Training a Single Bandit Arm: Supplementary Material

A Proof of Proposition 1

Proof of Proposition 1. For any policy π, we have that

$$\begin{aligned}
\mathcal{R}_T(\pi, \boldsymbol{\nu}) &= \mathrm{E}\big(\max_{i\in[K]} \overline{U}^i_T\big) \\
&= \mathrm{E}\big(\max_{i\in[K]} \big(\sum_{t=1}^{T} U^i_{n^i_{t-1}+1}\mathbb{1}_{\{I_t=i\}}\big)\big) \\
&\overset{(a)}{\leq} \mathrm{E}\left(\sum_{t=1}^{T}\max_{i\in[K]}\left(U^i_{n^i_{t-1}+1}\mathbb{1}_{\{I_t=i\}}\right)\right) \\
&= \sum_{t=1}^{T}\mathrm{E}\left(\max_{i\in[K]}\left(U^i_{n^i_{t-1}+1}\mathbb{1}_{\{I_t=i\}}\right)\right) \\
&\overset{(b)}{=} \sum_{t=1}^{T}\mathrm{E}\left(U^{I_t}_{n^{I_t}_{t-1}+1}\max_{i\in[K]}\left(\mathbb{1}_{\{I_t=i\}}\right)\right) \\
&= \sum_{t=1}^{T}\mathrm{E}\left(U^{I_t}_{n^{I_t}_{t-1}+1}\right) \\
&= \sum_{t=1}^{T}\mathrm{E}\left(\mathrm{E}\left(U^{I_t}_{n^{I_t}_{t-1}+1}\mid\mathcal{H}_t\right)\right) \\
&\overset{(c)}{=} \sum_{t=1}^{T}\mathrm{E}\left(\mu_{I_t}\right) \leq \mu_1 T.
\end{aligned}$$

Here, (a) is obtained due to pushing the max inside the sum; (b) is obtained because $U^i_{n^i_{t-1}+1} \geq 0$ for all i; and (c) holds because the reward for an arm in a period is independent of the past history of play and observations. Thus, the reward of $\mu_1 T$ is the highest that one can obtain under any policy. And this reward can, in fact, be obtained by the policy of always picking arm 1. This shows that

$$\sup_{\pi\in\Pi}\mathcal{R}_T(\pi,\boldsymbol{\nu}) = \mathcal{R}^*_T(\boldsymbol{\nu}).$$

□

B Proof of Theorem 1

The proof of Theorem 1 relies on the following key result.

Proposition 2. *Consider a class $\mathcal{V} = \mathcal{M}^K$ of K-armed stochastic bandits and let $(\pi_T)_{T\in\mathbb{N}}$ be a consistent sequence of policies for $\mathcal{V}$. Then, for all $\alpha \in (0,1]$ and $\boldsymbol{\nu} \in \mathcal{V}$ such that the optimal arm k^* is unique,*

$$\liminf_{T\to\infty} \frac{\mathrm{E}_{\boldsymbol{\nu}}\left[n^i_{\lceil T^\alpha\rceil}\right]}{\log(T)} \geq \frac{\alpha}{d_{\inf}(\nu_i, \mu^*, \mathcal{M})}$$

holds for each suboptimal arm $i \neq k^*$ in $\boldsymbol{\nu}$, where μ^* is the highest mean.

Proof of Proposition 2. In what follows, we denote $\mathrm{P}_{\boldsymbol{\nu}}$ to be the probability distribution induced by the policy π on events until time T under bandit $\boldsymbol{\nu}$, and we let $\mathrm{E}_{\boldsymbol{\nu}}$ denote the corresponding expectation.

Let $\mathrm{Reg}_{\mathrm{SUM},T}(\pi, \boldsymbol{\nu})$ denote the expected regret of the sum objective after T pulls of policy π under the bandit instance $\boldsymbol{\nu}$, which can be defined as

$$\mathrm{Reg}_{\mathrm{SUM},T}(\pi, \boldsymbol{\nu}) = \mu^* T - \mathrm{E}_{\boldsymbol{\nu}}\big(\sum_{t=1}^{T} X_t\big) \tag{5}$$

$$= \mu^* T - \mathrm{E}_{\boldsymbol{\nu}}\big(\sum_{i=1}^{K} \overline{U}^i_T\big), \tag{6}$$

where $X_t = U^{I_t}_{n^{I_t}_t}$, which is the reward due to the arm pulled at time t, and $\overline{U}^i_t = \sum_{n=1}^{n^i_t} U^i_n$, which is the cumulative reward obtained from arm i until time t. We need the following two lemmas for our proof.

Lemma 1. *Fix $\alpha \in (0,1]$ and a policy π. Consider a K-armed bandit instance $\boldsymbol{\nu}$ with $\mu^* \triangleq \mu_1 \geq \mu_2 \geq \cdots \geq \mu_K$. Fix a suboptimal arm i and let $A_i = \left\{n^i_{\lceil T^\alpha\rceil} > \frac{T^\alpha}{2}\right\}$. Then,*

$$\mathrm{Reg}_{\mathrm{SUM},T}(\pi, \boldsymbol{\nu}) > \mathrm{P}_{\boldsymbol{\nu}}(A_i)\frac{T^\alpha \Delta_i}{2}.$$

Lemma 2. *Fix $\alpha \in (0,1]$ and a policy π. Consider a K-armed bandit instance $\boldsymbol{\nu}$ with $\mu^* \triangleq \mu_1 \geq \mu_2 \geq \cdots \geq \mu_K$. Fix a suboptimal arm i and construct another K-armed bandit instance $\boldsymbol{\nu}'$ satisfying $\mu'_i > \mu^* = \mu_1 \geq \mu_2 \geq \cdots \geq \mu_{i-1} \geq \mu_{i+1} \geq \cdots \geq \mu_K$. Let $A^c_i = \left\{n^i_{\lceil T^\alpha\rceil} \leq \frac{T^\alpha}{2}\right\}$. Then,*

$$\mathrm{Reg}_{\mathrm{SUM},T}(\pi, \boldsymbol{\nu}') \geq \mathrm{P}_{\boldsymbol{\nu}'}(A^c_i)\frac{T^\alpha(\mu'_i - \mu^*)}{2}.$$

The proof of Lemma 1 is presented below at the end of this section. The proof of Lemma 2 is similar and hence is omitted.

Fix $\alpha \in (0,1]$. We proceed by constructing a second bandit $\boldsymbol{\nu}'$. Fix a suboptimal arm i, i.e., $\Delta_i > 0$, and let $\nu'_j = \nu_j$ for $j \neq i$ and pick a $\nu'_i \in \mathcal{M}$ such that $\mathrm{D}(\nu_i, \nu'_i) \leq d_i + \epsilon$ and $\mu'_i > \mu^*$ for some arbitrary $\epsilon > 0$.

Let μ_i (μ') be the mean of arm i in $\boldsymbol{\nu}$ ($\boldsymbol{\nu}'$) and $d_i \triangleq d_{\inf}(\nu_i, \mu^*, \mathcal{M})$. Recall that $d_{\inf}(\nu, \mu^*, \mathcal{M}) = \inf_{\nu'\in\mathcal{M}} \{\mathrm{D}(\nu, \nu') : \mu(\nu') > \mu^*\}$ where $\mu(\nu)$ denotes the mean of distribution ν.

Since any lower bound on the regret for the sum objective implies the same lower bound on the max objective,

using Lemma 1 and Lemma 2, we have the following:

$$\begin{aligned}
\mathrm{Reg}_T(\pi,\boldsymbol{\nu}) + \mathrm{Reg}_T(\pi,\boldsymbol{\nu'}) &\geq \mathrm{Reg}_{\mathrm{SUM},T}(\pi,\boldsymbol{\nu}) + \mathrm{Reg}_{\mathrm{SUM},T}(\pi,\boldsymbol{\nu'}) \\
&> \frac{T^\alpha}{2}\left(\mathrm{P}_{\boldsymbol{\nu}}\left(A_i\right)\Delta_i + \mathrm{P}_{\boldsymbol{\nu'}}\left(A_i^c\right)\left(\mu_i' - \mu^*\right)\right) \\
&\geq \frac{T^\alpha}{2}\min\{\Delta_i, (\mu_i' - \mu^*)\}\left(\mathrm{P}_{\boldsymbol{\nu}}\left(A_i\right) + \mathrm{P}_{\boldsymbol{\nu'}}\left(A_i^c\right)\right) \\
&= \frac{T^\alpha}{2}\min\{\Delta_i, (\mu_i' - \mu^*)\}\left(\overline{\mathrm{P}}_{\boldsymbol{\nu}}\left(A_i\right) + \overline{\mathrm{P}}_{\boldsymbol{\nu'}}\left(A_i^c\right)\right) \\
&\geq \frac{T^\alpha}{4}\min\{\Delta_i, (\mu_i' - \mu^*)\}\exp\left(-\mathrm{E}_{\boldsymbol{\nu}}\left[n^i_{\lceil T^\alpha\rceil}\right](d_i+\epsilon)\right). \qquad (7)
\end{aligned}$$

Here, $\overline{\mathrm{P}}_{\boldsymbol{\nu}}$ ($\overline{\mathrm{P}}_{\boldsymbol{\nu'}}$) is the probability distribution induced by the policy π on events until time $\lceil T^\alpha\rceil$ under bandit $\boldsymbol{\nu}$ ($\boldsymbol{\nu'}$). The equality then results from the fact that the two events $\{n^i_{\lceil T^\alpha\rceil} > \frac{T^\alpha}{2}\}$ and $\{n^i_{\lceil T^\alpha\rceil} \leq \frac{T^\alpha}{2}\}$ depend only on the play until time $\lceil T^\alpha\rceil$. The last inequality follows from using the Bretagnolle-Huber inequality and divergence decomposition (see Theorem 14.2 and Lemma 15.1 in Lattimore and Szepesvári (2018), respectively) combined with the fact that $\mathrm{D}(\nu_i, \nu_i') \leq d_i + \epsilon$:

$$\overline{\mathrm{P}}_{\boldsymbol{\nu}}\left(A_i\right) + \overline{\mathrm{P}}_{\boldsymbol{\nu'}}\left(A_i^c\right) \geq \frac{1}{2}\exp\left(-\mathrm{D}\left(\overline{\mathrm{P}}_{\boldsymbol{\nu}}, \overline{\mathrm{P}}_{\boldsymbol{\nu'}}\right)\right) \geq \frac{1}{2}\exp\left(-\mathrm{E}_{\boldsymbol{\nu}}\left[n^i_{\lceil T^\alpha\rceil}\right](d_i+\epsilon)\right), \qquad (8)$$

where the events A_i and A_i^c are defined as they have been in Lemmas 1 and 2 for the fixed arm i.

Rearranging Equation 7, we obtain

$$\frac{\mathrm{E}_{\boldsymbol{\nu}}\left[n^i_{\lceil T^\alpha\rceil}\right]}{\log(T)} > \frac{1}{d_i+\epsilon}\frac{\log\left(\frac{T^\alpha \min\{\Delta_i,\ \mu_i'-\mu^*\}}{4(\mathrm{Reg}_T(\pi,\boldsymbol{\nu})+\mathrm{Reg}_T(\pi,\boldsymbol{\nu'}))}\right)}{\log(T)}, \qquad (9)$$

and taking the limit inferior yields

$$\begin{aligned}
\liminf_{T\to\infty}\frac{\mathrm{E}_{\boldsymbol{\nu}}\left[n^i_{\lceil T^\alpha\rceil}\right]}{\log(T)} &> \frac{1}{d_i+\epsilon}\liminf_{T\to\infty}\frac{\log\left(\frac{T^\alpha \min\{\Delta_i,\ \mu_i'-\mu^*\}}{4(\mathrm{Reg}_T(\pi,\boldsymbol{\nu})+\mathrm{Reg}_T(\pi,\boldsymbol{\nu'}))}\right)}{\log(T)}, \\
&= \frac{1}{d_i+\epsilon}\liminf_{T\to\infty}\frac{\alpha\log(T) + \log(\beta_i) - \log(4) - \log\left(\mathrm{Reg}_T(\pi,\boldsymbol{\nu}) + \mathrm{Reg}_T(\pi,\boldsymbol{\nu'})\right)}{\log(T)} \\
&= \frac{1}{d_i+\epsilon}\left(\alpha - \limsup_{T\to\infty}\frac{\log\left(\mathrm{Reg}_T(\pi,\boldsymbol{\nu}) + \mathrm{Reg}_T(\pi,\boldsymbol{\nu'})\right)}{\log(T)}\right) \\
&\geq \frac{\alpha}{d_i+\epsilon}, \qquad (10)
\end{aligned}$$

where $\beta_i = \min\{\Delta_i, \mu_i' - \mu^*\}$.

Since π is a consistent policy over the class $\mathcal{V}$, we can find a constant c_p for any $p > 0$ such that $\mathrm{Reg}_T(\pi,\boldsymbol{\nu}) + \mathrm{Reg}_T(\pi,\boldsymbol{\nu'}) \leq c_p T^p$, which implies

$$\limsup_{T\to\infty}\frac{\log\left(\mathrm{Reg}_T(\pi,\boldsymbol{\nu}) + \mathrm{Reg}_T(\pi,\boldsymbol{\nu'})\right)}{\log(T)} \leq \limsup_{T\to\infty}\frac{p\log(T) + \log(c_p)}{\log(T)} = p. \qquad (11)$$

Then, Equation 10 follows from Equation 11 and the fact that $p > 0$ is arbitrary. Since $\epsilon > 0$ is arbitrary as well, we have

$$\liminf_{T\to\infty}\frac{\mathrm{E}_{\boldsymbol{\nu}}\left[n^i_{\lceil T^\alpha\rceil}\right]}{\log(T)} \geq \frac{\alpha}{d_i} \qquad (12)$$

for each $i \neq k^*$, i.e., each suboptimal arm i in $\boldsymbol{\nu}$. □

We present the proof of Lemma 1 before proceeding with the proof of Theorem 1.

Proof of Lemma 1. Recall that I_t is the arm pulled at time t and X_t is the reward due to arm pulled at time t, i.e., $X_t \sim \nu_{I_t}$. Then, due to, e.g., Lemma 4.5 in Lattimore and Szepesvári (2018), we can decompose the expected regret as

$$\mathrm{Reg}_{\mathrm{SUM},T}(\pi, \boldsymbol{\nu}) = \sum_{\substack{j=1 \\ j\neq i}}^{K} \Delta_j \mathrm{E}_{\boldsymbol{\nu}}(n_T^j) + \Delta_i \mathrm{E}_{\boldsymbol{\nu}}(n_T^i). \tag{13}$$

Due to the non-negativity of expected number of pulls and the suboptimality gaps, we have

$$\mathrm{Reg}_{\mathrm{SUM},T}(\pi, \boldsymbol{\nu}) \geq \Delta_i \mathrm{E}_{\boldsymbol{\nu}}(n_T^i).$$

Now, we look at $\mathrm{E}_{\boldsymbol{\nu}}(n_T^i)$:

$$\begin{aligned}
\mathrm{E}_{\boldsymbol{\nu}}(n_T^i) &= \mathrm{E}_{\boldsymbol{\nu}}\big(n_T^i \mid A_i\big)\mathrm{P}_{\boldsymbol{\nu}}(A_i) + \mathrm{E}_{\boldsymbol{\nu}}\big(n_T^i \mid A_i^c\big)\mathrm{P}_{\boldsymbol{\nu}}(A_i^c) \\
&\geq \mathrm{E}_{\boldsymbol{\nu}}\big(n_T^i \mid A_i\big)\mathrm{P}_{\boldsymbol{\nu}}(A_i) \\
&\overset{(a)}{>} \frac{T^\alpha}{2}\mathrm{P}_{\boldsymbol{\nu}}(A_i),
\end{aligned} \tag{14}$$

where (a) is due to event $A_i = \left\{ n^i_{\lceil T^\alpha \rceil} > \frac{T^\alpha}{2} \right\}$. Finally, we have

$$\mathrm{Reg}_{\mathrm{SUM},T}(\pi, \boldsymbol{\nu}) > \mathrm{P}_{\boldsymbol{\nu}}(A_i)\frac{T^\alpha \Delta_i}{2}. \tag{15}$$

□

Proof of Theorem 1. Let k^* denote the unique optimal arm in $\boldsymbol{\nu}$ and, without loss of generality, let $k^* = 1$, i.e., $\mu^* = \mu_1$. Let I^* denote the arm with the highest cumulative reward after T pulls and recall that n_T^i denotes the number of pulls spent on arm i until time T. Since all of the following expectations are over $\boldsymbol{\nu}$, we drop the subscript of $\boldsymbol{\nu}$ hereafter. We first look at the expected regret:

$$\mathrm{Reg}_T(\pi, \boldsymbol{\nu}) = \mu^*\, T - \mathrm{E}\left[\max\left(\overline{U}_T^1, \overline{U}_T^2, \ldots, \overline{U}_T^K\right)\right] \tag{16}$$

$$\overset{(a)}{=} \mathrm{E}\left[\sum_{t=1}^{T} U_t^1\right] - \mathrm{E}\left[\max\left(\overline{U}_T^1, \overline{U}_T^2, \ldots, \overline{U}_T^K\right) \mathbb{1}_{\{I^*=1\}}\right] - \mathrm{E}\left[\max\left(\overline{U}_T^1, \overline{U}_T^2, \ldots, \overline{U}_T^K\right) \mathbb{1}_{\{I^*\neq 1\}}\right] \tag{17}$$

$$\overset{(b)}{=} \mathrm{E}\left[\sum_{t=1}^{T} U_t^1\ \mathbb{1}_{\{I^*=1\}}\right] + \mathrm{E}\left[\sum_{t=1}^{T} U_t^1\ \mathbb{1}_{\{I^*\neq 1\}}\right] - \mathrm{E}\left[\sum_{t=1}^{n_T^1} U_t^1\ \mathbb{1}_{\{I^*=1\}}\right] - \sum_{i\neq 1}\mathrm{E}\left[\sum_{t=1}^{n_T^i} U_t^i\ \mathbb{1}_{\{I^*=i\}}\right] \tag{18}$$

$$= \mathrm{E}\left[\left(\sum_{t=1}^{T} U_t^1 - \sum_{t=1}^{n_T^1} U_t^1\right)\ \mathbb{1}_{\{I^*=1\}}\right] + \sum_{i\neq 1}\mathrm{E}\left[\left(\sum_{t=1}^{T} U_t^1 - \sum_{t=1}^{n_T^i} U_t^i\right)\ \mathbb{1}_{\{I^*=i\}}\right] \tag{19}$$

$$\overset{(c)}{\geq} \mathrm{E}\left[\left(\sum_{t=1}^{T} U_t^1 - \sum_{t=1}^{n_T^1} U_t^1\right)\ \mathbb{1}_{\{I^*=1\}}\right] + \sum_{i\neq 1}\mathrm{E}\left[\left(\sum_{t=1}^{T} U_t^1 - \sum_{t=1}^{T} U_t^i\right)\ \mathbb{1}_{\{I^*=i\}}\right] \tag{20}$$

$$= \mathrm{E}\left[\left(\sum_{t=n_T^1+1}^{T} U_t^1\right)\ \mathbb{1}_{\{I^*=1\}}\right] + \sum_{i\neq 1}\mathrm{E}\left[\left(\sum_{t=1}^{T} U_t^1 - \sum_{t=1}^{T} U_t^i\right)\ \mathbb{1}_{\{I^*=i\}}\right] \tag{21}$$

$$\overset{(d)}{=} \mu^*\mathrm{E}\left[(T - n_T^1)\ \mathbb{1}_{\{I^*=1\}}\right] + \sum_{i\neq 1}\mathrm{E}\left[\left(\sum_{t=1}^{T} U_t^1 - \sum_{t=1}^{T} U_t^i\right)\ \mathbb{1}_{\{I^*=i\}}\right] \tag{22}$$

$$\overset{(e)}{=} \mu^* \sum_{i\neq 1}\mathrm{E}\left[n_T^i\ \mathbb{1}_{\{I^*=1\}}\right] + \sum_{i\neq 1}\mathrm{E}\left[\left(\sum_{t=1}^{T} U_t^1 - \sum_{t=1}^{T} U_t^i\right)\ \mathbb{1}_{\{I^*=i\}}\right]. \tag{23}$$

Here, (a) is due to the fact that $\mathrm{E}\left[U_t^1\right] = \mu^*$ for $t \in [T]$. (b) follows from the definition of I^*. (c) results from $n_T^i \leq T$ for $i \in [K]$. (d) is due to the fact that the future rewards from the first arm is independent of the past history of play and observations of policy π. Finally, (e) follows from the identity $T = \sum_{i=1}^K n_T^i$.

We first focus on bounding the second term in the Expression 23. In order to do that, for each suboptimal arm i, $i \neq 1$, define a "good" event

$$G_i = \left\{ \overline{U}_T^1 > \overline{U}_T^i + \frac{T\Delta_i}{2} \right\}.$$

Notice that, for $i \neq 1$, $\Delta_i > 0$.

We proceed by showing that event G_i occurs with high probability. To that end, consider the complement event

$$\begin{aligned} P\left(G_i^c\right) &= P\left(\overline{U}_T^1 \leq \overline{U}_T^i + \frac{T\Delta_i}{2} \right) \\ &= P\left(\frac{\overline{U}_T^1 - \overline{U}_T^i}{T} - (\mu_1 - \mu_i) \leq \frac{\Delta_i}{2} - (\mu_1 - \mu_i) \right). \end{aligned} \tag{24}$$

By Hoeffding's inequality,

$$P\left(G_i^c\right) \leq \exp\left(-\frac{2T^2 \left(\frac{\Delta_i}{2}\right)^2}{4T} \right) = \exp\left(-\frac{T\Delta_i^2}{8} \right),$$

since $-1 \leq U_{t_1}^1 - U_{t_2}^i \leq 1$ for any pair $t_1, t_2 \in [T]$. We thus also have that

$$P(I^* = i, G_i^c) \leq \exp\left(-\frac{T\Delta_i^2}{8} \right). \tag{25}$$

We then have

$$\begin{aligned} &\mathrm{E}\left[\left(\sum_{t=1}^T U_t^1 - \sum_{t=1}^T U_t^i \right) \mathbb{1}_{\{I^*=i\}} \right] \\ &= \mathrm{E}\left[\left(\sum_{t=1}^T U_t^1 - \sum_{t=1}^T U_t^i \right) \mid I^* = i, G_i \right] P(I^* = i, G_i) + \mathrm{E}\left[\left(\sum_{t=1}^T U_t^1 - \sum_{t=1}^T U_t^i \right) \mid I^* = i, G_i^c \right] P(I^* = i, G_i^c) \\ &\geq \frac{T\Delta_i}{2} P(I^* = i, G_i) - TP(I^* = i, G_i^c) \\ &\geq \frac{T\Delta_i}{2} P(I^* = i, G_i) - \mathrm{O}(1) \qquad (26) \\ &\geq 0 - \mathrm{O}(1). \qquad (27) \end{aligned}$$

Thus the second term in (23) is lower bounded by a (instance-dependent) constant.

Next, we bound the first term in (23). To do so, we first need an upper bound on $P(I^* = i)$ for any $i \neq 1$. By consistency of policy π, we have that $\mathrm{Reg}_T(\pi, \boldsymbol{\nu}) \leq \mathrm{o}(T^p)$ for every $p > 0$. Thus from (23) and (26), for any $i \neq 1$, we have that

$$\mathrm{o}(T^p) \geq \mathrm{E}\left[\left(\sum_{t=1}^T U_t^1 - \sum_{t=1}^T U_t^i \right) \mathbb{1}_{\{I^*=i\}} \right] \geq \frac{T\Delta_i}{2} P(I^* = i, G_i) - \mathrm{O}(1). \tag{28}$$

This implies that for any $i \neq 1$,

$$P(I^* = i, G_i) \leq \mathrm{o}(T^{p-1}), \tag{29}$$

for every $p > 0$. Finally, (25) and (29) together imply that, for any $i \neq 1$, $P(I^* = i) = P(I^* = i, G_i) + P(I^* = i, G_i^c) \leq \mathrm{o}(T^{p-1})$ for every $p > 0$.

Finally, we are ready to derive a lower bound on the first term in the expression (23). For any $\alpha \in (0,1)$, we have

$$\begin{aligned}
\mathrm{E}[n^i_{\lceil T^\alpha \rceil}] &= \mathrm{E}\left[n^i_{\lceil T^\alpha \rceil}\ \mathbb{1}_{\{I^*=1\}}\right] + \mathrm{E}\left[n^i_{\lceil T^\alpha \rceil}\ \mathbb{1}_{\{I^*\neq 1\}}\right] \\
&\leq \mathrm{E}\left[n^i_{\lceil T^\alpha \rceil}\ \mathbb{1}_{\{I^*=1\}}\right] + \lceil T^\alpha \rceil P(I^* \neq 1) \\
&\leq \mathrm{E}\left[n^i_T\ \mathbb{1}_{\{I^*=1\}}\right] + \lceil T^\alpha \rceil P(I^* \neq 1) \\
&\leq \mathrm{E}\left[n^i_T\ \mathbb{1}_{\{I^*=1\}}\right] + \mathrm{o}(T^{\alpha+p-1}), \qquad (30)
\end{aligned}$$

for every $p > 0$. But then from Proposition 2, we have

$$\begin{aligned}
\frac{\alpha}{d_i} &\leq \liminf_{T\to\infty} \frac{\mathrm{E}[n^i_{\lceil T^\alpha \rceil}]}{\log T} \qquad (31) \\
&\leq \liminf_{T\to\infty} \frac{\mathrm{E}\left[n^i_T\ \mathbb{1}_{\{I^*=1\}}\right]}{\log T} + \liminf_{T\to\infty} \frac{\mathrm{o}(T^{\alpha+p-1})}{\log T}. \qquad (32)
\end{aligned}$$

By choosing a p such that $0 < p < 1-\alpha$, we have that $\liminf_{T\to\infty} \frac{\mathrm{o}(T^{\alpha+p-1})}{\log T} = 0$. And thus, for every $\alpha \in (0,1)$, we have

$$\liminf_{T\to\infty} \frac{\mathrm{E}\left[n^i_T\ \mathbb{1}_{\{I^*=1\}}\right]}{\log T} \geq \frac{\alpha}{d_i}, \qquad (33)$$

which implies that

$$\liminf_{T\to\infty} \frac{\mathrm{E}\left[n^i_T\ \mathbb{1}_{\{I^*=1\}}\right]}{\log T} \geq \frac{1}{d_i}. \qquad (34)$$

Finally, putting everything together, from (23), (27), and (34), we have

$$\begin{aligned}
\liminf_{T\to\infty} \frac{\mathrm{Reg}_T(\pi, \boldsymbol{\nu})}{\log T} &\geq \liminf_{T\to\infty} \mu^* \sum_{i\neq 1} \frac{\mathrm{E}\left[n^i_T\ \mathbb{1}_{\{I^*=1\}}\right]}{\log T} - \liminf_{T\to\infty} \sum_{i\neq 1} \frac{\mathrm{O}(1)}{\log T} \\
&\geq \sum_{i\neq 1} \frac{\mu^*}{d_i}. \qquad (35)
\end{aligned}$$

Plugging in the definition of d_i and substituting k^* back in place give the desired result. □

C Proof of Theorem 2

Proof of Theorem 2. First we fix a policy $\pi \in \Pi$. Let $\Delta \stackrel{\Delta}{=} (K-1)^{1/3}/(2T^{1/3})$. We construct two bandit environments with different reward distributions for each of the arms and show that π cannot perform well in both environments simultaneously.

We first specify the reward distribution for the arms in the base environment, denoted as the bandit $\boldsymbol{\nu} = \{\nu_1, \ldots, \nu_K\}$. Assume that the reward for all of the arms have the Bernoulli distribution, i.e., $\nu_i \sim \text{Bernoulli}(\mu_i)$. We let $\mu_1 = \frac{1}{2} + \Delta$, and $\mu_i = \frac{1}{2}$ for $2 \leq i \leq K$. We let $\mathrm{P}_{\boldsymbol{\nu}}$ denote the probability distribution induced over events until time T under policy π in this first environment, i.e., in bandit $\boldsymbol{\nu}$. Let $\mathrm{E}_{\boldsymbol{\nu}}$ denote the expectation under $\mathrm{P}_{\boldsymbol{\nu}}$.

Define $n^i_{\lceil \Delta T \rceil}$ as the (random) number of pulls spent on arm $i \in \{1, \ldots, K\}$ until time $\lceil \Delta T \rceil$ (note that $\sum_{i=1}^K n^i_{\lceil \Delta T \rceil} = \lceil \Delta T \rceil$) under policy π. Specifically, $n^1_{\lceil \Delta T \rceil}$ is the total (random) number of pulls spent on the first arm under policy π until time $\lceil \Delta T \rceil$. Under policy π, let l^* denote the arm in the set $[K] \setminus \{1\}$ that is pulled the least in expectation until time $\lceil \Delta T \rceil$, i.e., $l^* \in \arg\min_{2 \leq i \leq K} \mathrm{E}_{\boldsymbol{\nu}}(n^i_{\lceil \Delta T \rceil})$. Then clearly, we have that $\mathrm{E}_{\boldsymbol{\nu}}(n^{l^*}_{\lceil \Delta T \rceil}) \leq \frac{\lceil \Delta T \rceil}{K-1}$.

Having defined l^*, we can now define the second environment, denoted as the bandit $\boldsymbol{\nu'} = \{\nu'_1, \ldots, \nu'_K\}$. Again, assume that the reward for all of the arms have the Bernoulli distribution, i.e., $\nu'_i \sim \text{Bernoulli}(\mu'_i)$. We let $\mu'_1 = \frac{1}{2} + \Delta$, $\mu'_i = \frac{1}{2}$ for $[2 \leq i \leq K] \setminus \{l^*\}$, and $\mu'_{l^*} = \frac{1}{2} + 2\Delta$. We let $\mathrm{P}_{\boldsymbol{\nu'}}$ denote the probability distribution induced over events until time T under policy π in this second environment, i.e., in bandit $\boldsymbol{\nu'}$. Let $\mathrm{E}_{\boldsymbol{\nu'}}$ denote the expectation under $\mathrm{P}_{\boldsymbol{\nu'}}$.

With some abuse of notation, for any event B, we define:

$$\mathrm{Reg}_T(\pi, \boldsymbol{\nu}, B) = \mu^* T \mathrm{P}_{\boldsymbol{\nu}}(B) - \mathrm{E}_{\boldsymbol{\nu}}\big(\max\big(\overline{U}^1_T, \overline{U}^2_T, \ldots, \overline{U}^K_T\big)\mathbb{1}_B\big). \tag{36}$$

It is then clear that $\mathrm{Reg}_T(\pi, \boldsymbol{\nu}) = \mathrm{Reg}_T(\pi, \boldsymbol{\nu}, B) + \mathrm{Reg}_T(\pi, \boldsymbol{\nu}, B^c)$. We need the following two results for our proof.

Lemma 3. *Fix a policy π. Consider the K-armed bandit instance $\boldsymbol{\nu}$ with Bernoulli rewards and mean vector $\boldsymbol{\mu} = (\frac{1}{2} + \Delta, \frac{1}{2}, \frac{1}{2}, \cdots, \frac{1}{2})$, where $\Delta < \frac{1}{2}$. Consider the event $A = \{n^1_{\lceil \Delta T \rceil} \leq \frac{\Delta T}{2}\}$. Then we have,*

$$\mathrm{Reg}_T(\pi, \boldsymbol{\nu}, A) \geq \frac{\Delta T}{4} \mathrm{P}_{\boldsymbol{\nu}}(A) - 2\sqrt{T \log(KT)} - 2.$$

The proof of Lemma 3 is presented below in this section. A similar argument shows the following.

Lemma 4. *Fix a policy π. Consider the K-armed bandit instance $\boldsymbol{\nu'}$ with Bernoulli rewards and mean vector $\boldsymbol{\mu'} = (\frac{1}{2} + \Delta, \frac{1}{2}, \frac{1}{2}, \cdots, \frac{1}{2}, \frac{1}{2} + 2\Delta)$, where $\Delta < \frac{1}{4}$. Consider the event $A^c = \{n^1_{\lceil \Delta T \rceil} > \frac{\Delta T}{2}\}$. Then we have,*

$$\mathrm{Reg}_T(\pi, \boldsymbol{\nu'}, A^c) \geq \frac{\Delta T}{4} \mathrm{P}_{\boldsymbol{\nu'}}(A^c) - 2\sqrt{T \log(KT)} - 2.$$

The proof of Lemma 4 is omitted since it is almost identical to that of Lemma 3. These two facts result in the following two inequalities:

$$\mathrm{Reg}_T(\pi, \boldsymbol{\nu}, A) \geq \mathrm{P}_{\boldsymbol{\nu}}\left(n^1_{\lceil \Delta T \rceil} \leq \frac{\Delta T}{2}\right) \Omega(\Delta T), \text{ and} \tag{37}$$

$$\mathrm{Reg}_T(\pi, \boldsymbol{\nu'}, A^c) \geq \mathrm{P}_{\boldsymbol{\nu'}}\left(n^1_{\lceil \Delta T \rceil} > \frac{\Delta T}{2}\right) \Omega(\Delta T). \tag{38}$$

Note that here we have ignored the lower order $\sqrt{T \log(KT)}$ terms since $\Delta T = \Theta(T^{2/3} K^{1/3})$. Now, using the Bretagnolle-Huber inequality (see Theorem 14.2 in Lattimore and Szepesvári (2018)), we have,

$$\mathrm{Reg}_T(\pi, \boldsymbol{\nu}, A) + \mathrm{Reg}_T(\pi, \boldsymbol{\nu'}, A^c) \geq \Omega(\Delta T) \left(\mathrm{P}_{\boldsymbol{\nu}}\left(n^1_{\lceil \Delta T \rceil} \leq \frac{\Delta T}{2}\right) + \mathrm{P}_{\boldsymbol{\nu'}}\left(n^1_{\lceil \Delta T \rceil} > \frac{\Delta T}{2}\right)\right) \tag{39}$$

$$= \Omega(\Delta T) \left(\overline{\mathrm{P}}_{\boldsymbol{\nu}}\left(n^1_{\lceil \Delta T \rceil} \leq \frac{\Delta T}{2}\right) + \overline{\mathrm{P}}_{\boldsymbol{\nu'}}\left(n^1_{\lceil \Delta T \rceil} > \frac{\Delta T}{2}\right)\right) \tag{40}$$

$$\geq \Omega(\Delta T) \exp\left(-\mathrm{D}\left(\overline{\mathrm{P}}_{\boldsymbol{\nu}}, \overline{\mathrm{P}}_{\boldsymbol{\nu'}}\right)\right). \tag{41}$$

Here, $\overline{\mathrm{P}}_{\boldsymbol{\nu}}$ ($\overline{\mathrm{P}}_{\boldsymbol{\nu}'}$) is the probability distribution induced by the policy π on events until time $\lceil \Delta T \rceil$ under bandit $\boldsymbol{\nu}$ ($\boldsymbol{\nu}'$). The first equality then results from the fact that the two events $\{n^1_{\lceil \Delta T \rceil} \leq \frac{\Delta T}{2}\}$ and $\{n^1_{\lceil \Delta T \rceil} > \frac{\Delta T}{2}\}$ depend only on the play until time $\lceil \Delta T \rceil$. In the second inequality, which results from the Bretagnolle-Huber inequality, $\mathrm{D}\left(\overline{\mathrm{P}}_{\boldsymbol{\nu}}, \overline{\mathrm{P}}_{\boldsymbol{\nu}'}\right)$ is the relative entropy, or the Kullback-Leibler (KL) divergence between the distributions $\overline{\mathrm{P}}_{\boldsymbol{\nu}}$ and $\overline{\mathrm{P}}_{\boldsymbol{\nu}'}$ respectively. We can upper bound $\mathrm{D}\left(\overline{\mathrm{P}}_{\boldsymbol{\nu}}, \overline{\mathrm{P}}_{\boldsymbol{\nu}'}\right)$ as,

$$\mathrm{D}\left(\overline{\mathrm{P}}_{\boldsymbol{\nu}}, \overline{\mathrm{P}}_{\boldsymbol{\nu}'}\right) = \mathrm{E}_{\boldsymbol{\nu}}(n^{l^*}_{\lceil \Delta T \rceil})\mathrm{D}\left(\nu_{l^*}, \nu'_{l^*}\right) \leq \frac{\lceil \Delta T \rceil}{K-1}\mathrm{D}\left(\nu_{l^*}, \nu'_{l^*}\right) \lessapprox \frac{8\Delta^3 T}{K-1}, \tag{42}$$

where ν_{l^*} (ν'_{l^*}) denotes the reward distribution of arm l^* in the first (second) environment. The first equality results from divergence decomposition (see Lemma 15.1 in Lattimore and Szepesvári (2018)) and the fact no arm other than l^* offers any distinguishability between $\boldsymbol{\nu}$ and $\boldsymbol{\nu}'$. The next inequality follows from the fact that $\mathrm{E}_{\boldsymbol{\nu}}[n^{l^*}_{\lceil \Delta T \rceil}] \leq (\lceil \Delta T \rceil)/(K-1)$, since by definition, l^* is the arm that is pulled the least in expectation until time $\lceil \Delta T \rceil$ in bandit $\boldsymbol{\nu}$ under π. Now $\mathrm{D}\left(\nu_{l^*}, \nu'_{l^*}\right)$ is simply the relative entropy between the distributions Bernoulli(1/2) and Bernoulli($1/2 + 2\Delta$), which, by elementary calculations, can be shown to be at most $8\Delta^2$, resulting in the final inequality. Thus, we finally have,

$$\mathrm{Reg}_T(\pi, \boldsymbol{\nu}, A) + \mathrm{Reg}_T(\pi, \boldsymbol{\nu}', A^c) \geq \Omega(\Delta T)\exp\left(-\frac{8\Delta^3 T}{K-1}\right).$$

Substituting $\Delta = (K-1)^{1/3}/(2T^{1/3})$ gives

$$\mathrm{Reg}_T(\pi, \boldsymbol{\nu}, A) + \mathrm{Reg}_T(\pi, \boldsymbol{\nu}', A^c) \geq \Omega\left((K-1)^{1/3}T^{2/3}\right). \tag{43}$$

Equation 43 along with

$$\mathrm{Reg}_T(\pi, \boldsymbol{\nu}, A^c) \geq -\mathrm{O}(\sqrt{T\log(KT)}) \text{ and} \tag{44}$$

$$\mathrm{Reg}_T(\pi, \boldsymbol{\nu}', A) \geq -\mathrm{O}(\sqrt{T\log(KT)}), \tag{45}$$

imply that

$$\mathrm{Reg}_T(\pi, \boldsymbol{\nu}) + \mathrm{Reg}_T(\pi, \boldsymbol{\nu}') \geq \Omega\left((K-1)^{1/3}T^{2/3}\right). \tag{46}$$

Finally, using $2\max\{a, b\} \geq a + b$ gives the desired lower bound on the regret.

Showing Equations 44 and 45 is an easy exercise:

$$\begin{aligned}
\mathrm{Reg}_T(\pi, \boldsymbol{\nu}, A^c) &= \mu^* T\mathrm{P}_{\boldsymbol{\nu}}(A^c) - \mathrm{E}_{\boldsymbol{\nu}}\big(\max\left(\overline{U}^1_T, \overline{U}^2_T, \ldots, \overline{U}^K_T\right)\mathbb{1}_{A^c}\big) \\
&\geq \mu^* T\mathrm{P}_{\boldsymbol{\nu}}(A^c) - \mathrm{E}_{\boldsymbol{\nu}}\big(\max\big(\sum_{t=1}^T U^1_t, \sum_{t=1}^T U^2_t, \ldots, \sum_{t=1}^T U^K_t\big)\mathbb{1}_{A^c}\big) \\
&\overset{(a)}{\geq} \mu^* T\mathrm{P}_{\boldsymbol{\nu}}(A^c) - \mu^* T\mathrm{P}_{\boldsymbol{\nu}}(A^c) - 2\sqrt{T\log(KT)} - 2 \\
&= -2\sqrt{T\log(KT)} - 2.
\end{aligned} \tag{47}$$

Here, (a) follows from an argument essentially identical to the one in the proof of Lemma 3 below and we do not repeat it here for brevity. Similarly, we can show that

$$\mathrm{Reg}_T(\pi, \boldsymbol{\nu}', A) \geq -2\sqrt{T\log(KT)} - 2. \tag{48}$$

□

Proof of Lemma 3. We first have that

$$\mathrm{E}_{\boldsymbol{\nu}}\big(\max\left(\overline{U}^1_T, \overline{U}^2_T, \ldots, \overline{U}^K_T\right)\mathbb{1}_A\big) = \mathrm{E}_{\boldsymbol{\nu}}\big(\max\big(\sum_{t=1}^{n^1_T} U^1_t, \sum_{t=1}^{n^2_T} U^2_t, \ldots, \sum_{t=1}^{n^K_T} U^K_t\big)\mathbb{1}_A\big) \tag{49}$$

$$\leq \mathrm{E}_{\boldsymbol{\nu}}\big(\max\big(\sum_{t=1}^{T-\lceil \frac{T\Delta}{2} \rceil} U^1_t, \sum_{t=1}^{T} U^2_t, \ldots, \sum_{t=1}^{T} U^K_t\big)\mathbb{1}_A\big). \tag{50}$$

Defining $T_1 = T - \lceil \frac{T\Delta}{2} \rceil$, and $T_i = T$ for all $i > 1$, consider the "good" event

$$G = \left\{ |\sum_{t=1}^{T_j} U_t^j - \mu_j T_j| \leq \sqrt{T \log(KT)} \text{ for all } j \right\}.$$

Since $U_t^j \in [0,1]$, by Hoeffding's inequality, we have that for any $T' \leq T$,

$$\begin{aligned}
\mathrm{P}_{\boldsymbol{\nu}} \left(|\sum_{t=1}^{T'} U_t^j - \mu_j T'| \leq \sqrt{T \log(KT)} \right) &\geq 1 - 2\exp(-\frac{2(\sqrt{T\log(KT)})^2}{T'}) \\
&\geq 1 - 2\exp(-\frac{2(\sqrt{T\log(KT)})^2}{T}) \\
&= 1 - \frac{2}{K^2T^2} \geq 1 - \frac{2}{KT}.
\end{aligned}$$

Hence, by the union bound we have that $P(G) \geq 1 - \frac{2}{T}$. Thus we finally have,

$$\begin{aligned}
&\mathrm{E}_{\boldsymbol{\nu}}\big(\max\big(\sum_{t=1}^{T-\lceil \frac{T\Delta}{2} \rceil} U_t^1, \sum_{t=1}^{T} U_t^2, \ldots, \sum_{t=1}^{T} U_t^K \big) \mathbb{1}_A \big) \\
&= \mathrm{E}_{\boldsymbol{\nu}}\big(\max\big(\sum_{t=1}^{T-\lceil \frac{T\Delta}{2} \rceil} U_t^1, \sum_{t=1}^{T} U_t^2, \ldots, \sum_{t=1}^{T} U_t^K \big) \mathbb{1}_{A,\,G} \big) \\
&\quad + \mathrm{E}_{\boldsymbol{\nu}}\big(\max\big(\sum_{t=1}^{T-\lceil \frac{T\Delta}{2} \rceil} U_t^1, \sum_{t=1}^{T} U_t^2, \ldots, \sum_{t=1}^{T} U_t^K \big) \mathbb{1}_A \mid G^c \big) \mathrm{P}_{\boldsymbol{\nu}}(G^c) \\
&\leq \mathrm{E}_{\boldsymbol{\nu}} \left(\left(\max_{i \in [K]} \mu_i T_i + 2\sqrt{T\log(KT)} \right) \mathbb{1}_{A,\,G} \right) + \frac{2}{T} \times T \\
&\leq \max\left((\frac{1}{2} + \Delta)(T - \frac{\Delta T}{2}), \frac{T}{2} \right) \mathrm{P}_{\boldsymbol{\nu}}(A) + 2\sqrt{T\log(KT)} + 2 \\
&\overset{(a)}{=} (\frac{1}{2} + \Delta)(T - \frac{\Delta T}{2}) \mathrm{P}_{\boldsymbol{\nu}}(A) + 2\sqrt{T\log(KT)} + 2.
\end{aligned} \tag{51}$$

Here, (a) follows from the fact that $\Delta < \frac{1}{2}$. Thus, from Equations 50 and 51, we finally have,

$$\begin{aligned}
\mathrm{Reg}_T(\pi, \boldsymbol{\nu}, A) &= (\frac{1}{2} + \Delta) T \mathrm{P}_{\boldsymbol{\nu}}(A) - \mathrm{E}_{\boldsymbol{\nu}}\big(\max\big(\overline{U}_T^1, \overline{U}_T^2, \ldots, \overline{U}_T^K \big) \mathbb{1}_A \big) \\
&\geq (\frac{1}{2} + \Delta) T \mathrm{P}_{\boldsymbol{\nu}}(A) - (\frac{1}{2} + \Delta)(T - \frac{\Delta T}{2}) \mathrm{P}_{\boldsymbol{\nu}}(A) - 2\sqrt{T\log(KT)} - 2 \\
&= (\frac{1}{2} + \Delta) \frac{\Delta T}{2} \mathrm{P}_{\boldsymbol{\nu}}(A) - 2\sqrt{T\log(KT)} - 2 \\
&\geq \frac{\Delta T}{4} \mathrm{P}_{\boldsymbol{\nu}}(A) - 2\sqrt{T\log(KT)} - 2.
\end{aligned} \tag{52}$$

□

D Proof of Theorem 3

The proof of Theorem 3 utilizes two technical lemmas. The first one is the following.

Lemma 5. *Let $\delta \in (0,1)$, and $X_1, X_2, \ldots,$ be a sequence of independent 0-mean 1-Sub-Gaussian random variables. Let $\bar{\mu}_t = \frac{1}{t}\sum_{s=1}^{t} X_s$. Then for any $x > 0$,*

$$P\left(\exists\, t > 0 : \bar{\mu}_t + \sqrt{\frac{4}{t}\log^+\left(\frac{1}{\delta t^{3/2}}\right)} + x < 0\right) \leq \frac{39\delta}{x^3}.$$

Its proof is similar to the proof of Lemma 9.3 in Lattimore and Szepesvári (2018), which we present below for completeness.

Proof of Lemma 5. We have,

$$\begin{aligned}
&P\left(\exists\, t > 0 : \bar{\mu}_t + \sqrt{\frac{4}{t}\log^+\left(\frac{1}{\delta t^{3/2}}\right)} + x < 0\right) \\
&= P\left(\exists\, t > 0 : t\bar{\mu}_t + \sqrt{4t\log^+\left(\frac{1}{\delta t^{3/2}}\right)} + tx < 0\right) \\
&\leq \sum_{i=0}^{\infty} P\left(\exists\, t \in \left[2^i, 2^{i+1}\right] : t\bar{\mu}_t + \sqrt{4t\log^+\left(\frac{1}{\delta t^{3/2}}\right)} + tx < 0\right) \\
&\leq \sum_{i=0}^{\infty} P\left(\exists\, t \in \left[0, 2^{i+1}\right] : t\bar{\mu}_t + \sqrt{2^{i+2}\log^+\left(\frac{1}{\delta 2^{(i+1)\cdot 3/2}}\right)} + 2^i x < 0\right) \\
&\leq \sum_{i=0}^{\infty} \exp\left(-\frac{\left(\sqrt{2^{i+2}\log^+\left(\frac{1}{\delta 2^{(i+1)\cdot 3/2}}\right)} + 2^i x\right)^2}{2^{i+2}}\right) \\
&\leq \delta \sum_{i=0}^{\infty} 2^{(i+1)\cdot 3/2}\exp\left(-2^{i-2}x^2\right), \qquad (53)
\end{aligned}$$

where the first inequality follows from a union bound on a geometric grid. The second inequality is used to set up the argument to apply Theorem 9.2 in Lattimore and Szepesvári (2018) and the third inequality is due to its application. The fourth inequality follows from $(a+b)^2 \geq a^2+b^2$ for $a, b \geq 0$. Then, using a property of unimodal functions ($\sum_{j=c}^{d} f(j) \leq \max_{i\in[c,d]} f(i) + \int_c^d f(i)di$ for a unimodal function f), the term $2^{(i+1)\cdot 3/2}\exp\left(-2^{i-2}x^2\right)$ can be upper bounded by $\frac{42\delta}{e^{3/2}x^3} + \delta\int_0^{\infty}(2^{3/2})^{i+1}\exp(-x^2 2^{i-2})di$. Evaluating the integral to $\frac{8\sqrt{2\pi}}{\log(2)}\frac{1}{x^3}$, we get

$$P\left(\exists\, t > 0 : \bar{\mu}_t + \sqrt{\frac{4}{t}\log\frac{1}{\delta t^{3/2}}} + x < 0\right) \leq \frac{39\delta}{x^3}. \qquad (54)$$

□

The second result we need is Lemma 8.2 from Lattimore and Szepesvári (2018), which we present below for completeness.

Lemma 6. *Lattimore and Szepesvári (2018) Let $X_1, X_2, \cdots,$ be a sequence of independent 0-mean 1-Sub-Gaussian random variables. Let $\bar{\mu}_t = \frac{1}{t}\sum_{s=1}^{t} X_s$. Let $\epsilon > 0$, and $a > 0$, and define*

$$\kappa = \sum_{t=1}^{T} \mathbb{1}\{\bar{\mu}_t + \sqrt{\frac{2a}{t}} > \epsilon\}.$$

Then $\mathrm{E}[\kappa] \leq 1 + \frac{2}{\epsilon^2}(a + \sqrt{a\pi} + 1)$.

Proof of Theorem 3. Let 1 denote the first arm and i^* denote the arm used in the Commit phase of ADA-ETC. We first define a random variable that quantifies the lowest value of the index of arm 1 can take with respect to its true mean across τ pulls.

$$\Delta \triangleq \left(\mu_1 - \min_{n\leq\tau}\left(\bar{\mu}_n^1 + \sqrt{\frac{4}{n}\log\left(\frac{T}{Kn^{3/2}}\right)}\mathbb{1}_{\{n<\tau\}}\right)\right)^+.$$

The following bound is instrumental for our analysis. For any $x \geq 0$,

$$\begin{aligned}
P(\Delta > x) &= P\left(\exists\, n \leq \tau : \bar{\mu}_n^1 + \sqrt{\frac{4}{n}\log\left(\frac{T}{Kn^{3/2}}\right)}\mathbb{1}_{\{n<\tau\}} < \mu_1 - x\right) \\
&\leq P\left(\exists\, n < \tau : \bar{\mu}_n^1 + \sqrt{\frac{4}{n}\log\left(\frac{T}{Kn^{3/2}}\right)} < \mu_1 - x\right) + P\left(\bar{\mu}_\tau^1 < \mu_1 - x\right) \\
&\overset{(a)}{\leq} \min(1, \frac{39K}{Tx^3} + \exp(-2\tau x^2)) \qquad (55) \\
&\overset{(b)}{\leq} \min(1, \frac{40K}{Tx^3}). \qquad (56)
\end{aligned}$$

Here, (a) follows from Lemma 5 and Hoeffding's inequality, and (b) follows by the definition of τ and since $\exp(-2\alpha^{2/3}) \leq 1/\alpha$ for all $\alpha \geq 0$.

We next decompose the regret into the regret from wasted pulls in the Explore phase and the regret from committing to a suboptimal arm in the Commit phase. Let ω be the random time when the Explore phase ends. Let r_ω^i be the reward earned from arm i until time ω. Then the expected regret in the event that $\{i^* = i\}$ is bounded by:

$$\mathrm{E}\left(\left(T\mu_1 - (T - \sum_{j\neq i} n_\omega^j - n_\omega^i)\mu_i - r_\omega^i\right)\mathbb{1}_{\{i^*=i\}}\right). \qquad (57)$$

Note that this expression assumes that the cumulative reward of arm i will be chosen to compete against $T\mu_1$ at the end of time T; however, if there is an arm with a higher cumulative reward, then the resulting regret can only be lower. Thus the total expected regret is bounded by:

$$\begin{aligned}
&\sum_{i=1}^K \mathrm{E}\left(\left(T\mu_1 - (T - \sum_{j\neq i} n_\omega^j - n_\omega^i)\mu_i - r_\omega^i\right)\mathbb{1}_{\{i^*=i\}}\right) \\
&\overset{(a)}{\leq} \sum_{i=1}^K \mathrm{E}(T\Delta_i \mathbb{1}_{\{i^*=i\}}) + \mu_1 \sum_{i=1}^K \mathrm{E}(n_\omega^i \mathbb{1}_{\{i^*\neq i\}}) + \sum_{i=1}^K \mathrm{E}\Big((n_\omega^i\mu_i - r_\omega^i)\mathbb{1}_{\{i^*=i\}}\Big) \\
&\overset{(b)}{=} \sum_{i=1}^K \mathrm{E}(T\Delta_i \mathbb{1}_{\{i^*=i\}}) + \mu_1 \sum_{i=1}^K \mathrm{E}(n_\omega^i \mathbb{1}_{\{i^*\neq i\}}) + \sum_{i=1}^K \mathrm{E}\Big((\tau\mu_i - r_\omega^i)\mathbb{1}_{\{i^*=i\}}\Big) \\
&= \sum_{i=1}^K \mathrm{E}(T\Delta_i \mathbb{1}_{\{i^*=i\}}) + \mu_1 \sum_{i=1}^K \mathrm{E}(n_\omega^i \mathbb{1}_{\{i^*\neq i\}}) + \sum_{i=1}^K P(i^*=i)(\tau\mu_i - \mathrm{E}(r_\omega^i \mid i^*=i)) \\
&\overset{(c)}{=} \sum_{i=1}^K \mathrm{E}(T\Delta_i \mathbb{1}_{\{i^*=i\}}) + \mu_1 \sum_{i=1}^K \mathrm{E}(n_\omega^i \mathbb{1}_{\{i^*\neq i\}}) + \sum_{i=1}^K P(i^*=i)(\tau\mu_i - \sum_{n=1}^\tau \mathrm{E}(U_n^i \mid i^*=i)) \\
&\overset{(d)}{\leq} \underbrace{\sum_{i=1}^K \mathrm{E}(T\Delta_i \mathbb{1}_{\{i^*=i\}})}_{\text{Regret from misidentifications in Commit phase}} + \underbrace{\mu_1 \sum_{i=1}^K \mathrm{E}(n_\omega^i \mathbb{1}_{\{i^*\neq i\}})}_{\text{Regret from wasted pulls in the Explore phase}}. \qquad (58)
\end{aligned}$$

Here, (a) results from rearranging terms, and from the fact that $\mu_i \leq \mu_1$. Both (b) and (c) result from the fact that in the event that $\{i^* = i\}$, $n_\omega^i = \tau$. (d) holds since, by a standard stochastic dominance argument, $\tau\mu_i \leq \sum_{n=1}^\tau \mathrm{E}(U_n^i \mid i^* = i)$.

We bound these two terms one by one.

Regret from Explore. First, note that an instance-independent bound on the regret from Explore is simply $K\tau = K\lceil \frac{T^{2/3}}{K^{2/3}} \rceil = \mathrm{O}(K^{1/3}T^{2/3})$, which is the maximum number of pulls possible before ADA-ETC enters the Commit phase. Hence, we now focus on deriving an instance-dependent bound. We have that

$$\begin{aligned}\mathrm{E}(\sum_{i=1}^{K} n_\omega^i \mathbb{1}_{\{i^* \neq i\}}) &\leq \mathrm{E}(\sum_{i=2}^{K} n_\omega^i) + \tau P(i^* \neq 1) \\ &= \mathrm{E}(\sum_{i \geq 2 : \Delta \leq \frac{\Delta_i}{2}} n_\omega^i) + \mathrm{E}(\sum_{i \geq 2 : \Delta > \frac{\Delta_i}{2}} n_\omega^i) + \tau P(i^* \neq 1). \end{aligned} \tag{59}$$

We first bound the first term. Define the random variable

$$\eta_i = \sum_{n=1}^{\tau} \mathbb{1}\left\{ \bar{\mu}_n^i + \sqrt{\frac{4}{n}\log\left(\frac{T}{Kn^{3/2}}\right)} \mathbb{1}_{\{n<\tau\}} \geq \mu_i + \frac{\Delta_i}{2} \right\}.$$

Then in the event that $\Delta \leq \frac{\Delta_i}{2}$, we have that $n_\omega^i \leq \eta_i$. We also have that $n_\omega^i \leq \tau$. And thus in the event that $\Delta \leq \frac{\Delta_i}{2}$, we have $n_\omega^i \leq \min(\eta_i, \tau)$. Hence the first term above is bounded as:

$$\sum_{i=2}^{K} P(\Delta \leq \frac{\Delta_i}{2}) \mathrm{E}(\min(\eta_i, \tau)) \leq \sum_{i=2}^{K} P(\Delta \leq \frac{\Delta_i}{2}) \min(\mathrm{E}(\eta_i), \tau) \leq \sum_{i=2}^{K} \min(\mathrm{E}(\eta_i), \tau).$$

We can now bound $\mathrm{E}(\eta_i)$ as follows:

$$\begin{aligned}\mathrm{E}(\eta_i) &\leq 1 + \mathrm{E}\left(\sum_{n=1}^{\tau-1} \mathbb{1}\left\{ \bar{\mu}_n^i + \sqrt{\frac{4}{n}\log\left(\frac{T}{Kn^{3/2}}\right)} \geq \mu_i + \frac{\Delta_i}{2} \right\}\right) \\ &= 1 + \mathrm{E}\left(\sum_{n=1}^{\tau-1} \mathbb{1}\left\{ \bar{\mu}_n^i + \sqrt{\frac{4}{n}\log^+\left(\frac{T}{Kn^{3/2}}\right)} \geq \mu_i + \frac{\Delta_i}{2} \right\}\right) \\ &\overset{(a)}{\leq} 1 + \frac{1}{\Delta_i^2} + \mathrm{E}\left(\sum_{n=1}^{\tau-1} \mathbb{1}\left\{ \bar{\mu}_n^i + \sqrt{\frac{4}{n}\log^+\left(\frac{T\Delta_i^3}{K}\right)} \geq \mu_i + \frac{\Delta_i}{2} \right\}\right) \\ &\overset{(b)}{\leq} 1 + \frac{1}{\Delta_i^2} + \frac{8}{\Delta_i^2}\left(2\log^+\left(\frac{T\Delta_i^3}{K}\right) + \sqrt{2\pi\log^+\left(\frac{T\Delta_i^3}{K}\right)} + 1\right) \\ &\leq \frac{10}{\Delta_i^2} + \frac{16}{\Delta_i^2}\log^+\left(\frac{T\Delta_i^3}{K}\right) + \frac{24}{\Delta_i^2}\sqrt{\log^+\left(\frac{T\Delta_i^3}{K}\right)}. \end{aligned} \tag{60}$$

Here, (a) is due to lower bounding $1/n^{3/2}$ by Δ_i^3, and adding $1/\Delta^2$ for the first $1/\Delta^2$ time periods where this lower bound doesn't hold. (b) is due to Lemma 6. The final inequality results from the fact that $\Delta_i \leq 1$ and from trivially bounding $2\pi \leq 9$. Thus, we finally have,

$$\mathrm{E}(\sum_{i \geq 2 : \Delta \leq \frac{\Delta_i}{2}} n_\omega^i) \leq \sum_{i=2}^{K} \min\left(\frac{10}{\Delta_i^2} + \frac{16}{\Delta_i^2}\log^+\left(\frac{T\Delta_i^3}{K}\right) + \frac{24}{\Delta_i^2}\sqrt{\log^+\left(\frac{T\Delta_i^3}{K}\right)}, \tau\right). \tag{61}$$

We now focus on the second term in Equation 59. Note that we have $n_\omega^i \leq \tau$, and hence,

$$\mathrm{E}(\sum_{i \geq 2 : \Delta > \frac{\Delta_i}{2}} n_\omega^i) \leq \tau \sum_{i=2}^{K} P(\Delta > \frac{\Delta_i}{2}) \leq \tau \sum_{i=2}^{K} \min(1, \frac{320K}{T\Delta_i^3}). \tag{62}$$

Here the second inequality follows from Equation 56. Next, we focus on the third term in Equation 59. We have:

$$\begin{aligned} P(i^* \neq 1) &= P(i^* \neq 1 \text{ and } \Delta \leq \frac{\Delta_2}{2}) + P(i^* \neq 1 \text{ and } \Delta > \frac{\Delta_2}{2}) \\ &\leq \min\left(1, \sum_{i=2}^{K} P(i^* = i \text{ and } \Delta \leq \frac{\Delta_2}{2}) + P(\Delta > \frac{\Delta_2}{2})\right) \\ &\leq \min\left(1, \sum_{i=2}^{K} P(i^* = i \text{ and } \Delta \leq \frac{\Delta_2}{2}) + \frac{320K}{T\Delta_2^3}\right). \end{aligned} \tag{63}$$

Here the final inequality again follows from Equation 56. Now in the event that $\Delta \leq \Delta_2/2$, $i^* = i$ implies that there is some $n \leq \tau$ such that $\text{LCB}_n^i = \bar{\mu}_n^i - \bar{\mu}_n^i \mathbb{1}_{\{n<\tau\}} > \mu_i + \Delta_i/2$. Thus, we have,

$$\begin{aligned} \sum_{i=2}^{K} P(i^* = i \text{ and } \Delta \leq \frac{\Delta_2}{2}) &\leq \sum_{i=2}^{K} P\left(\exists\, n \leq \tau : \bar{\mu}_n^i - \bar{\mu}_n^i \mathbb{1}_{\{n<\tau\}} > \mu_i + \Delta_i/2\right) \\ &= \sum_{i=2}^{K} P(\bar{\mu}_\tau^i > \mu_i + \Delta_i/2) \\ &\overset{(a)}{\leq} \sum_{i=2}^{K} \exp(-\frac{\tau\Delta_i^2}{2}) \overset{(b)}{\leq} \sum_{i=2}^{K} \frac{8K}{T\Delta_i^3}. \end{aligned} \tag{64}$$

Here, (a) follows from Hoeffding's inequality, and (b) follows from the definition of τ and the fact that $\exp(-\alpha^{2/3}/2) \leq 8/\alpha$ for $\alpha \geq 0$. Thus we finally have

$$\begin{aligned} \tau P(i^* \neq 1) &\leq \tau \min(1, \sum_{i=2}^{K} \frac{8K}{T\Delta_i^3} + \frac{320K}{T\Delta_2^3}) \\ &\leq \tau \min(1, \sum_{i=2}^{K} \frac{328K}{T\Delta_i^3}). \end{aligned} \tag{65}$$

Thus, combining Equations 61, 62, and 65, we have that the regret from the Explore phase is bounded by

$$\begin{aligned} &\mu_1 \sum_{i=2}^{K} \min\left(\frac{10}{\Delta_i^2} + \frac{16}{\Delta_i^2}\log^+\left(\frac{T\Delta_i^3}{K}\right) + \frac{24}{\Delta_i^2}\sqrt{\log^+\left(\frac{T\Delta_i^3}{K}\right)}, \tau\right) \\ &+ \mu_1\tau \sum_{i=2}^{K} \min(1, \frac{320K}{T\Delta_i^3}) + \mu_1\tau \min(1, \sum_{i=2}^{K} \frac{328K}{T\Delta_i^3}) \\ &\leq \mu_1 \sum_{i=2}^{K} \min\left(\frac{10}{\Delta_i^2} + \frac{16}{\Delta_i^2}\log^+\left(\frac{T\Delta_i^3}{K}\right) + \frac{24}{\Delta_i^2}\sqrt{\log^+\left(\frac{T\Delta_i^3}{K}\right)}, \tau\right) \\ &+ \mu_1\tau \sum_{i=2}^{K} \min(2, \frac{628K}{T\Delta_i^3}). \end{aligned} \tag{66}$$

Here the inequality results from the fact that $\min(1, a) + \min(1, b) \leq \min(2, a + b)$ for a, $b > 0$. This finishes our derivation of a distribution dependent bound on the regret from the Explore phase. We next focus on the regret arising from misidentification in the Commit phase.

Regret from Commit. This regret is upper bounded by

$$\text{E}(\sum_{i:\Delta \leq \frac{\Delta_i}{2}} \mathbb{1}_{\{i^*=i\}} T\Delta_i) + \text{E}(\sum_{i:\Delta > \frac{\Delta_i}{2}} \mathbb{1}_{\{i^*=i\}} T\Delta_i). \tag{67}$$

We now get instance dependent and independent bounds on each of the first two terms.

An instance dependent bound on $\mathbf{E}(\sum_{i:\Delta\leq\frac{\Delta_i}{2}}\mathbb{1}_{\{i^*=i\}}T\Delta_i)$. In the event that $\Delta\leq\Delta_i/2$, $i^*=i$ implies that there is some $n\leq\tau$ such that $\text{LCB}_n^i=\bar{\mu}_n^i-\bar{\mu}_n^i\mathbb{1}_{\{n<\tau\}}>\mu_i+\Delta_i/2$. Thus, we have,

$$\mathrm{E}(\sum_{i:\Delta\leq\frac{\Delta_i}{2}}\mathbb{1}_{\{i^*=i\}}T\Delta_i)\leq\sum_{i=2}^{K}P\left(\exists\, n\leq\tau:\bar{\mu}_n^i-\bar{\mu}_n^i\mathbb{1}_{\{n<\tau\}}>\mu_i+\Delta_i/2\right)T\Delta_i. \tag{68}$$

Now, we have,

$$P(\exists\, n\leq\tau:\bar{\mu}_n^i-\bar{\mu}_n^i\mathbb{1}_{\{n<\tau\}}>\mu_i+\Delta_i/2)=P(\bar{\mu}_\tau^i>\mu_i+\Delta_i/2)\leq\exp(-\frac{\tau\Delta_i^2}{2}). \tag{69}$$

Here the final inequality follows from Hoeffding's inequality. Thus we finally have,

$$\mathrm{E}(\sum_{i:\Delta\leq\frac{\Delta_i}{2}}\mathbb{1}_{\{i^*=i\}}T\Delta_i)\leq\sum_{i=2}^{K}\exp(-\frac{\tau\Delta_i^2}{2})T\Delta_i. \tag{70}$$

An instance independent bound on $\mathbf{E}(\sum_{i:\Delta\leq\frac{\Delta_i}{2}}\mathbb{1}_{\{i^*=i\}}T\Delta_i)$. We have

$$\begin{aligned}
\mathrm{E}(\sum_{i:\Delta<\frac{\Delta_i}{2}}\mathbb{1}_{\{i^*=i\}}T\Delta_i)&=T^{2/3}K^{1/3}\sqrt{2\log K}+\mathrm{E}(\sum_{i:\Delta<\frac{\Delta_i}{2};\Delta_i\geq\frac{K^{1/3}\sqrt{2\log K}}{T^{1/3}}}\mathbb{1}_{\{i^*=i\}}T\Delta_i)\\
&\overset{(a)}{\leq}T^{2/3}K^{1/3}\sqrt{2\log K}+\mathrm{E}(\sum_{i:\Delta_i\geq\frac{K^{1/3}\sqrt{2\log K}}{T^{1/3}}}\exp(-\frac{\tau\Delta_i^2}{2})T\Delta_i)\\
&\overset{(b)}{\leq}T^{2/3}K^{1/3}\sqrt{2\log K}+T^{2/3}K^{1/3}\sqrt{2\log K}.
\end{aligned} \tag{71}$$

Here, (a) follows for the same reason as the derivation of the bound in Equation 70. Next, observe that the function $\exp(-\frac{\tau x^2}{2})x$ is maximized at $x^*=\sqrt{2/\tau}=\sqrt{2}K^{1/3}/T^{1/3}$. But since $\Delta_i\geq\sqrt{2\log K}K^{1/3}/T^{1/3}\geq\sqrt{2}K^{1/3}/T^{1/3}$, by the unimodality of $\exp(-\frac{\tau x^2}{2})x$, we have

$$\exp(-\frac{\tau\Delta_i^2}{2})T\Delta_i\leq\exp(-\log K)T^{2/3}K^{1/3}\sqrt{2\log K}=\frac{1}{K}T^{2/3}K^{1/3}\sqrt{2\log K}.$$

Hence (b) follows.

An instance dependent bound on $\mathbf{E}(\sum_{i:\Delta>\frac{\Delta_i}{2}}\mathbb{1}_{\{i^*=i\}}T\Delta_i)$.

$$\begin{aligned}
\mathrm{E}(\sum_{i:\Delta>\frac{\Delta_i}{2}}\mathbb{1}_{\{i^*=i\}}T\Delta_i)&\leq\mathrm{E}(\max_{i\in[K]}T\Delta_i\mathbb{1}_{\{\Delta>\frac{\Delta_i}{2}\}})\\
&=P(\Delta>\frac{\Delta_K}{2})T\Delta_K+\sum_{i=1}^{K-1}P(\frac{\Delta_{i+1}}{2}\geq\Delta>\frac{\Delta_i}{2})T\Delta_i\\
&=P(\Delta>\frac{\Delta_K}{2})T\Delta_K+\sum_{i=1}^{K-1}\left(P(\Delta>\frac{\Delta_i}{2})-P(\Delta>\frac{\Delta_{i+1}}{2})\right)T\Delta_i\\
&=\sum_{i=2}^{K}P(\Delta>\frac{\Delta_i}{2})T(\Delta_i-\Delta_{i-1})\\
&\leq\sum_{i=2}^{K}\min(1,\frac{320K}{T\Delta_i^3})T(\Delta_i-\Delta_{i-1}).
\end{aligned} \tag{72}$$

Here the final inequality again follows from Equation 56.

An instance independent bound on $\mathbf{E}(\sum_{i:\Delta>\frac{\Delta_i}{2}} \mathbb{1}_{\{i^*=i\}} T\Delta_i)$. We have,

$$\mathrm{E}(\sum_{i:\Delta>\frac{\Delta_i}{2}} \mathbb{1}_{\{i^*=i\}} T\Delta_i) \leq \mathrm{E}(2T\Delta \sum_{i=1}^{K} \mathbb{1}_{\{i^*=i\}}) = \mathrm{E}(2T\Delta) = 2T\mathrm{E}(\Delta). \tag{73}$$

We then look at $\mathrm{E}(\Delta)$. We have,

$$\mathrm{E}(\Delta) = \int_0^\infty \mathrm{P}\left(\Delta > x\right) dx \leq \int_0^\infty \min\left(1, \frac{40K}{Tx^3}\right) dx.$$

This integral evaluates to

$$\int_0^{\frac{(40K)^{1/3}}{T^{1/3}}} dx + \int_{\frac{(40K)^{1/3}}{T^{1/3}}}^\infty \frac{40K}{Tx^3} dx \leq 2\frac{(40K)^{1/3}}{T^{1/3}}.$$

Combining these results, we have

$$\mathrm{E}(\Delta) \leq 2\frac{(40K)^{1/3}}{T^{1/3}}. \tag{74}$$

Thus we finally have,

$$\mathrm{E}(\sum_{i:\Delta>\frac{\Delta_i}{2}} \mathbb{1}_{\{i^*=i\}} T\Delta_i) \leq 4(40K)^{1/3}T^{2/3}. \tag{75}$$

The final instance-dependent bound follows from Equations 66, 70, and 72. The instance-independent bound follows from the fact that the regret from the Explore phase is at most $K\tau = \mathrm{O}(T^{2/3}K^{1/3})$ and from Equations 71 and 75. □