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 $\psi^\delta(n)$ is chosen in accordance with Remark 1 and with $L = \psi^\delta(n)$:

$$\sum_{t=1}^n \mu_{I_t} \geq (1 - \alpha)\mu_0 t \quad \text{for all } t \in \{1, \ldots, n\},$$

$$\tilde{R}_n \leq \sum_{i=1}^n \left(4L \Delta_i + \Delta_i\right) + \frac{2(K+1)\Delta_0}{\alpha\mu_0} + \frac{6L}{\alpha\mu_0} \sum_{i=1}^K \max\{\Delta_i, \Delta_0 - \Delta_i\},$$

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

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

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

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

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

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, $\mu^* \leq \theta_i(t) \leq \hat{\mu}_i(t-1) + \sqrt{L/T_i(t-1)} \leq \mu_i + 2\sqrt{L/T_i(t-1)},$ which implies that arm $i$ has not been chosen too often; in particular we obtain

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

and the regret satisfies

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

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 $J_\tau = 0$ is never the UCB choice; the default arm was only played because $\xi_\tau < 0$:

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

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

$$\alpha T_0(\tau - 1)\mu_0$$

$$< (1 - \alpha)\mu_0 + \sum_{i=1}^K T_i(\tau - 1) \left((1 - \alpha)\mu_0 - \lambda_i(\tau)\right)$$

$$\leq (1 - \alpha)\mu_0$$

$$+ \sum_{i=1}^K T_i(\tau - 1) \left((1 - \alpha)\mu_0 - \mu_i + \sqrt{L/T_i(\tau - 1)}\right)$$

$$\leq 1 + \sum_{i=1}^K S_i.$$ (19)

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

$$S_i = T_i(\tau - 1) \left((1 - \alpha)\mu_0 - \mu_i + \sqrt{L/T_i(\tau - 1)}\right)$$

is a bound on the decrease in $\xi_i$ 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

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

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

$$S_i \leq \frac{L}{4a_i} = \frac{L}{4(\Delta_0 + \alpha\mu_0 - \Delta_i)}.$$ (21)

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

$$S_i \leq \frac{6L}{\max\{\Delta_i, \Delta_0 - \Delta_i\}} + 2.$$ (22)

Summarizing (20) to (22) gives

$$S_i \leq \frac{6L}{\max\{\Delta_i, \Delta_0 - \Delta_i\}} + 2.$$ (23)

Continuing from (19), we get

$$T_0(n) = T_0(n-1) + 1$$

$$\leq K + 2 + \sum_{i=1}^K \frac{6L}{\max\{\Delta_i, \Delta_0 - \Delta_i\}}.$$ (24)
We can now upper bound the regret by
\[
\tilde{R}_n \leq \sum_{i>0, \Delta_i>0} \left( \frac{4L}{\Delta_i} + \Delta_i + \frac{2(K+1)\Delta_0}{\alpha \mu_0} \right) + \frac{6L}{\alpha \mu_0} \sum_{i=1}^{K} \frac{\Delta_0}{\max\{\Delta_i, \Delta_0 - \Delta_i\}}.
\]
(10)

We will now show (11). To bound the regret due to the non-default arms, Jensen’s inequality gives
\[
\left( \sum_{i>0} T_i(n) \Delta_i \right)^2 \leq m^2 \sum_{i>0} \frac{T_i(n)}{m} \Delta_i^2,
\]
where \(m \leq n\) is the number of times non-default arms were chosen. Combining this with \(\Delta_i^2 \leq 4L/T_i(n)\) for sub-optimal arms from (17) gives
\[
\sum_{i>0} T_i(n) \Delta_i \leq 2\sqrt{mKL} \in O(\sqrt{nKL}).
\]
To bound the regret due to the default arm, observe that \(\max\{\Delta_i, \Delta_0 - \Delta_i\} \geq \Delta_0/2\) 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 \(\mu_0\) 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
\[
\tilde{R}_n \leq \sum_{i>0} \left( \frac{4L}{\Delta_i} + \Delta_i + \frac{2(K+1)\Delta_0}{\alpha \mu_0} \right) + \frac{7L}{\alpha \mu_0} \sum_{i=1}^{K} \frac{\Delta_0}{\max\{\Delta_i, \Delta_0 - \Delta_i\}}.
\]
(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 \(P\{F\} \geq 1 - \delta\).

To show that the modified algorithm satisfies the constraint (5), we write the budget (6) as
\[
\tilde{Z}_t = \sum_{i=1}^{K} T_i(t-1) \mu_i + \mu J_t + (T_0(t-1) - (1-\alpha)t) \mu_0
\]
when the UCB arm \(J_t\) is chosen and show that it is indeed lower-bounded by
\[
\xi_t^* = \sum_{i=1}^{K} T_i(t-1) \lambda_i(t) + \lambda J_t(t)
\]
+ \((T_0(t-1) - (1-\alpha)t) \theta_0(t)\).
(14)

This is apparent if \(T_0(t-1) < (1-\alpha)t\), since the last term in (14) is then negative and \(\theta_0(t) \geq \mu_0\). On the other hand, if \(T_0(t-1) \geq (1-\alpha)t\) then the constraint is still satisfied:
\[
\sum_{i=1}^{K} \mu_i \geq T_0(t-1) \mu_0 \geq (1-\alpha) \mu_0 t.
\]

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 \(T_i(n) \leq 4L/\Delta_i^2 + 1\). If this also holds for \(i = 0\) or if \(\Delta_0 = 0\) then \(\tilde{R}_n \leq \sum_{i>0} (4L/\Delta_i^2 + \Delta_i)\) 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 \(\tau\) be the last round in which \(I_\tau = 0\). We can ignore the possibility that \(J_\tau = 0\), since then the above bound on \(T_0(n)\) would apply even to the default arm, contradicting our assumption above. Thus we can assume that the default arm was played because \(\xi^*_t < 0\):
\[
\sum_{i=1}^{K} T_i(\tau - 1) \lambda_i(\tau) + \lambda J_\tau(\tau)
\]
+ \((T_0(\tau - 1) - (1-\alpha)\tau) \theta_0(\tau) < 0,
\]
in which we drop \(\lambda J_\tau(\tau)\), replace \(\tau\) with \(\sum_{i=0}^{K} T_i(\tau - 1) + 1\), and rearrange the terms to get
\[
\alpha T_0(\tau - 1) \theta_0(\tau) < (1-\alpha) \theta_0(\tau)
\]
+ \(\sum_{i=1}^{K} T_i(\tau - 1)(1-\alpha) \theta_0(\tau) - \lambda_i(\tau)\).
(23)

We lower-bound the left-hand side of (23) using \(\theta_0(\tau) \geq \mu_0\), whereas we upper-bound the right-hand side using
\[
\theta_0(\tau) \leq \mu_0 + \sqrt{\frac{L}{T_0(\tau - 1)}} \leq \mu_0 + \frac{\Delta_0}{2},
\]
which comes from \(T_0(\tau - 1) \geq 4L/\Delta_0^2\). Combining these in (23) with the lower confidence bound \(\lambda_i(\tau) \geq \mu_i - \sqrt{L/T_0(\tau - 1)}\) gives
\[
\alpha \mu_0 T_0(\tau - 1) < (1-\alpha) \left( \mu_0 + \frac{\Delta_0}{2} \right)
\]
+ \(\sum_{i=1}^{K} T_i(\tau - 1) \left( (1-\alpha) \left( \mu_0 + \frac{\Delta_0}{2} \right) \right)
\]
- \(\mu_i + \sqrt{\frac{L}{T_0(\tau - 1)}}\)
\[
= (1-\alpha) \left( \mu_0 + \frac{\Delta_0}{2} \right) + \sum_{i=1}^{K} S_i
\]
\[
\leq 1 + \sum_{i=1}^{K} S_i,
\]
(24)
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 \( \xi'_i \) 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
\[
S_i \leq \frac{6L}{\Delta_i} + 2, \quad \text{when } a_i \geq 0;
\]
\[
S_i \leq \frac{2L}{\Delta_i} + 1, \quad \text{otherwise};
\]
and in the latter case, using \( ax^2 + bx \leq -b^2/4a \) gives
\[
S_i \leq -\frac{L}{4a_i} = \frac{L}{4((1 + \alpha)\Delta_0/2 + \alpha \mu_0 - \Delta_i)}.
\]
Summarizing (25) to (27) gives
\[
S_i \leq \frac{6L}{\max\{\Delta_i, 24((1 + \alpha)\Delta_0/2 + \alpha \mu_0 - \Delta_i)\}} + 2
\]
\[
\leq \frac{7L}{\max\{\Delta_i, \Delta_0 - \Delta_i\}} + 2.
\]
Continuing with (24), if \( T_0(n) > \frac{4L}{\Delta_0} + 1 \), we get
\[
T_0(n) = T_0(\tau - 1) + 1
\]
\[
\leq \frac{2K + 2}{\alpha \mu_0} + \frac{1}{\alpha \mu_0} \sum_{i=1}^{K} \max\{\Delta_i, \Delta_0 - \Delta_i\}.
\]
We can now upper bound the regret by
\[
\tilde{R}_n \leq \sum_{i: \Delta_i > 0} \left( \frac{4L}{\Delta_i} + \Delta_i \right) + \frac{2(K + 1)\Delta_0}{\alpha \mu_0}
\]
\[
+ \frac{7L}{\alpha \mu_0} \sum_{i=1}^{K} \max\{\Delta_i, \Delta_0 - \Delta_i\}.
\]
\[
(15) \quad \square
\]

C. Proof of Theorem 7

**Theorem 7.** Any \( \hat{R}^\delta_1 \)-admissible algorithm \( \mathcal{A} \), when adapted with our safe-playing strategy, satisfies the constraint (2) and has a regret bound of \( R_n \leq t_0 + \hat{R}^\delta_1 \) with probability at least \( 1 - \delta \) where \( t_0 = \max\{t | \alpha \mu_0 t \leq \hat{R}^\delta_1 + \mu_0 \} \).

**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 \( 1 - \delta \),
\[
\max_{i \in \{0, \ldots, K\}} \sum_{s=1}^{t} 1\{Z'_s \geq 0\} (X_{s,i} - X_{s,i_t}) \leq \hat{R}^\delta_B(t)
\]
where \( B(t) = \sum_{s=1}^{t} 1\{Z'_s \geq 0\} \). Let \( \tau \) be the last round in which the algorithm plays safe.
\[
\mu_0 B(\tau - 1)
\]
\[
\leq \max_{i \in \{0, \ldots, K\}} \sum_{s=1}^{\tau - 1} 1\{Z'_s \geq 0\} X_{s,i}
\]
\[
\leq \hat{R}^\delta_B(\tau - 1) + \sum_{s=1}^{\tau - 1} 1\{Z'_s \geq 0\} X_{s,i_t}
\]
\[
= \hat{R}^\delta_B(\tau - 1) + \sum_{s=1}^{\tau - 1} X_{s,i_s} - \mu_0(\tau - 1 - B(\tau - 1))
\]
\[
\leq \hat{R}^\delta_B(\tau - 1) + (1 - \alpha)\mu_0 \tau - \mu_0(\tau - 1 - B(\tau - 1)),
\]
which indicates \( \alpha \mu_0 \tau \leq \hat{R}^\delta_1 + \mu_0 \) and thus \( \tau \leq t_0 \). It follows that \( R_n \leq t_0 + \hat{R}^\delta_1 \).

**D. Proof of Theorem 9**

**Theorem 9.** Suppose for any \( \mu_i \in [0, 1] \) \((i > 0)\) and \( \mu_0 \) satisfying
\[
\min\{\mu_0, 1 - \mu_0\} \geq \max\{1/2\sqrt{\epsilon}, \sqrt{\epsilon + 1/2} \sqrt{K/n}, \}
\]
an algorithm satisfies \( E_{\mu} \sum_{t=1}^{n} X_{t,i} \geq (1 - \alpha)\mu_0 n \). Then there is some \( \mu \in [0, 1]^K \) such that its expected regret satisfies \( E_{\mu} R_n \geq B \) where
\[
B = \max \left\{ \frac{K}{(16\epsilon + 8)\alpha \mu_0}, \frac{\sqrt{Kn}}{\sqrt{16\epsilon + 8}} \right\}.
\]

**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 \( E_{\mu} R_n > B \) 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 \( E_{\mu} R_n \leq B \) for all \( \mu \in [0, 1]^K \).

For some \( \Delta > 0 \), define environment \( \mu \in [0, 1]^K \) such that \( \mu_i = \mu_0 - \Delta \) for all \( i \in [K] \). For now, assume that \( \mu_0 \) and \( \Delta \) are such that \( \mu_i \geq 0 \); we will get back to this condition later. Also define environment \( \mu^{(i)} \) for each \( i = 1, \ldots, K \) 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.
\section*{Conservative Bandits}

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

In this case, $B = \sqrt{\frac{Kn}{16e + 8}}$ 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 $P_{\mu}(A_i) \geq 1/2$:

\[
P_{\mu}(T_i \leq 2B/\Delta) = 1 - P_{\mu}(T_i > 2B/\Delta)
\]

\[
\geq 1 - \frac{E_{\mu}[T_i]}{2B} \geq 1 - \frac{E_{\mu}[R_n]}{2B} \geq \frac{1}{2}.
\]

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

\[
P_{\mu^{(i)}}(T_i \leq 2B/\Delta) = P_{\mu^{(i)}}(n - T_i \geq n - 2B/\Delta)
\]

\[
\leq \frac{E_{\mu^{(i)}}[n - T_i]}{n - 2B/\Delta} \leq \frac{B}{\Delta n - 2B} = \frac{1}{4e}.
\]

Note that $\mu$ and $\mu^{(i)}$ differ only in the $i$th component: $\mu_i = \mu_0 - \Delta$ whereas $\mu_i^{(i)} = \mu_0 + \Delta$. Then the KL divergence between the reward distributions of the $i$th arms is $\text{KL}(\mu_i, \mu_i^{(i)}) = (2\Delta)^2/2 = 2\Delta^2$. 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

\[
E_{\mu}[T_i] \cdot \text{KL}(\mu_i; \mu_i^{(i)}) \geq d(P_{\mu}(A_i), P_{\mu^{(i)}}(A_i)) \geq 1\frac{\log(1/4e)}{2} = \frac{1}{2}
\]

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

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

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

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 $P_{\mu}(A_i) \geq 1/2$:

\[
P_{\mu}(T_i \leq 2\alpha \mu_0 n/\Delta) = 1 - P_{\mu}(T_i > 2\alpha \mu_0 n/\Delta)
\]

\[
\geq 1 - \frac{\Delta E_{\mu}[T_i]}{2\alpha \mu_0 n} \geq 1 - \frac{E_{\mu}[R_n]}{2\alpha \mu_0 n} \geq \frac{1}{2},
\]

where we use the fact that

\[
E_{\mu}[R_n] = n\mu_0 - E_{\mu}\left[\sum_{t=1}^{n} X_{t,t}\right]
\]

\[
\leq n\mu_0 - (1 - \alpha)\mu_0 n = \alpha \mu_0 n.
\]

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

\[
P_{\mu^{(i)}}(T_i \leq 2\alpha \mu_0 n/\Delta)
\]

\[
= P_{\mu^{(i)}}(n - T_i \geq n - 2\alpha \mu_0 n/\Delta)
\]

\[
\leq \frac{E_{\mu^{(i)}}[n - T_i]}{n - 2\alpha \mu_0 n/\Delta}
\]

\[
\leq \frac{B}{\Delta n - 2\alpha \mu_0 n}
\]

\[
= \frac{K}{(4e + 2)K - (32e + 16)\alpha^2 \mu_0^2 n} < \frac{1}{4e}.
\]

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

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

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

To summarize, we proved that

\[
E_{\mu}[R_n] \geq \begin{cases} \sqrt{Kn} & \text{when } \alpha \geq \frac{\sqrt{K}}{\mu_0 \sqrt{(16e + 8)n}} \\ \frac{\sqrt{16e + 8}}{K} & \text{when } \alpha \leq \frac{\sqrt{K}}{\mu_0 \sqrt{(16e + 8)n}} \end{cases}, \quad \text{otherwise},
\]

finishing the proof. \qed