SUPPLEMENTARY MATERIAL

A. Proof of Theorem 2

Theorem 2. *In any stochastic environment where the arms have expected rewards $\mu_i \in [0,1]$ with 1-subgaussian noise, Algorithm 1 satisfies the following with probability at least $1-\delta$ and for every time horizon n, when ψ^δ is chosen in accordance with Remark 1 and with $L = \psi^\delta(n)$:*

$$\sum_{s=1}^{t} \mu_{I_s} \geq (1-\alpha)\mu_0 t \qquad \text{for all } t \in \{1,\dots,n\}, \tag{5}$$

$$\widetilde{R}_n \leq \sum_{i>0:\Delta_i>0} \left(\frac{4L}{\Delta_i} + \Delta_i\right) + \frac{2(K+1)\Delta_0}{\alpha\mu_0} + \frac{6L}{\alpha\mu_0}\sum_{i=1}^{K} \frac{\Delta_0}{\max\{\Delta_i, \Delta_0 - \Delta_i\}}, \tag{10}$$

$$\widetilde{R}_n \in O\left(\sqrt{nKL} + \frac{KL}{\alpha\mu_0}\right). \tag{11}$$

Proof. By Remark 1, with probability $\mathbb{P}\{F\} \geq 1-\delta$ the confidence intervals are valid for all t and all arms $i \in \{1,\dots,K\}$:

$$\begin{aligned}|\hat{\mu}_i(t-1) - \mu_i| &\leq \sqrt{\psi^\delta(T_i(t-1))/T_i(t-1)} \\ &\leq \sqrt{L/T_i(t-1)};\end{aligned}$$

we will henceforth assume that this is the case (i.e. that F holds). By the definition of the confidence intervals and by the construction of Algorithm 1 we immediately satisfy the constraint

$$\sum_{t=1}^{n} \mu_{I_t} \geq (1-\alpha)n\mu_0 \qquad \text{for all } n.$$

We now bound the regret. Let $i > 0$ be the index of a sub-optimal arm and suppose $I_t = i$. Since the confidence intervals are valid,

$$\begin{aligned}\mu^* \leq \theta_i(t) &\leq \hat{\mu}_i(t-1) + \sqrt{L/T_i(t-1)} \\ &\leq \mu_i + 2\sqrt{L/T_i(t-1)}\,,\end{aligned}$$

which implies that arm i has not been chosen too often; in particular we obtain

$$T_i(n) \leq T_i(n-1) + 1 \leq \frac{4L}{\Delta_i^2} + 1. \tag{17}$$

and the regret satisfies

$$\widetilde{R}_n = \sum_{i=0}^{K} T_i(n)\Delta_i \leq \sum_{i>0:\Delta_i>0} \left(\frac{4L}{\Delta_i} + \Delta_i\right) + T_0(n)\Delta_0.$$

If $\Delta_0 = 0$ then the theorem holds trivially; we therefore assume that $\Delta_0 > 0$ and find an upper bound for $T_0(n)$.

Let $\tau = \max\{t \leq n \mid I_t = 0\}$ be the last round in which the default arm is played. Since F holds and $\theta_0(t) = \mu_0 < \mu^* < \max_i \theta_i(t)$, it follows that $J_t = 0$ is never the UCB choice; the default arm was only played because $\xi_\tau < 0$:

$$\sum_{i=0}^{K} T_i(\tau-1)\lambda_i(\tau) + \lambda_{J_\tau}(\tau) - (1-\alpha)\mu_0\tau < 0 \tag{18}$$

By dropping $\lambda_{J_\tau}(\tau)$, replacing τ with $\sum_{i=0}^{K} T_i(\tau-1)+1$, and rearranging the terms in (18), we get

$$\begin{aligned}&\alpha T_0(\tau-1)\mu_0 \\ &< (1-\alpha)\mu_0 + \sum_{i=1}^{K} T_i(\tau-1)\left((1-\alpha)\mu_0 - \lambda_i(\tau)\right) \\ &\leq (1-\alpha)\mu_0 \\ &\quad + \sum_{i=1}^{K} T_i(\tau-1)\left((1-\alpha)\mu_0 - \mu_i + \sqrt{\frac{L}{T_i(\tau-1)}}\right) \\ &\leq 1 + \sum_{i=1}^{K} S_i\,.\end{aligned} \tag{19}$$

where $a_i = (1-\alpha)\mu_0 - \mu_i$ and

$$\begin{aligned}S_i &= T_i(\tau-1)\cdot\left((1-\alpha)\mu_0 - \mu_i + \sqrt{L/T_i(\tau-1)}\right) \\ &= a_i T_i(\tau-1) + \sqrt{LT_i(\tau-1)}\end{aligned}$$

is a bound on the decrease in ξ_t in the first $\tau-1$ rounds due to choosing arm i. We will now bound S_i for each $i > 0$.

The first case is $a_i \geq 0$, i.e. $\Delta_i \geq \Delta_0 + \alpha\mu_0$. Then (17) gives $T_i(\tau-1) \leq 4L/\Delta_i^2 + 1$ and we get

$$S_i \leq \frac{4La_i}{\Delta_i^2} + \frac{2L}{\Delta_i} + 2 \leq \frac{6L}{\Delta_i} + 2\,. \tag{20}$$

The other case is $a_i < 0$, i.e. $\Delta_i < \Delta_0 + \alpha\mu_0$. Then

$$S_i \leq \sqrt{LT_i(\tau-1)} \leq \frac{2L}{\Delta_i} + 1, \tag{21}$$

and by using $ax^2 + bx \leq -b^2/4a$ for $a < 0$ we have

$$S_i \leq -\frac{L}{4a_i} = \frac{L}{4(\Delta_0 + \alpha\mu_0 - \Delta_i)}. \tag{22}$$

Summarizing (20) to (22) gives

$$S_i \leq \frac{6L}{\max\{\Delta_i, \Delta_0 - \Delta_i\}} + 2\,.$$

Continuing from (19), we get

$$\begin{aligned}T_0(n) &= T_0(\tau-1) + 1 \\ &\leq \frac{2K+2}{\alpha\mu_0} + \frac{1}{\alpha\mu_0}\sum_{i=1}^{K} \frac{6L}{\max\{\Delta_i, \Delta_0 - \Delta_i\}}\,.\end{aligned}$$

We can now upper bound the regret by

$$\widetilde{R}_n \le \sum_{i>0:\Delta_i>0} \left(\frac{4L}{\Delta_i} + \Delta_i\right) + \frac{2(K+1)\Delta_0}{\alpha\mu_0} + \frac{6L}{\alpha\mu_0}\sum_{i=1}^{K}\frac{\Delta_0}{\max\{\Delta_i, \Delta_0 - \Delta_i\}}. \quad (10)$$

We will now show (11). To bound the regret due to the non-default arms, Jensen's inequality gives

$$\left(\sum_{i>0} T_i(n)\Delta_i\right)^2 \le m^2 \sum_{i>0} \frac{T_i(n)}{m}\Delta_i^2,$$

where $m \le n$ is the number of times non-default arms were chosen. Combining this with $\Delta_i^2 \le 4L/T_i(n)$ for sub-optimal arms from (17) gives

$$\sum_{i>0} T_i(n)\Delta_i \le 2\sqrt{mKL} \in O(\sqrt{nKL}).$$

To bound the regret due to the default arm, observe that $\max\{\Delta_i, \Delta_0 - \Delta_i\} \ge \Delta_0/2$ and thus $T_0(n)\Delta_0 \in O(KL/\alpha\mu_0)$. Combining these two bounds gives (11). □

B. Proof of Theorem 5

Theorem 5. *Algorithm 1, modified as above to work without knowing μ_0 but otherwise the same conditions as Theorem 2, satisfies with probability $1-\delta$ and for all time horizons n the constraint (5) and the regret bound*

$$\widetilde{R}_n \le \sum_{i:\Delta_i>0} \left(\frac{4L}{\Delta_i} + \Delta_i\right) + \frac{2(K+1)\Delta_0}{\alpha\mu_0} + \frac{7L}{\alpha\mu_0}\sum_{i=1}^{K}\frac{\Delta_0}{\max\{\Delta_i, \Delta_0 - \Delta_i\}}. \quad (15)$$

Proof. We proceed very similarly to the proof of Theorem 2 in Appendix A. As we did there, we assume that F holds: the confidence intervals are valid for all rounds and all arms (including the default), which happens with probability $\mathbb{P}\{F\} \ge 1-\delta$.

To show that the modified algorithm satisfies the constraint (5), we write the budget (6) as

$$\widetilde{Z}_t = \sum_{i=1}^{K} T_i(t-1)\mu_i + \mu_{J_t} + (T_0(t-1) - (1-\alpha)t)\mu_0$$

when the UCB arm J_t is chosen and show that it is indeed lower-bounded by

$$\xi'_t = \sum_{i=1}^{K} T_i(t-1)\lambda_i(t) + \lambda_{J_t}(t) + (T_0(t-1) - (1-\alpha)t)\theta_0(t). \quad (14)$$

This is apparent if $T_0(t-1) < (1-\alpha)t$, since the last term in (14) is then negative and $\theta_0(t) \ge \mu_0$. On the other hand, if $T_0(t-1) \ge (1-\alpha)t$ then the constraint is still satisfied:

$$\sum_{s=1}^{t} \mu_{I_s} \ge T_0(t-1)\mu_0 \ge (1-\alpha)\mu_0 t.$$

We now upper-bound the regret. As in the earlier proof, we can show that for any arm $i > 0$ with $\Delta_i > 0$ we have $T_i(n) \le 4L/\Delta_i^2 + 1$. If this also holds for $i = 0$ or if $\Delta_0 = 0$ then $\widetilde{R}_n \le \sum_{i:\Delta>0}(4L/\Delta_i + \Delta_i)$ and the theorem holds trivially. From now on we only consider the case when $\Delta_0 > 0$ and $T_0(n) > 4L/\Delta_0^2 + 1$. As before, we will proceed to upper-bound $T_0(n)$.

Let τ be the last round in which $I_\tau = 0$. We can ignore the possibility that $J_\tau = 0$, since then the above bound on $T_i(n)$ would apply even to the default arm, contradicting our assumption above. Thus we can assume that the default arm was played because $\xi'_\tau < 0$:

$$\sum_{i=1}^{K} T_i(\tau-1)\lambda_i(\tau) + \lambda_{J_\tau}(\tau) + \big(T_0(\tau-1) - (1-\alpha)\tau\big)\theta_0(\tau) < 0,$$

in which we drop $\lambda_{J_\tau}(\tau)$, replace τ with $\sum_{i=0}^{K} T_i(\tau-1)+1$, and rearrange the terms to get

$$\alpha T_0(\tau-1)\theta_0(\tau) < (1-\alpha)\theta_0(\tau) + \sum_{i=1}^{K} T_i(\tau-1)\big((1-\alpha)\theta_0(\tau) - \lambda_i(\tau)\big). \quad (23)$$

We lower-bound the left-hand side of (23) using $\theta_0(\tau) \ge \mu_0$, whereas we upper-bound the right-hand side using

$$\theta_0(\tau) \le \mu_0 + \sqrt{\frac{L}{T_0(\tau-1)}} \le \mu_0 + \frac{\Delta_0}{2},$$

which comes from $T_0(\tau-1) \ge 4L/\Delta_0^2$. Combining these in (23) with the lower confidence bound $\lambda_i(\tau) \ge \mu_i - \sqrt{L/T_i(\tau-1)}$ gives

$$\begin{aligned}
\alpha\mu_0 T_0(\tau-1) &< (1-\alpha)\left(\mu_0 + \frac{\Delta_0}{2}\right) \\
&\quad + \sum_{i=1}^{K} T_i(\tau-1)\left((1-\alpha)\left(\mu_0 + \frac{\Delta_0}{2}\right) - \mu_i + \sqrt{\frac{L}{T_i(\tau-1)}}\right) \\
&= (1-\alpha)\left(\mu_0 + \frac{\Delta_0}{2}\right) + \sum_{i=1}^{K} S_i \\
&\le 1 + \sum_{i=1}^{K} S_i, \qquad (24)
\end{aligned}$$

where $a_i = (1-\alpha)(\mu_0 + \Delta_0/2) - \mu_i$ and

$$S_i = a_i T_i(\tau-1) + \sqrt{L T_i(\tau-1)}$$

is a bound on the decrease in ξ'_t in the first $\tau-1$ rounds due to choosing arm i. We will now bound S_i for each $i > 0$.

Analogously to the previous proof, we get the bounds

$$S_i \le \frac{6L}{\Delta_i} + 2, \quad \text{when } a_i \ge 0\,; \tag{25}$$

$$S_i \le \frac{2L}{\Delta_i} + 1\,, \quad \text{otherwise;} \tag{26}$$

and in the latter case, using $ax^2 + bx \le -b^2/4a$ gives

$$S_i \le -\frac{L}{4a_i} = \frac{L}{4\big((1+\alpha)\Delta_0/2 + \alpha\mu_0 - \Delta_i\big)}\,. \tag{27}$$

Summarizing (25) to (27) gives

$$\begin{aligned} S_i &\le \frac{6L}{\max\big\{\Delta_i, 24\big((1+\alpha)\Delta_0/2 + \alpha\mu_0 - \Delta_i\big)\big\}} + 2 \\ &\le \frac{7L}{\max\{\Delta_i, \Delta_0 - \Delta_i\}} + 2\,. \end{aligned}$$

Continuing with (24), if $T_0(n) > \frac{4L}{\Delta_0^2} + 1$, we get

$$\begin{aligned} T_0(n) &= T_0(\tau-1) + 1 \\ &\le \frac{2K+2}{\alpha\mu_0} + \frac{1}{\alpha\mu_0}\sum_{i=1}^{K} \frac{7L}{\max\{\Delta_i, \Delta_0 - \Delta_i\}}\,. \end{aligned}$$

We can now upper bound the regret by

$$\begin{aligned} \widetilde{R}_n &\le \sum_{i:\Delta_i>0}\left(\frac{4L}{\Delta_i} + \Delta_i\right) + \frac{2(K+1)\Delta_0}{\alpha\mu_0} \\ &\quad + \frac{7L}{\alpha\mu_0}\sum_{i=1}^{K}\frac{\Delta_0}{\max\{\Delta_i, \Delta_0 - \Delta_i\}}\,. \end{aligned} \tag{15}$$

□

C. Proof of Theorem 7

Theorem 7. *Any $\hat{R}^\delta_t$-admissible algorithm $\mathcal{A}$, when adapted with our safe-playing strategy, satisfies the constraint (2) and has a regret bound of $R_n \le t_0 + \hat{R}^\delta_n$ with probability at least $1-\delta$ where $t_0 = \max\{t \mid \alpha\mu_0 t \le \hat{R}^\delta_t + \mu_0\}$.*

Proof of Theorem 7. It is clear from the description of the safe-playing strategy that it is indeed safe: the constraint (2) is always satisfied.

The algorithm plays safe when the following quantity, which is a lower bound on the budget Z_t, is negative:

$$Z'_t = Z_t - X_{t,I_t} = \sum_{s=1}^{t-1} X_{s,I_s} - (1-\alpha)\mu_0 t$$

To upper bound the regret, consider only the rounds in which our safe-playing strategy does not interfere with playing $\mathcal{A}$'s choice of arm. Then with probability $1-\delta$,

$$\max_{i\in\{0,\dots,K\}} \sum_{s=1}^{t} \mathbb{1}\{Z'_s \ge 0\}\,(X_{s,i} - X_{s,I_s}) \le \hat{R}^\delta_{B(t)}$$

where $B(t) = \sum_{s=1}^{t}\mathbb{1}\{Z'_s \ge 0\}$. Let τ be the last round in which the algorithm plays safe.

$$\begin{aligned} &\mu_0 B(\tau-1) \\ &\quad \le \max_i \sum_{s=1}^{\tau-1} \mathbb{1}\{Z'_s \ge 0\}\,X_{s,i} \\ &\quad \le \hat{R}^\delta_{B(\tau-1)} + \sum_{s=1}^{\tau-1}\mathbb{1}\{Z'_s \ge 0\}\,X_{s,I_s} \\ &\quad = \hat{R}^\delta_{B(\tau-1)} + \sum_{s=1}^{\tau-1} X_{s,I_s} - \mu_0(\tau - 1 - B(\tau-1)) \\ &\quad \le \hat{R}^\delta_{B(\tau-1)} + (1-\alpha)\mu_0\tau - \mu_0(\tau - 1 - B(\tau-1))\,, \end{aligned}$$

which indicates $\alpha\mu_0\tau \le \hat{R}^\delta_\tau + \mu_0$ and thus $\tau \le t_0$. It follows that $R_n \le t_0 + \hat{R}^\delta_n$. □

D. Proof of Theorem 9

Theorem 9. *Suppose for any $\mu_i \in [0,1]$ $(i > 0)$ and μ_0 satisfying*

$$\min\{\mu_0, 1-\mu_0\} \ge \max\Big\{1/2\sqrt{\alpha}, \sqrt{e+1/2}\Big\}\sqrt{K/n},$$

an algorithm satisfies $\mathbb{E}_\mu \sum_{t=1}^{n} X_{t,I_t} \ge (1-\alpha)\mu_0 n$. Then there is some $\mu \in [0,1]^K$ such that its expected regret satisfies $\mathbb{E}_\mu R_n \ge B$ where

$$B = \max\left\{\frac{K}{(16e+8)\alpha\mu_0}, \frac{\sqrt{Kn}}{\sqrt{16e+8}}\right\}. \tag{16}$$

Proof of Theorem 9. Pick any algorithm. We want to show that the algorithm's regret on some environment is at least as large as B. If $\mathbb{E}_\mu R_n > B$ for some $\mu \in [0,1]^K$, there is nothing to be proven. Hence, without loss of generality, we can assume that the algorithm is *consistent* in the sense that $\mathbb{E}_\mu R_n \le B$ for all $\mu \in [0,1]^K$.

For some $\Delta > 0$, define environment $\mu \in \mathbb{R}^K$ such that $\mu_i = \mu_0 - \Delta$ for all $i \in [K]$. For now, assume that μ_0 and Δ are such that $\mu_i \ge 0$; we will get back to this condition later. Also define environment $\mu^{(i)}$ for each $i = 1, \dots, K$ by

$$\mu^{(i)}_j = \begin{cases} \mu_0 + \Delta, & \text{for } j = i\,; \\ \mu_0 - \Delta, & \text{otherwise.} \end{cases}$$

In this proof, we use $T_i = T_i(n)$ to denote the number of times arm i was chosen in the first n rounds. We distinguish two cases, based on how large the exploration budget is.

Case 1: $\alpha \geq \dfrac{\sqrt{K}}{\mu_0\sqrt{(16e+8)n}}$.

In this case, $B = \frac{\sqrt{Kn}}{\sqrt{16e+8}}$ and we use $\Delta = (4e+2)B/n$. For each $i \in [K]$ define event $A_i = \{T_i \leq 2B/\Delta\}$. First we prove that $\mathbb{P}_\mu(A_i) \geq 1/2$:

$$\begin{aligned}\mathbb{P}_\mu\{T_i \leq 2B/\Delta\} &= 1 - \mathbb{P}_\mu\{T_i > 2B/\Delta\} \\ &\geq 1 - \frac{\Delta \mathbb{E}_\mu[T_i]}{2B} \\ &\geq 1 - \frac{\mathbb{E}_\mu[R_n]}{2B} \geq \frac{1}{2}\,.\end{aligned}$$

Next we prove that $\mathbb{P}_{\mu^{(i)}}(A_i) \leq 1/4e$:

$$\begin{aligned}\mathbb{P}_{\mu^{(i)}}\{T_i \leq 2B/\Delta\} &= \mathbb{P}_{\mu^{(i)}}\{n - T_i \geq n - 2B/\Delta\} \\ &\leq \frac{\mathbb{E}_{\mu^{(i)}}[n - T_i]}{n - 2B/\Delta} \\ &\leq \frac{B}{\Delta n - 2B} = \frac{1}{4e}\,.\end{aligned}$$

Note that μ and $\mu^{(i)}$ differ only in the ith component: $\mu_i = \mu_0 - \Delta$ whereas $\mu_i^{(i)} = \mu_0 + \Delta$. Then the KL divergence between the reward distributions of the ith arms is $\mathrm{KL}(\mu_i, \mu_i^{(i)}) = (2\Delta)^2/2 = 2\Delta^2$. Define the *binary relative entropy* to be

$$d(x,y) = x \log \frac{x}{y} + (1-x) \log \frac{1-x}{1-y};$$

it satisfies $d(x,y) \geq (1/2)\log(1/4y)$ for $x \in [1/2, 1]$ and $y \in (0,1)$. By a standard change of measure argument (see, e.g., Kaufmann et al., 2015, Lemma 1) we get that

$$\begin{aligned}\mathbb{E}_\mu[T_i] \cdot \mathrm{KL}(\mu_i; \mu_i^{(i)}) &\geq d(\mathbb{P}_\mu(A_i), \mathbb{P}_{\mu^{(i)}}(A_i)) \\ &\geq \frac{1}{2} \log \frac{1}{4(1/4e)} = \frac{1}{2}\end{aligned}$$

and so $\mathbb{E}_\mu[T_i] \geq 1/4\Delta^2$ for each $i \in [K]$. Hence

$$\mathbb{E}_\mu[R_n] = \Delta \sum_{i \in [K]} \mathbb{E}_\mu[T_i] \geq \frac{K}{4\Delta} = \frac{\sqrt{Kn}}{\sqrt{16e+8}} = B\,.$$

Case 2: $\alpha < \dfrac{\sqrt{K}}{\mu_0\sqrt{(16e+8)n}}$.

In this case, $B = \frac{K}{(16e+8)\alpha\mu_0}$ and we use $\Delta = K/4\alpha\mu_0 n$. For each i define the event $A_i = \{T_i \leq 2\alpha\mu_0 n/\Delta\}$. First we prove that $\mathbb{P}_\mu(A_i) \geq 1/2$:

$$\begin{aligned}\mathbb{P}_\mu\{T_i \leq 2\alpha\mu_0 n/\Delta\} &= 1 - \mathbb{P}_\mu\{T_i > 2\alpha\mu_0 n/\Delta\} \\ &\geq 1 - \frac{\Delta \mathbb{E}_\mu[T_i]}{2\alpha\mu_0 n} \\ &\geq 1 - \frac{\mathbb{E}_\mu[R_n]}{2\alpha\mu_0 n} \geq \frac{1}{2}\,,\end{aligned}$$

where we use the fact that

$$\begin{aligned}\mathbb{E}_\mu[R_n] &= n\mu_0 - \mathbb{E}_\mu\Big[\sum_{t=1}^n X_{t,I_t}\Big] \\ &\leq n\mu_0 - (1-\alpha)\mu_0 n = \alpha\mu_0 n.\end{aligned}$$

Next, we show that $\mathbb{P}_{\mu^{(i)}}(A_i) < 1/4e$:

$$\begin{aligned}&\mathbb{P}_{\mu^{(i)}}\{T_i \leq 2\alpha\mu_0 n/\Delta\} \\ &\quad= \mathbb{P}_{\mu^{(i)}}\{n - T_i \geq n - 2\alpha\mu_0 n/\Delta\} \\ &\quad\leq \frac{\mathbb{E}_{\mu^{(i)}}[n - T_i]}{n - 2\alpha\mu_0 n/\Delta} \\ &\quad\leq \frac{B}{\Delta n - 2\alpha\mu_0 n} \\ &\quad= \frac{K}{(4e+2)K - (32e+16)\alpha^2\mu_0^2 n} < \frac{1}{4e}\,.\end{aligned}$$

As in the other case, we have $\mathbb{E}_\mu[T_i] > 1/4\Delta^2$ for each $i \in [K]$. Therefore

$$\mathbb{E}_\mu[R_n] = \Delta \sum_{i \in [K]} \mathbb{E}_\mu[T_i] > \frac{K}{4\Delta} = \alpha\mu_0 n,$$

which contradicts the fact that $\mathbb{E}_\mu[R_n] \leq \alpha\mu_0 n$. So there does not exist an algorithm whose worst-case regret is smaller than B.

To summarize, we proved that

$$\mathbb{E}_\mu R_n \geq \begin{cases} \dfrac{\sqrt{Kn}}{\sqrt{16e+8}}, & \text{when } \alpha \geq \dfrac{\sqrt{K}}{\mu_0\sqrt{(16e+8)n}} \\ \dfrac{K}{(16e+8)\alpha\mu_0}, & \text{otherwise,} \end{cases}$$

finishing the proof. □