}\sum_{j=k_0+1}^{K}(\Delta_j-\Delta_{j-1})\mathbb{I}_{\{A_s^*<z_j\}}
\end{aligned}$$

We will rewrite the sum over s of such terms as a sum over r, number of times that an arm in $[k_0+1,K]$ has been pulled. Note that if we know that suboptimal arms were pulled at least r times before a time s, we get $A_s^*\geq B_r^*$.

$$\begin{aligned}
&\sum_{s\geq 0}\mathbb{I}_{\{\pi_s\in[k_0+1,K]\}}\sum_{k=k_0+1}^{K}(\Delta_k-\Delta_{k-1})\mathbb{I}_{\{A_s^*<z_k\}}\\
&\qquad\leq\sum_{r\geq 0}\sum_{k=k_0+1}^{K}(\Delta_k-\Delta_{k-1})\mathbb{I}_{\{B_r^*<z_k\}}\,.
\end{aligned}$$

Intuitively, for each r, this is of the form $(\Delta_{k_r}-\Delta_{k_0})$ for some $k_r\geq k_0$ and is thus of the order of one Δ_{k_r}.

The sum describing the optimal arm is thus bounded as

$$\begin{aligned}
&\mathbb{E}[\sum_{r\geq 0}\sum_{k=k_0+1}^{K}(\Delta_k-\Delta_{k-1})\mathbb{I}_{\{B_r^*<z_k\}}]\\
&=\sum_{r\geq 0}\sum_{k=k_0+1}^{K}(\Delta_k-\Delta_{k-1})\mathbb{P}\{B_r^*<z_k\}\\
&\leq\sum_{r\geq 0}\sum_{k=k_0+1}^{K}(\Delta_k-\Delta_{k-1})\mathbb{P}\{\exists s\geq 1,\exists t'\geq r+s,\overline{X}_s^*+\epsilon_{t',s}<z_k\}\,.
\end{aligned}$$

We use the monotonicity of $u\mapsto\epsilon_{u,s}$ to simplify the event,

$$\begin{aligned}
&\sum_{r\geq 0}\sum_{k=k_0+1}^{K}(\Delta_k-\Delta_{k-1})\mathbb{P}\{\exists s\geq 1,\exists t'\geq r+s,\overline{X}_s^*+\epsilon_{t',s}<z_k\}\\
&\leq\sum_{k=k_0+1}^{K}(\Delta_k-\Delta_{k-1})\sum_{r\geq 0}\mathbb{P}\{\exists s\geq 1,\overline{X}_s^*+\epsilon_{r+s,s}^*<z_k\}\,.
\end{aligned}$$

Step 3: bounding the probability that a suboptimal arm is overestimated.

$$\begin{aligned}
&\sum_{j=k_0}^{K}\sum_{k=j+1}^{K}(\Delta_k-\Delta_{k_0})\mathbb{I}_{\{\pi_s=k,A_s^*\in[z_{j+1},z_j)\}}\\
&=\sum_{k=k_0+1}^{K}\sum_{j=k_0}^{k-1}(\Delta_k-\Delta_{k_0})\mathbb{I}_{\{\pi_s=k,A_s^*\in[z_{j+1},z_j)\}}\\
&=\sum_{k=k_0+1}^{K}(\Delta_k-\Delta_{k_0})\mathbb{I}_{\{\pi_s=k\}}\sum_{j=k_0}^{k-1}\mathbb{I}_{\{A_s^*\in[z_{j+1},z_j)\}}\\
&=\sum_{k=k_0+1}^{K}(\Delta_k-\Delta_{k_0})\mathbb{I}_{\{\pi_s=k,A_s^*\geq z_k\}}
\end{aligned}$$

As we did for the other sum, we replace the sum over the time of such terms by sums over the number of pulls of the arms. We denote by $P_r^{(k)}$ the event "arm k was pulled for the r^{th} time".

$$\begin{aligned}
&\sum_{s\geq 0}\sum_{k=k_0+1}^{K}(\Delta_k-\Delta_{k_0})\mathbb{I}_{\{\pi_s=k,A_s^*\geq z_k\}}\\
&\leq\sum_{k=k_0+1}^{K}(\Delta_k-\Delta_{k_0})\sum_{r\geq 0}\mathbb{I}_{\{P_r^{(k)},\text{ at time }t_r,\text{ and }A_{t_r}^*\geq z_k\}}\\
&\leq\sum_{r\geq 0}\sum_{k=k_0+1}^{K}(\Delta_k-\Delta_{k_0})\mathbb{I}_{\{P_r^{(k)},\text{ at time }t_r,\text{ and }\overline{X}_r^{(k)}+\epsilon_{t_r,r}\geq z_k\}}\\
&\leq\sum_{r\geq 0}\sum_{k=k_0+1}^{K}(\Delta_k-\Delta_{k_0})\mathbb{I}_{\{P_r^{(k)}\text{ and }\exists t'\geq r,\overline{X}_r^{(k)}+\epsilon_{t',r}\geq z_k\}}\\
&\leq\sum_{r\geq 0}\sum_{k=k_0+1}^{K}(\Delta_k-\Delta_{k_0})\mathbb{I}_{\{\exists t'\geq r,\overline{X}_r^{(k)}+\epsilon_{t',r}\geq z_k\}}\,.
\end{aligned}$$

For this sum, for each r we get a sum that intuitively can be of order $\sum_{k=k_0+1}^{K}(\Delta_k-\Delta_{k_0})$, that is roughly K times larger than the sum depending on the optimal arm.

The sum describing the suboptimal arms is thus bounded

as

$$\begin{aligned}
&\mathbb{E}[\sum_{r\geq 0}\sum_{k=k_0+1}^{K}(\Delta_k-\Delta_{k_0})\mathbb{I}_{\{\exists t'\geq r,\overline{X}_r^{(k)}+\epsilon_{t',r}\geq z_k\}}]\\
&=\sum_{r\geq 0}\sum_{k=k_0+1}^{K}(\Delta_k-\Delta_{k_0})\mathbb{P}\{\exists t'\geq r,\overline{X}_r^{(k)}+\epsilon_{t',r}\geq z_k\}\\
&\leq\sum_{r\geq 0}\sum_{k=k_0+1}^{K}\Delta_k\mathbb{P}\{\exists t'\geq r,\overline{X}_r^{(k)}+\epsilon_{t',r}\geq z_k\}\\
&\leq\sum_{k=k_0+1}^{K}\Delta_k\sum_{r\geq 0}\mathbb{P}\{\overline{X}_r^{(k)}+\epsilon_{t,r}\geq z_k\}\,,
\end{aligned}$$

where we used the monotonicity of $u\mapsto\epsilon_{u,s}$ in the last inequality.

Step 4: Controlling the probabilities. Putting the two previous steps together we get for the expected regret the inequality

$$\begin{aligned}
\mathbb{E}R_t\leq\ & t\Delta_{k_0}\\
&+\sum_{k=k_0+1}^{K}(\Delta_k-\Delta_{k-1})\sum_{r\geq 0}\mathbb{P}\{\exists s\geq 1,\overline{X}_s^*+\epsilon^*_{r+s,s}<z_k\}\\
&+\sum_{k=k_0+1}^{K}\Delta_k\sum_{r\geq 0}\mathbb{P}\{\overline{X}_r^{(k)}+\epsilon_{t,r}\geq z_k\}\,.
\end{aligned}$$

The next step is to control the sums of probabilities, which are small for r big enough. To this effect we cut the sums in two, a first part for small r for which the probability is upper bounded by 1 and a second part for big r. As noted previously, intuitively the first sum tend to be K times smaller than the second one. Thus we cut the sums at indices that differ by a factor K.

Let $\tilde{r}_k$ be the largest integer such that $\tilde{r}_k\leq\frac{K}{2\Delta_k^2}+1$ and $\tilde{r}'_k$ the largest integer such that $\tilde{r}'_k\leq\frac{(1+80\alpha)\overline{\log}(\frac{2t\Delta_k^2}{K})}{2\Delta_k^2}$.

$$\begin{aligned}
\mathbb{E}R_t\leq\ & t\Delta_{k_0}+\sum_{k=k_0+1}^{K}(\Delta_k-\Delta_{k-1})\tilde{r}_k\\
&+\sum_{k=k_0+1}^{K}(\Delta_k-\Delta_{k-1})\sum_{r>\tilde{r}_k}\mathbb{P}\{\exists s\geq 1,\overline{X}_s^*+\epsilon^*_{r+s,s}<z_k\}\\
&+\sum_{k=k_0+1}^{K}\Delta_k\tilde{r}'_k+\sum_{k=k_0+1}^{K}\Delta_k\sum_{r>\tilde{r}'_k}\mathbb{P}\{\overline{X}_r^{(k)}+\epsilon_{t,r}\geq z_k\}\\
&=t\Delta_{k_0}+A+B+C+D\,,
\end{aligned}$$

where A,B,C,D are the four sums of the previous lines.

Bounding term A. $\tilde{r}_k\leq\frac{K}{2\Delta_k^2}+1$ and thus

$$\begin{aligned}
A&\leq\sum_{k=k_0+1}^{K}(\Delta_k-\Delta_{k-1})(\frac{K}{2\Delta_k^2}+1)\\
&\leq\Delta_K+\sum_{k=k_0+1}^{K}(\Delta_k-\Delta_{k-1})\frac{K}{2\Delta_k^2}\\
&\leq\Delta_K+\frac{K}{\Delta_{k_0+1}}\,.
\end{aligned}$$

Bounding term B. With lemma 4,

$$B\leq\sum_{k=k_0+1}^{K}(\Delta_k-\Delta_{k-1})\frac{K}{\Delta_k^2}\log(\frac{2et\Delta_k^2}{K})\frac{4(1+\alpha)^{3/2}}{\alpha^2\log(1+\alpha)}$$

We compute the sum,

$$\begin{aligned}
&\sum_{k=k_0+1}^{K}(\Delta_k-\Delta_{k-1})\log(\frac{2et\Delta_k^2}{K})\frac{1}{\Delta_k^2}\\
&=\sum_{k=k_0+2}^{K}(\Delta_k-\Delta_{k-1})\log(\frac{2et\Delta_k^2}{K})\frac{1}{\Delta_k^2}\\
&\quad+\log(\frac{2et\Delta_{k_0+1}^2}{K})\frac{1}{\Delta_{k_0+1}^2}\\
&\leq\int_{\Delta_{k_0+1}}^{1}\log(\frac{2et}{K}x^2)\frac{1}{x^2}\,\mathrm{d}x+\log(\frac{2et\Delta_{k_0+1}^2}{K})\frac{1}{\Delta_{k_0+1}^2}\\
&=\left[\log(\frac{2et}{K}x^2)\frac{1}{x}+\frac{2}{x}\right]_1^{\Delta_{k_0+1}}+\log(\frac{2et\Delta_{k_0+1}^2}{K})\frac{1}{\Delta_{k_0+1}^2}\\
&\leq 2\log(\frac{2et\Delta_{k_0+1}^2}{K})\frac{1}{\Delta_{k_0+1}}+\frac{2}{\Delta_{k_0+1}}\\
&=2\log(\frac{2t\Delta_{k_0+1}^2}{K})\frac{1}{\Delta_{k_0+1}}+\frac{4}{\Delta_{k_0+1}}\,.
\end{aligned}$$

Then we have

$$\begin{aligned}
B&\leq\frac{8(1+\alpha)}{\alpha^2\log(1+\alpha)}\left(\log(\frac{2t\Delta_{k_0+1}^2}{K})\frac{K}{\Delta_{k_0+1}}+\frac{2K}{\Delta_{k_0+1}}\right)\\
&\leq\frac{8(1+\alpha)}{\alpha^2\log(1+\alpha)}\left(\overline{\log}(\frac{2t\Delta_{k_0+1}^2}{K})\frac{K}{\Delta_{k_0+1}}+\frac{2K}{\Delta_{k_0+1}}\right)
\end{aligned}$$

Bounding term C. $\tilde{r}'_k\leq\frac{(1+80\alpha)\overline{\log}(\frac{2t\Delta_k^2}{K})}{2\Delta_k^2}$ and then

$$\begin{aligned}
C&\leq\sum_{k=k_0+1}^{K}\Delta_k\frac{(1+80\alpha)\overline{\log}(\frac{2t\Delta_k^2}{K})}{2\Delta_k^2}\\
&\leq(1/2+40\alpha)\frac{K}{\Delta_{k_0+1}}\overline{\log}(\frac{2t\Delta_{k_0+1}^2}{K})\,.
\end{aligned}$$

Bounding term D. Since for $r > \tilde{r}'_k$, $r > \frac{(1+80\alpha)\overline{\log}(\frac{2t\Delta_k^2}{K})}{2\Delta_k^2} \geq \frac{1}{2\Delta_k^2}$, it is not difficult to see that $\epsilon_{t,r} \leq \Delta_k\sqrt{\frac{1+\alpha}{1+80\alpha}}$ and thus that lemma 2 applies.

$$D \leq \sum_{k=k_0+1}^{K} \frac{32}{\alpha^2\Delta_k} \leq \frac{32K}{\alpha^2\Delta_{k_0+1}} .$$

Step 5: Putting things together. We get the following bound for the regret, for any k_0,

$$\begin{aligned}\mathbb{E}R_t \leq\ & t\Delta_{k_0} \\ &+ \left(\frac{8(1+\alpha)^{3/2}}{\alpha^2\log(1+\alpha)}+1/2+40\alpha\right)\frac{K}{\Delta_{k_0+1}}\overline{\log}(\frac{2t\Delta_{k_0+1}^2}{K}) \\ &+ \left(\frac{16(1+\alpha)^{3/2}}{\alpha^2\log(1+\alpha)}+1+\frac{32}{\alpha^2}\right)\frac{K}{\Delta_{k_0+1}} + \Delta_K .\end{aligned}$$

We can then get two particular upper-bounds. The first one is obtained with k_0 the number of the last optimal arm,

$$\mathbb{E}R_t \leq C_\alpha\frac{K}{\Delta_{\min}}\overline{\log}(\frac{2t\Delta_{\min}^2}{K}) + C'_\alpha\frac{K}{\Delta_{\min}} + \Delta_K ,$$

then one upper-bound independent of the distributions, by taking k_0 such that $\Delta_{k_0} \leq \sqrt{\frac{K}{t}} < \Delta_{k_0+1}$,

$$\mathbb{E}R_t \leq \sqrt{Kt}(1 + C_\alpha\log(2) + C'_\alpha) + \Delta_K .$$

The values of C_α and C'_α are

$$\begin{aligned}C_\alpha &= \frac{8(1+\alpha)^{3/2}}{\alpha^2\log(1+\alpha)} + 40\alpha + 1/2 \\ C'_\alpha &= \frac{16(1+\alpha)^{3/2}}{\alpha^2\log(1+\alpha)} + 1 + \frac{32}{\alpha^2} .\end{aligned}$$

For $\alpha = 1.35$, the maximum value allowed by lemma 2, $C_\alpha \leq 75, C'_\alpha \leq 60$ and $(1 + C_\alpha\log(2) + C'_\alpha) \leq 113$. Hence the result.

□

4. Technical Lemmas

This section is dedicated to three bounds of quantities used in the proofs for MOSS-anytime and single-pull UCB2.

Lemma 2 (Suboptimal arms)**.** *For $k \in \{1, \dots, K\}$,*

$$\sum_{r=0}^{+\infty}\mathbb{P}\left\{\overline{X}_r^{(k)} + \Delta_k\sqrt{\frac{1+\alpha}{1+80\alpha}} \geq \mu^* - \alpha\frac{\Delta_k}{2}\right\} \leq \frac{32}{\alpha^2\Delta_k^2} .$$

Proof.

$$\begin{aligned}&\mathbb{P}\left\{\overline{X}_r^{(k)} + \Delta_k\sqrt{\frac{1+\alpha}{1+80\alpha}} \geq \mu^* - \alpha\frac{\Delta_k}{2}\right\} \\ &= \mathbb{P}\left\{\overline{X}_r^{(k)} - \mu^{(k)} \geq \Delta_k\left(1 - \frac{\alpha}{2} - \sqrt{\frac{1+\alpha}{1+80\alpha}}\right)\right\}\end{aligned}$$

To be able to apply Hoeffding's inequality, we need $1 - \frac{\alpha}{2} - \sqrt{\frac{1+\alpha}{1+80\alpha}} \geq 0$. This is in particular true for $\alpha \leq 1.35$. In this case $(1 - \frac{\alpha}{2} - \sqrt{\frac{1+\alpha}{1+80\alpha}}) \geq \frac{\alpha}{8}$ and

$$\begin{aligned}&\mathbb{P}\{\overline{X}_r^{(k)} + \Delta_k\sqrt{\frac{1+\alpha}{1+80\alpha}} \geq \mu^* - \delta\frac{\Delta_k}{2}\} \\ &\leq \mathbb{P}\{\overline{X}_r^{(k)} - \mu^{(k)} \geq \Delta_k\frac{\alpha}{8}\} \leq \exp(-r\frac{\alpha^2\Delta_k^2}{32}) .\end{aligned}$$

Moreover,

$$\begin{aligned}&\sum_{r\geq 1}\mathbb{P}\{\overline{X}_r^{(k)} + \Delta_k\sqrt{\frac{1+\alpha}{1+80\alpha}} \geq \mu^* - \delta\frac{\Delta_k}{2}\} \\ &\leq \sum_{r\geq 1}\exp(-r\frac{\alpha^2\Delta_k^2}{32}) \leq \sum_{r\geq 1}\int_r^{r+1}\exp(-x\frac{\alpha^2\Delta_k^2}{32})\,\mathrm{d}x \\ &\leq \int_0^{+\infty}\exp(-x\frac{\alpha^2\Delta_k^2}{32})\,\mathrm{d}x \leq \frac{32}{\alpha^2\Delta_k^2} .\end{aligned}$$

□

Lemma 3 (Optimal arm bound for single-pull UCB2)**.** *For $\delta > 0$ and $\eta \in (0, \sqrt{1+\alpha} - 1)$,*

$$\begin{aligned}&\sum_{r\geq\tilde{r}'_k}\mathbb{P}\left\{\exists s \geq 1, \overline{X}_s^* + \epsilon_{r+s,s} \leq \mu^* - \delta\frac{\Delta_k}{2}\right\} \\ &\leq \frac{2(1+\eta)^3}{\delta^2\Delta_k^2\log(1+\eta)(\frac{1+\alpha}{(1+\eta)^2} - 1)}\end{aligned}$$

Proof. We start by working on $\mathbb{P}\left\{\exists s \geq 1, \overline{X}_s^* + \epsilon_{r+s,s} \leq \mu^* - \delta\frac{\Delta_k}{2}\right\}$ and introduce a time grid for the pulls of the optimal arm similar to the $\tau(i)$ of the proof of UCB2 in (Auer et al., 2002). Let $\eta > 0$.

$$
\begin{aligned}
&\mathbb{P}\left\{\exists s \geq 1, \overline{X}_s^* + \epsilon_{r+s,s} \leq \mu^* - \delta\frac{\Delta_k}{2}\right\} \\
&= \mathbb{P}\left\{\exists s \geq 1, \sum_{u=1}^{s}(X_u^* - \mu^*) \leq -s(\epsilon_{r+s,s} + \delta\frac{\Delta_k}{2})\right\} \\
&\leq \sum_{i\geq 1}\mathbb{P}\left\{\exists s \in [(1+\eta)^{i-1},(1+\eta)^i), \sum_{u=1}^{s}(X_u^* - \mu^*)\right. \\
&\qquad\qquad \left. \leq -(1+\eta)^{i-1}(\epsilon_{r+(1+\eta)^i,(1+\eta)^i} + \delta\frac{\Delta_k}{2})\right\}
\end{aligned}
$$

(where we used that for a given u the function $h : s \mapsto \epsilon_{u+s,s} = \sqrt{\frac{(1+\alpha)\overline{\log}(\frac{eu}{s}+e)}{2s}}$ is decreasing)

$$
\begin{aligned}
&\leq \sum_{i\geq 1}\mathbb{P}\left\{\exists s \leq (1+\eta)^i, \sum_{u=1}^{s}(X_u^* - \mu^*)\right. \\
&\qquad\qquad \left. \leq -(1+\eta)^{i-1}(\epsilon_{r+(1+\eta)^i,(1+\eta)^i} + \delta\frac{\Delta_k}{2})\right\} \\
&\leq \sum_{i\geq 1}\exp\left(-2\frac{(1+\eta)^{2(i-1)}}{\lfloor(1+\eta)^i\rfloor}(\frac{\delta^2\Delta_k^2}{4} + \epsilon^2_{r+(1+\eta)^i,(1+\eta)^i})\right)
\end{aligned}
$$

(maximal Hoeffding's inequality)

$$
\begin{aligned}
&\leq \sum_{i\geq 1}\exp\left(-(1+\eta)^i\frac{1}{(1+\eta)^2}\frac{\delta^2\Delta_k^2}{2}\right. \\
&\qquad\left. -\frac{1+\alpha}{(1+\eta)^2}\overline{\log}(e + \frac{er}{(1+\eta)^i})\right)
\end{aligned}
$$

With η such that $1+\alpha > (1+\eta)^2$, the sum of the probabilities is then bounded as

$$
\begin{aligned}
&\sum_{r\geq\tilde{r}_k}\mathbb{P}\left\{\exists s \geq 1, \overline{X}_s^* + \epsilon_{r+s,s} \leq \mu^* - \delta\frac{\Delta_k}{2}\right\} \\
&\leq \sum_{r\geq\tilde{r}_k}\sum_{i\geq 1}\exp\left(-(1+\eta)^i\frac{1}{(1+\eta)^2}\frac{\delta^2\Delta_k^2}{2} -\right. \\
&\qquad\left. \frac{1+\alpha}{(1+\eta)^2}\overline{\log}(e + \frac{er}{(1+\eta)^i})\right) \\
&= \sum_{i\geq 1}\exp(-(1+\eta)^i\frac{\delta^2\Delta_k^2}{2(1+\eta)^2}) \\
&\qquad\times\sum_{r\geq\tilde{r}_k}\exp\left(-\frac{1+\alpha}{(1+\eta)^2}\log(e + \frac{er}{(1+\eta)^i})\right) \\
&\leq \sum_{i\geq 1}\exp(-(1+\eta)^i\frac{\delta^2\Delta_k^2}{2(1+\eta)^2}) \\
&\qquad\times\sum_{r\geq\tilde{r}_k}\left(1 + \frac{r}{(1+\eta)^i}\right)^{-\frac{1+\alpha}{(1+\eta)^2}} \\
&\leq \sum_{i\geq 1}\exp(-(1+\eta)^i\frac{\delta^2\Delta_k^2}{2(1+\eta)^2}) \\
&\qquad\times\int_0^{+\infty}\left(1 + \frac{x}{(1+\eta)^i}\right)^{-\frac{1+\alpha}{(1+\eta)^2}}\mathrm{d}x \\
&\leq \sum_{i\geq 1}\exp(-(1+\eta)^i\frac{\delta^2\Delta_k^2}{2(1+\eta)^2})(1+\eta)^i\frac{1}{\frac{1+\alpha}{(1+\eta)^2}-1} \\
&\leq \frac{1}{\frac{1+\alpha}{(1+\eta)^2}-1}\sum_{i\geq 1}(1+\eta)^i\exp(-(1+\eta)^i\frac{\delta^2\Delta_k^2}{2(1+\eta)^2})\,.
\end{aligned}
$$

Let $B = \frac{\delta^2\Delta_k^2}{2(1+\eta)^2}$, then

$$
\begin{aligned}
&\sum_{i\geq 1}(1+\eta)^i\exp(-(1+\eta)^i\frac{\delta^2\Delta_k^2}{2(1+\eta)^2}) \\
&\leq (1+\eta)\int_0^{+\infty}(1+\eta)^x\exp(-B(1+\eta)^x)\,\mathrm{d}x \\
&= \frac{(1+\eta)}{\log(1+\eta)}\int_1^{+\infty}\exp(-Bz)\,\mathrm{d}z \\
&= \frac{e^{-B}(1+\eta)}{B\log(1+\eta)} \\
&= \frac{2(1+\eta)^3}{\delta^2\Delta_k^2\log(1+\eta)}\exp\left(-\frac{\delta^2\Delta_k^2}{2(1+\eta)^2}\right) \\
&\leq \frac{2(1+\eta)^3}{\delta^2\Delta_k^2\log(1+\eta)}\,.
\end{aligned}
$$

Hence the result.

□

Lemma 4 (Optimal arm bound for MOSS-anytime)**.** *For any suboptimal arm k, for $\delta > 0$,*

$$
\begin{aligned}
&\sum_{r>\tilde{r}_k}\mathbb{P}\left\{\exists s \geq 1, \overline{X}_s^* + \epsilon_{r+s,s} \leq \mu^* - \delta\frac{\Delta_k}{2}\right\} \\
&\leq \frac{K}{\Delta_k^2}\log(\frac{2et\Delta_k^2}{K})\frac{4(1+\alpha)^{3/2}}{\delta^2\log(1+\alpha)}
\end{aligned}
$$

Proof. We want to bound a sum starting at $\tilde{r}_k + 1$. Here we use $\tilde{r}_k$ the largest integer such that $\tilde{r}_k \leq \frac{K}{2\Delta_k^2} + 1$. Thus $\tilde{r}_k \geq \frac{K}{2\Delta_k^2}$. We start by working on $\mathbb{P}\left\{\exists s \geq 1, \overline{X}_s^* + \epsilon_{r+s,s} \leq \mu^* - \delta\frac{\Delta_k}{2}\right\}$ for a fixed r. We use a peeling argument and Hoeffding's maximal inequality to control this probability. Let $\eta = \sqrt{1+\alpha} - 1$, such that $(1+\eta)^2 = (1+\alpha)$.

We recall the notation $\widehat{\log}(x) = \max(0, \log(x))$.

$$\mathbb{P}\left\{\exists s \geq 1, \overline{X}_s^* + \epsilon_{r+s,s} \leq \mu^* - \delta\frac{\Delta_k}{2}\right\}$$
$$= \mathbb{P}\left\{\exists s \geq 1, \sum_{u=1}^{s}(X_u^* - \mu^*) \leq -s(\epsilon_{r+s,s} + \delta\frac{\Delta_k}{2})\right\}$$
$$\leq \sum_{i\geq 1}\mathbb{P}\left\{\exists s \in [(1+\eta)^{i-1},(1+\eta)^i), \sum_{u=1}^{s}(X_u^* - \mu^*) \leq -(1+\eta)^{i-1}(\epsilon_{r+(1+\eta)^i,(1+\eta)^i} + \delta\frac{\Delta_k}{2})\right\}$$

(where we used that for a given u the function $h : s \mapsto \epsilon_{u+s,s} = \sqrt{\frac{(1+\alpha)\widehat{\log}(\frac{u}{Ks}+\frac{1}{K})}{2s}}$ is decreasing)

$$\leq \sum_{i\geq 1}\mathbb{P}\left\{\exists s \leq (1+\eta)^i, \sum_{u=1}^{s}(X_u^* - \mu^*) \leq -(1+\eta)^{i-1}(\epsilon_{r+(1+\eta)^i,(1+\eta)^i} + \delta\frac{\Delta_k}{2})\right\}$$
$$\leq \sum_{i\geq 1}\exp\left(-2\frac{(1+\eta)^{2(i-1)}}{\lfloor(1+\eta)^i\rfloor}(\frac{\delta^2\Delta_k^2}{4} + \epsilon^2_{r+(1+\eta)^i,(1+\eta)^i})\right)$$

(maximal Hoeffding's inequality)

$$\leq \sum_{i\geq 1}\exp\left(-(1+\eta)^i\frac{1}{(1+\eta)^2}\frac{\delta^2\Delta_k^2}{2} - \frac{1+\alpha}{(1+\eta)^2}\widehat{\log}(\frac{1}{K} + \frac{r}{K(1+\eta)^i})\right) .$$

Thus we can bound the sum of the probabilities,

$$\sum_{r>\tilde{r}_k}\mathbb{P}\left\{\exists s \geq 1, \overline{X}_s^* + \epsilon_{r+s,s} \leq \mu^* - \delta\frac{\Delta_k}{2}\right\}$$
$$\leq \sum_{r>\tilde{r}_k}\sum_{i\geq 1}\exp\left(-(1+\eta)^i\frac{1}{(1+\alpha)}\frac{\delta^2\Delta_k^2}{2} - \widehat{\log}(\frac{1}{K} + \frac{r}{K(1+\eta)^i})\right)$$
$$= \sum_{i\geq 1}\exp(-(1+\eta)^i\frac{\delta^2\Delta_k^2}{2(1+\alpha)}) \times \sum_{r>\tilde{r}_k}\exp\left(-\widehat{\log}(\frac{1}{K} + \frac{r}{K(1+\eta)^i})\right) .$$

Let $\overline{r}_i$ be the smallest integer r such that $\frac{1}{K} + \frac{r}{K(1+\eta)^i} \geq 1$. $\overline{r}_i$ is the smallest integer such that $\widehat{\log}(\frac{1}{K} + \frac{r}{K(1+\eta)^i}) = \log(\frac{1}{K} + \frac{r}{K(1+\eta)^i})$. For $r < \overline{r}_i$, $\exp\left(-\widehat{\log}(\frac{1}{K} + \frac{r}{K(1+\eta)^i})\right) = 1$. We have $\overline{r}_i \leq (K-1)(1+\eta)^i + 1 \leq K(1+\eta)^i$.

$$\sum_{i\geq 1}\exp(-(1+\eta)^i\frac{\delta^2\Delta_k^2}{2(1+\alpha)}) \times \sum_{r>\tilde{r}_k}\exp\left(-\widehat{\log}(\frac{1}{K} + \frac{r}{K(1+\eta)^i})\right)$$
$$\leq \sum_{i\geq 1}\exp(-(1+\eta)^i\frac{\delta^2\Delta_k^2}{2(1+\alpha)}) \times \left(\overline{r}_i + \sum_{r\geq\max(\overline{r}_i,\tilde{r}_k+1)}\left(\frac{1}{K} + \frac{r}{K(1+\eta)^i}\right)^{-1}\right)$$
$$\leq \sum_{i\geq 1}\exp(-(1+\eta)^i\frac{\delta^2\Delta_k^2}{2(1+\alpha)}) \times \left(\overline{r}_i + \int_{\tilde{r}_k}^{t}\left(\frac{1}{K} + \frac{x}{K(1+\eta)^i}\right)^{-1}\mathrm{d}x\right) ,$$

We use $\tilde{r}_k \geq \frac{K}{2\Delta_k^2}$ and compute an upper bound for the integral,

$$\sum_{i\geq 1}\exp(-(1+\eta)^i\frac{\delta^2\Delta_k^2}{2(1+\alpha)}) \times \left(\overline{r}_i + \int_{\tilde{r}_k}^{t}\left(\frac{1}{K} + \frac{x}{K(1+\eta)^i}\right)^{-1}\mathrm{d}x\right)$$
$$\leq \sum_{i\geq 1}\exp(-(1+\eta)^i\frac{\delta^2\Delta_k^2}{2(1+\alpha)}) \times \left(\overline{r}_i + \int_{\frac{K}{2\Delta_k^2}}^{t}\left(\frac{1}{K} + \frac{x}{K(1+\eta)^i}\right)^{-1}\mathrm{d}x\right)$$
$$= \sum_{i\geq 1}\exp(-(1+\eta)^i\frac{\delta^2\Delta_k^2}{2(1+\alpha)}) \times \left(\overline{r}_i + K(1+\eta)^i\int_{\frac{K}{2\Delta_k^2}}^{t}\frac{1}{x+(1+\eta)^i}\,\mathrm{d}x\right)$$
$$\leq \sum_{i\geq 1}\exp(-(1+\eta)^i\frac{\delta^2\Delta_k^2}{2(1+\alpha)}) \times \left(\overline{r}_i + K(1+\eta)^i\int_{\frac{K}{2\Delta_k^2}}^{t}\frac{1}{x}\,\mathrm{d}x\right)$$
$$\leq \sum_{i\geq 1}\exp(-(1+\eta)^i\frac{\delta^2\Delta_k^2}{2(1+\alpha)}) \times \left(K(1+\eta)^i + K(1+\eta)^i\log\frac{2t\Delta_k^2}{K}\right)$$
$$= K\log(\frac{2et\Delta_k^2}{K})\sum_{i\geq 1}(1+\eta)^i\exp(-(1+\eta)^i\frac{\delta^2\Delta_k^2}{2(1+\alpha)}) .$$

Let $B = \frac{\delta^2\Delta_k^2}{2(1+\alpha)}$.

$$
\begin{aligned}
&\sum_{i\geq 1}(1+\eta)^i \exp(-(1+\eta)^i \frac{\delta^2\Delta_k^2}{2(1+\alpha)}) \\
&\leq \int_1^{+\infty} (1+\eta)^x \exp(-B(1+\eta)^{x-1})\,\mathrm{d}x \\
&\leq (1+\eta)\int_0^{+\infty} (1+\eta)^x \exp(-B(1+\eta)^x)\,\mathrm{d}x \\
&= \frac{1+\eta}{\log(1+\eta)}\int_1^{+\infty} \exp(-Bz)\,\mathrm{d}z, \text{ with } z=(1+\eta)^x \\
&= \frac{e^{-B}(1+\eta)}{B\log(1+\eta)} \\
&= \frac{2(1+\alpha)^{3/2}}{\delta^2\Delta_k^2\log(1+\eta)}\exp\left(-\frac{\delta^2\Delta_k^2}{2(1+\alpha)^{3/2}}\right) \\
&\leq \frac{4(1+\alpha)^{3/2}}{\delta^2\Delta_k^2\log(1+\alpha)}
\end{aligned}
$$

Thus we proved

$$
\begin{aligned}
&\sum_{r>\tilde{r}_k}\mathbb{P}\left\{\exists s\geq 1, \overline{X}_s^* + \epsilon_{r+s,s} \leq \mu^* - \delta\frac{\Delta_k}{2}\right\} \\
&\leq \frac{K}{\Delta_k^2}\log(\frac{2et\Delta_k^2}{K})\frac{4(1+\alpha)^{3/2}}{\delta^2\log(1+\alpha)}
\end{aligned}
$$

□

References

Audibert, Jean-yves and Bubeck, Sbastien. Regret Bounds and Minimax Policies under Partial Monitoring. *Journal of Machine Learning Research*, 11:2785–2836, 2010.

Auer, Peter, Cesa-Bianchi, Nicolo, and Fischer, Paul. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.