Supplementary Material for ”Combinatorial multi-armed bandit: general framework, results and applications”, by Wei Chen, Yajun Wang, and Yang Yuan.

A. Full proof of Theorem 1

We use the following two well known bounds in our proofs.

Lemma 1 (Chernoff-Hoeffding bound)**.** *Let $X_1, \cdots, X_n$ be random variables with common support $[0,1]$ and $\mathbb{E}[X_i] = \mu$. Let $S_n = X_1 + \cdots + X_n$. Then for all $t \geq 0$,*

$$\Pr[S_n \geq n\mu + t] \leq e^{-2t^2/n} \text{ and } \Pr[S_n \leq n\mu - t] \leq e^{-2t^2/n}$$

Lemma 2 (Bernstein inequality)**.** *Let $X_1, \ldots, X_n$ be independent zero-mean random variables. If for all $1 \leq i \leq n, |X_i| \leq k$, then for all $t > 0$,*

$$\Pr\left[\left|\sum_{i=1}^{n} X_i\right| > t\right] \leq \exp\left\{-\frac{t^2/2}{\sum_{i=1}^{n}\mathbb{E}[X_i^2] + kt/3}\right\}.$$

For a given underlying arm $i \in [m]$, Let $\mathcal{S}_{i,B}$ be the set of all bad super arms containing arm i. We sort all bad super arms in $\mathcal{S}_{i,B}$ as $S_{i,\mathrm{B}}^1, S_{i,\mathrm{B}}^2, \ldots, S_{i,\mathrm{B}}^{K_i}$ in increasing order of their expected rewards, where K_i is the number of bad super arms containing arm i. Define

$$\Delta^{i,j} = \alpha \cdot \mathrm{opt}_{\boldsymbol{\mu}} - r_{\boldsymbol{\mu}}(S_{i,\mathrm{B}}^j). \tag{14}$$

Thus $\Delta_{\max}^i = \Delta^{i,1}$ and $\Delta_{\min}^i = \Delta^{i,K_i}$. Recall that $\Delta_{\max} = \max_{i\in[m]} \Delta_{\max}^i$.

Theorem 1 (restated) *The (α, β)-approximation regret of the CUCB algorithm in n rounds using an (α, β)-approximation oracle is at most*

$$\begin{aligned}&\sum_{i\in[m],\Delta_{\min}^i>0}\left(\frac{6\ln n}{(f^{-1}(\Delta_{\min}^i))^2}\cdot\Delta_{\min}^i + \int_{\Delta_{\min}^i}^{\Delta_{\max}^i}\frac{6\ln n}{(f^{-1}(x))^2}\mathrm{d}x\right) + \left(\frac{\pi^2}{3}+1\right)\cdot m\cdot\Delta_{\max}\\ \leq &\sum_{i\in[m],\Delta_{\min}^i>0}\frac{6\ln n}{(f^{-1}(\Delta_{\min}^i))^2}\cdot\Delta_{\max}^i + \left(\frac{\pi^2}{3}+1\right)\cdot m\cdot\Delta_{\max},\end{aligned}$$

where $f(\cdot)$ is the bounded smoothness function.

Our proof depends on the fact that with high probability, the entire process behaves nicely. In other words, the empirical mean of X_i is close to the actual expectation μ_i.

Definition 1 (Nice process)**.** *The process is* nice *at time horizon t if:*

$$\forall i \in [m], \mid \hat{\mu}_{i,T_{i,t-1}} - \mu_i \mid < \sqrt{\frac{3\ln t}{2T_{i,t-1}}}.$$

Lemma 3. *The probability that the process is* nice *at time t is at least $1 - 2mt^{-2}$.*

Proof. By Chernoff-Hoeffding bound in Lemma 1, for any $i \in [m]$,

$$\begin{aligned}&\Pr\left[\mid \hat{\mu}_{i,T_{i,t-1}} - \mu_i \mid \geq \sqrt{\frac{3\ln t}{2T_{i,t-1}}}\right] &(15)\\ =&\sum_{s=1}^{t-1}\Pr\left[\left\{\mid \hat{\mu}_{i,s} - \mu_i \mid \geq \sqrt{\frac{3\ln t}{2s}}, T_{i,t-1} = s\right\}\right]\\ \leq&\sum_{s=1}^{t-1}\Pr\left[\left\{\mid \hat{\mu}_{i,s} - \mu_i \mid \geq \sqrt{\frac{3\ln t}{2s}}\right\}\right]\\ \leq& t\cdot 2e^{-3\ln t} = \frac{2}{t^2}. &(16)\end{aligned}$$

The lemma follows by taking union bound on i. □

We now briefly explain the idea to prove Theorem 1, based on the refinement of the idea used to prove the simplified regret bound in Eq.(4). In the proof of Eq.(4), we essentially show that if all arms are sufficiently sampled with respect to $\Delta_{\min}$, that is, sampled at least $\frac{6\ln t}{(f^{-1}(\Delta_{\min}))^2}$ times, then the sample means are close enough to their true mean values. As a result, by the monotonicity and bounded smoothness properties of the expected reward function and by the property of the approximation oracle, we know that the probability that we hit a bad super arm is very small. On the other hand, in a bad round, if the underlying arms are not sufficiently sampled with respect to $\Delta_{\min}$, we incur a regret of $\Delta_{\max}$. Notice that there is a discrepancy in the analysis, i.e., the sufficiency of sampling is defined on $\Delta_{\min}$ while the regret is counted as $\Delta_{\max}$. This makes our analysis of regret bound in Eq.(4) not tight enough.

In this section, we refine the previous analysis. In particular, for each arm i, it has a series of bad super arms $S^1_{i,\mathrm{B}}, S^2_{i,\mathrm{B}}, \ldots, S^{K_i}_{i,\mathrm{B}}$ containing i, and for each $S^l_{i,\mathrm{B}}$, we define sufficient sampling of i with respect to $S^l_{i,\mathrm{B}}$ (or equivalently with respect to $\Delta^{i,l}$) as i being sampled $\frac{6\ln n}{(f^{-1}(\Delta^{i,l}))^2}$ times and i's counter N_i being incremented in these sampled instances, where n is the time horizon. We show that when i is sufficiently sampled with respect to $S^l_{i,\mathrm{B}}$, the probability that $S^l_{i,\mathrm{B}}$ is selected by the oracle is very small. On the other hand, in a bad round when i's counter N_i is incremented, if i is under-sampled with respect to $S^l_{i,\mathrm{B}}$, we incur a regret of at most $\Delta^{i,j}$ for some $j \leq l$. In this way, we reduce the discrepancy between $\Delta_{\min}$ and $\Delta_{\max}$ to a much tighter $\Delta^{i,l}$ and $\Delta^{i,j}$, which enables us to prove the much tighter bound given in Theorem 1.

Proof of Theorem 1. For variable T_i, let $T_{i,t}$ be the value of T_i at the end of round t, that is, $T_{i,t}$ is the number of times arm i is played in the first t rounds. For variable $\hat{\mu}_i$, let $\hat{\mu}_{i,s}$ be the value of $\hat{\mu}_i$ after arm i is played s times, that is, $\hat{\mu}_{i,s} = (\sum_{j=1}^{s} X_{i,j})/s$. Then, the value of variable $\hat{\mu}_i$ at the end of round t is $\hat{\mu}_{i,T_{i,t}}$. For variable $\bar{\mu}_i$, let $\bar{\mu}_{i,t}$ be the value of $\bar{\mu}_i$ at the end of round t, and let $\bar{\boldsymbol{\mu}}_t = (\mu_{1,t}, \mu_{2,t}, \ldots, \mu_{m,t})$ be the input vector to the oracle at round t. Then, according to line 6 of Algorithm 1, we have

$$\bar{\mu}_{i,t} = \hat{\mu}_{i,T_{i,t-1}} + \sqrt{\frac{3\ln t}{2T_{i,t-1}}}. \tag{17}$$

For the proof, we maintain counter N_i for each arm i after the m initialization rounds. Let $N_{i,t}$ be the value of N_i after the t-th round and $N_{i,m} = 1$. Note that $\sum_i N_{i,m} = m$. Counters $\{N_i\}_{i=1}^m$ are updated in the following way.

For a round $t > m$, let S_t be the super arm selected in round t by the oracle (line 7 of Algorithm 1). Round t is a *bad* round if the oracle selects a super arm $S_t \in \mathcal{S}_\mathrm{B}$, which is not an α-approximate super arm with respect to the true mean vector $\boldsymbol{\mu}$. If round t is bad, let $i = \operatorname{argmin}_{j\in S_t} N_{j,t-1}$. We increment N_i by one, i.e., $N_{i,t} = N_{i,t-1} + 1$. In other words, we find the arm i with smallest counter in S_t and increase its counter. If i is not unique, we pick an arbitrary arm with smallest counters in S_t. By definition $N_{i,t} \leq T_{i,t}$. Notice that in every bad round, exactly one counter in $\{N_i\}_{i=1}^m$ is increased.

Each time N_i gets updated, one of the bad super arms containing i is played. We further divide counter N_i into more counters $\{N_i^l\}_{l=1}^{K_i}$, whose value at round n, $N^l_{i,n}$ is defined as follows:

$$\forall l \in [K_i],\; N^l_{i,n} = \sum_{t=m+1}^{n} \mathbb{I}\{S_t = \mathcal{S}^l_{i,\mathrm{B}}, N_{i,t} > N_{i,t-1}\}.$$

Define $\ell_n(\Delta) = \frac{6\ln n}{(f^{-1}(\Delta))^2}$, i.e., the number of sampling that is considered *sufficient* for a super arm with reward Δ away from the α-approximation with respect to time horizon n. When counter N_i^l is incremented at time t, i.e., $S_t = \mathcal{S}^l_{i,\mathrm{B}}$ and $N_{i,t} > N_{i,t-1}$, we inspect the value $N_{i,t-1}$. Notice that every arm in S_t must have been played at least $N_{i,t-1}$ times by round t, since in our updating rule we choose the smallest counter value among arms in S_t to update, and i is the chosen one. If $N_{i,t-1} > \ell_n(\Delta^{i,l})$, we say that the bad arm $\mathcal{S}^l_{i,\mathrm{B}}$ is *sufficiently sampled*. Otherwise, it is *under-sampled*. Notice that our definition of sufficient sampling at a time t is with respect to the value $\ell_n(\Delta^{i,l})$ related to the time horizon n, not the current time t. This is a bit different from the proof of Eq.(4) in the main text, and it is needed in the analysis of the under-sampled case. In our analysis, the time t considered is at most n.

We write as

$$
\begin{aligned}
N_{i,n}^{l,suf} &= \sum_{t=m+1}^{n} \mathbb{I}\{S_t = \mathcal{S}_{i,\mathrm{B}}^l, N_{i,t} > N_{i,t-1}, N_{i,t-1} > \ell_n(\Delta^{i,l})\}, \\
N_{i,n}^{l,und} &= \sum_{t=m+1}^{n} \mathbb{I}\{S_t = \mathcal{S}_{i,\mathrm{B}}^l, N_{i,t} > N_{i,t-1}, N_{i,t-1} \le \ell_n(\Delta^{i,l})\}.
\end{aligned}
$$

Then we have $N_{i,n} = 1 + \sum_{l\in[K_i]}(N_{i,n}^{l,suf} + N_{i,n}^{l,und})$. Using this notation, the total reward at time horizon n is at least

$$
n \cdot \alpha \cdot \mathrm{opt}_{\boldsymbol{\mu}} - \mathbb{E}\left[\sum_{i\in[m]}\left(\Delta^{i,1} + \sum_{l\in[K_i]}(N_{i,n}^{l,suf} + N_{i,n}^{l,und}) \cdot \Delta^{i,l}\right)\right], \tag{18}
$$

where $\Delta^{i,1}$ is for the initialization.

We claim that it is unlikely that a bad super arm is played when all the underlying arms are sufficiently sampled. More specifically, we have the following claim.

Claim 1. *For any time horizon $n > m$,*

$$
\mathbb{E}\left[\sum_{i\in[m]}\sum_{l\in[K_i]} N_{i,n}^{l,suf}\right] \le (1-\beta)n + \frac{\pi^2}{3}\cdot m. \tag{19}
$$

Proof. By the definition of $N_{i,n}^{l,suf}$, it is sufficient to show that for any $n \ge t > m$,

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{i\in[m],l\in[K_i]} \mathbb{I}\{S_t = \mathcal{S}_{i,\mathrm{B}}^l, N_{i,t} > N_{i,t-1}, N_{i,t-1} > \ell_n(\Delta^{i,l})\}\right] \\
=&\sum_{i\in[m],l\in[K_i]} \Pr\{S_t = \mathcal{S}_{i,\mathrm{B}}^l, N_{i,t} > N_{i,t-1}, \forall s \in \mathcal{S}_{i,\mathrm{B}}^l, N_{s,t-1} > \ell_n(\Delta^{i,l})\} \\
\le&(1-\beta) + 2mt^{-2}.
\end{aligned}
$$

Define $\Lambda_{i,t} = \sqrt{\frac{3\ln t}{2T_{i,t-1}}}$ (a random variable since $T_{i,t-1}$ is a random variable) and $\Lambda_t = \max\{\Lambda_{i,t} \mid i \in S_t\}$. Define $\Lambda^{i,l} = \sqrt{\frac{3\ln t}{2\ell_n(\Delta^{i,l})}}$ (not a random variable).

Let $\mathcal{N}_t$ indicate the event that the process is *nice* at time t. Let F_t be the event that the oracle fails to produce an α-approximate answer with respect to input vector $\bar{\boldsymbol{\mu}}_t$ in round t. We have $\Pr[F_t] = \mathbb{E}[\mathbb{I}\{F_t\}] \le 1-\beta$.

Notice that $\bar{\mu}_{i,t} = \hat{\mu}_{i,t} + \sqrt{\frac{3\ln t}{2T_{i,t-1}}}$. We have the following properties.

$$
\begin{aligned}
\mathcal{N}_t \Rightarrow \forall i \in [m],\ \bar{\mu}_{i,t} - \mu_i > 0, \qquad (20)\\
\mathcal{N}_t \Rightarrow \forall i \in S_t,\ \bar{\mu}_{i,t} - \mu_i < 2\Lambda_t, \qquad (21)\\
\forall i \in [m], \forall l \in [K_i], \left\{S_t = S_{i,\mathrm{B}}^l, N_{i,t} > N_{i,t-1}, \forall s \in S_t, N_{s,t-1} > \ell_n(\Delta^{i,l})\right\} \Rightarrow \Lambda^{i,l} > \Lambda_t. \qquad (22)
\end{aligned}
$$

For any particular $i \in [m]$ and $l \in [K_i]$, if $\left\{\mathcal{N}_t, \neg F_t, S_t = \mathcal{S}_{i,\mathrm{B}}^l, N_{i,t} > N_{i,t-1}, \forall s \in S_t, N_{s,t-1} > \ell_n(\Delta^{i,l})\right\}$ holds at time t, we have the following properties:

$$\begin{aligned}
r_{\boldsymbol{\mu}}(S_t) + f(2\Lambda^{i,l}) >& r_{\boldsymbol{\mu}}(S_t) + f(2\Lambda_t) && \text{strict monotonicity of } f(\cdot) \text{ and Eq.(22)} \\
\geq& r_{\bar{\boldsymbol{\mu}}_t}(S_t) && \text{bounded smoothness property and Eq.(21)} \\
\geq& \alpha \cdot \mathrm{opt}_{\bar{\boldsymbol{\mu}}_t} && \neg F_t \Rightarrow S_t \text{ is an } \alpha \text{ approximation w.r.t } \bar{\boldsymbol{\mu}}_t \\
\geq& \alpha \cdot r_{\bar{\boldsymbol{\mu}}_t}(S^*_{\boldsymbol{\mu}}) && \text{definition of } \mathrm{opt}_{\bar{\boldsymbol{\mu}}_t} \\
\geq& \alpha \cdot r_{\boldsymbol{\mu}}(S^*_{\boldsymbol{\mu}}) = \alpha \cdot \mathrm{opt}_{\boldsymbol{\mu}}. && \text{monotonicity of } r_{\boldsymbol{\mu}}(S) \text{ and Eq.(20)}
\end{aligned}$$

So we have

$$r_{\boldsymbol{\mu}}(S^l_{i,\mathrm{B}}) + f(2\Lambda^{i,l}) > \alpha \cdot \mathrm{opt}_{\boldsymbol{\mu}}. \tag{23}$$

Since $\ell_n(\Delta^{i,l}) = \frac{6 \ln n}{(f^{-1}(\Delta^{i,l}))^2}$, we have $2\Lambda^{i,l} = f^{-1}(\Delta^{i,l}) \cdot \sqrt{\frac{\ln t}{\ln n}}$, which implies $f(2\Lambda^{i,l}) \leq \Delta^{i,l}$ by the monotonicity of $f(\cdot)$ and $t \leq n$. Therefore, Eq. (23) contradicts the definition of $\Delta^{i,l}$ in Eq.(14). In other words,

$$\begin{aligned}
&\forall i \in [m]\, \forall l \in [K_i],\ \Pr\left\{\mathcal{N}_t, \neg F_t, S_t = S^l_{i,\mathrm{B}}, N_{i,t} > N_{i,t-1}, \forall s \in S_t,\ N_{s,t-1} > \ell_n(\Delta^{i,l})\right\} = 0 \\
&\Rightarrow \Pr\left\{\mathcal{N}_t, \neg F_t, \exists i \in [m], \exists l \in [K_i], S_t = S^l_{i,\mathrm{B}}, N_{i,t} > N_{i,t-1}, \forall s \in S_t,\ N_{s,t-1} > \ell_n(\Delta^{i,l})\right\} = 0 \\
&\Rightarrow \Pr\left\{\exists i \in [m], \exists l \in [K_i], S_t = S^l_{i,\mathrm{B}}, N_{i,t} > N_{i,t-1}, \forall s \in S_t,\ N_{s,t-1} > \ell_n(\Delta^{i,l})\right\} \\
&\quad \leq \Pr[F_t \vee \neg \mathcal{N}_t] \leq (1-\beta) + 2mt^{-2} \qquad (24) \\
&\Rightarrow \sum_{i \in [m], l \in [K_i]} \Pr\left\{S_t = S^l_{i,\mathrm{B}}, N_{i,t} > N_{i,t-1}, \forall s \in S_t,\ N_{s,t-1} > \ell_n(\Delta^{i,l})\right\} \leq (1-\beta) + 2mt^{-2}. \qquad (25)
\end{aligned}$$

The second inequality in Eq. (24) uses the facts that $\Pr\{F_t\} = (1-\beta)$ and $\Pr\{\neg\mathcal{N}_t\} \leq 2mt^{-2}$ (Lemma 3). The left side of Eq. (25) equals the left side of Eq. (24), because the events $\{S_t = S^l_{i,\mathrm{B}}, N_{i,t} > N_{i,t-1}, \forall s \in S_t, N_{s,t-1} > \ell_t(\Delta^{i,l})\}$ for all $i \in [m]$ and $l \in [K_i]$ are mutually exclusive, which in turn is because in each round when S_t is bad, only one arm $i \in S_t$ gets to increment its counter N_i and thus $N_{i,t} > N_{i,t-1}$, and within arm i, only one index l satisfies $S_t = S^l_{i,\mathrm{B}}$. □

Now we consider the bad super arms that are under-sampled when played. For a particular arm i, its counter N_i will increase from 1 to $\ell_n(\Delta^{i,K_i})$. To simplify the notation, set $\ell_n(\Delta^{i,0}) = 0$. (Notice that $N_{i,m} = 1$ for all i.) Before we go into the details, we discuss the essential idea behind Eqn.(28). We break the range of the counter N_i into discrete segments, i.e., $(\ell_n(\Delta^{i,j-1}), \ell_n(\Delta^{i,j})]$ for $j \in [K_i]$. Let us assume that the round t is bad and N_i is incremented. Assume $N_{i,t-1} \in (\ell_n(\Delta^{i,j-1}), \ell_n(\Delta^{i,j})]$ for some j. Notice that we are only interested in the case that S_t is under-sampled. In particular, if this is indeed the case, $S_t = \mathcal{S}^{i,l}_{\mathrm{B}}$ for some $l \geq j$. (Otherwise, S_t is sufficiently sampled based on the counter value $N_{i,t-1}$.) Therefore, we will suffer a regret of $\Delta^{i,l} \leq \Delta^{i,j}$ (Eqn.(26)). Consequently, for counter $N_{i,t}$ in range $(\ell_n(\Delta^{i,j-1}), \ell_n(\Delta^{i,j})]$, we will suffer a total regret for those under-sampled arms at most $(\ell_n(\Delta^{i,j}) - \ell_n(\Delta^{i,j-1})) \cdot \Delta^{i,j}$ (Eqn.(27)) in rounds that $N_{i,t}$ is incremented.

We implement the argument rigorously as follows. For any arm i in $\{i \in [m] \mid \Delta^i_{\min} > 0\}$, we have,

$$\begin{aligned}
\sum_{l\in[K_i]} N_{i,n}^{l,und}\cdot\Delta^{i,l} &= \sum_{t=m+1}^{n}\sum_{l\in[K_i]}\mathbb{I}\{S_t = S_{i,\mathrm{B}}^l, N_{i,t} > N_{i,t-1}, N_{i,t-1} \le \ell_n(\Delta^{i,l})\}\cdot\Delta^{i,l} \\
&= \sum_{t=m+1}^{n}\sum_{l\in[K_i]}\sum_{j=1}^{l}\mathbb{I}\{S_t = S_{i,\mathrm{B}}^l, N_{i,t} > N_{i,t-1}, N_{i,t-1} \in (\ell_n(\Delta^{i,j-1}), \ell_n(\Delta^{i,j})]\}\cdot\Delta^{i,l} \\
&\le \sum_{t=m+1}^{n}\sum_{l\in[K_i]}\sum_{j=1}^{l}\mathbb{I}\{S_t = S_{i,\mathrm{B}}^l, N_{i,t} > N_{i,t-1}, N_{i,t-1} \in (\ell_n(\Delta^{i,j-1}), \ell_n(\Delta^{i,j})]\}\cdot\Delta^{i,\boldsymbol{j}} \qquad (26) \\
&\le \sum_{t=m+1}^{n}\sum_{l\in[K_i]}\sum_{j\in[K_i]}\mathbb{I}\{S_t = S_{i,\mathrm{B}}^l, N_{i,t} > N_{i,t-1}, N_{i,t-1} \in (\ell_n(\Delta^{i,j-1}), \ell_n(\Delta^{i,j})]\}\cdot\Delta^{i,j} \\
&= \sum_{t=m+1}^{n}\sum_{j\in[K_i]}\mathbb{I}\{S_t \in \mathcal{S}_{i,\mathrm{B}}, N_{i,t} > N_{i,t-1}, N_{i,t-1} \in (\ell_n(\Delta^{i,j-1}), \ell_n(\Delta^{i,j})]\}\cdot\Delta^{i,j} \\
&= \sum_{j\in[K_i]}\sum_{t=m+1}^{n}\mathbb{I}\{S_t \in \mathcal{S}_{i,\mathrm{B}}, N_{i,t} > N_{i,t-1}, N_{i,t-1} \in (\ell_n(\Delta^{i,j-1}), \ell_n(\Delta^{i,j})]\}\cdot\Delta^{i,j} \\
&\le \sum_{j\in[K_i]}(\ell_n(\Delta^{i,j}) - \ell_n(\Delta^{i,j-1}))\cdot\Delta^{i,j} \qquad (27) \\
&= \ell_n(\Delta^{i,K_i})\Delta^{i,K_i} + \sum_{j=1}^{K_i-1}\ell_n(\Delta^{i,j})\cdot(\Delta^{i,j} - \Delta^{i,j+1}) \\
&\le \ell_n(\Delta^{i,K_i})\Delta^{i,K_i} + \int_{\Delta^{i,K_i}}^{\Delta^{i,1}}\ell_n(x)\mathrm{d}x. \qquad (28)
\end{aligned}$$

Last inequality comes from the fact that $\ell_n(x)$ is decreasing. Notice that by our definition, for arms that $j \in [m] \setminus \{i \in [m] \mid \Delta_{\min}^i > 0\}$, the counter N_j remains one after the initialization. Since they do not contribute to any regret, we have $K_j = 0$ for all these arms. Therefore, combining with Eq.(19) and Eq.(28), the overall regret of our algorithm is

$$\begin{aligned}
Reg_{\boldsymbol{\mu},\alpha,\beta}^{A}(n) \le& n\cdot\alpha\cdot\beta\cdot\mathrm{opt}_{\boldsymbol{\mu}} - \left(\alpha\cdot n\cdot\mathrm{opt}_{\boldsymbol{\mu}} - \mathbb{E}\left[\sum_{i\in[m]}\left(\Delta^{i,1} + \sum_{l\in[K_i]}(N_{i,n}^{l,suf} + N_{i,n}^{l,und})\cdot\Delta^{i,l}\right)\right)\right] \\
\le& \Delta_{\max}\cdot\left(m + \mathbb{E}\left[\sum_{i\in[m],l\in[K_i]}N_{i,n}^{l,suf}\right]\right) + \sum_{i\in[m],\Delta_{\min}^i>0}\left(\ell_n(\Delta^{i,K_i})\Delta^{i,K_i} + \int_{\Delta^{i,K_i}}^{\Delta^{i,1}}\ell_n(x)\mathrm{d}x\right) \\
&- (1-\beta)\cdot n\cdot\alpha\cdot\mathrm{opt}_{\boldsymbol{\mu}} \\
\le& \left(\frac{\pi^2}{3} + 1\right)\cdot m\cdot\Delta_{\max} + \sum_{i\in[m],\Delta_{\min}^i>0}\left(\ell_n(\Delta^{i,K_i})\Delta^{i,K_i} + \int_{\Delta^{i,K_i}}^{\Delta^{i,1}}\ell_n(x)\mathrm{d}x\right).
\end{aligned}$$

The theorem follows directly. □

Improving the coefficient of the leading term. In general, we can set $\bar{\mu}_i = \hat{\mu}_i + \sqrt{y/(2T_i)}$ for some y in line 6 in the CUCB algorithm. The corresponding regret bound obtained is

$$\sum_{i\in[m],\Delta_{\min}^i>0}\left(\frac{2\cdot y}{(f^{-1}(\Delta_{\min}^i))^2}\cdot\Delta_{\min}^i + \int_{\Delta_{\min}^i}^{\Delta_{\max}^i}\frac{2\cdot y}{(f^{-1}(x))^2}\mathrm{d}x\right) + \left(1 + \sum_{t=m+1}^{n}\frac{2t}{e^{-y}}\right)\cdot m\cdot\Delta_{\max}.$$

What we need is to make sure the term $\sum_{t=m+1}^{n}\frac{2t}{e^{-y}}$ in the above regret bound converges. We can thus set y appropriately to guarantee convergence while improving the constant in the leading term. One way is setting $y = (1+c)\ln t$ with a constant $c > 1$, or equivalently setting $\bar{\mu}_i = \hat{\mu}_i + \sqrt{(1+c)\ln t/(2T_i)}$, so that $\sum_{t=m+1}^{n} \frac{2t}{e^{-y}} = 2\sum_{t=m+1}^{n} t^{-c} \leq 2\zeta(c)$, where $\zeta(c) = \sum_{t=1}^{\infty} \frac{1}{t^c}$ is the Riemann's zeta function, and has a finite value when $c > 1$. Then the regret bound is

$$\sum_{i\in[m],\Delta_{\min}^i>0} \left(\frac{2\cdot(1+c)\cdot\ln n}{(f^{-1}(\Delta_{\min}^i))^2} \cdot \Delta_{\min}^i + \int_{\Delta_{\min}^i}^{\Delta_{\max}^i} \frac{2\cdot(1+c)\cdot\ln n}{(f^{-1}(x))^2} \mathrm{d}x \right) + (2\cdot\zeta(c)+1)\cdot m\cdot\Delta_{\max}.$$

We can also further improve the constant factor from $2(1+c)$ to 4 by setting $\bar{\mu}_i = \hat{\mu}_i + \sqrt{\frac{2\ln t + \ln\ln t}{2T_i}}$ at the cost of a second order $\ln\ln n$ term as in (Garivier & Cappé, 2011), with regret at most

$$\sum_{i\in[m],\Delta_{\min}^i>0} \left(\frac{2\cdot(2\ln n+\ln\ln n)}{(f^{-1}(\Delta_{\min}^i))^2} \cdot \Delta_{\min}^i + \int_{\Delta_{\min}^i}^{\Delta_{\max}^i} \frac{2\cdot(2\ln n+\ln\ln n)}{(f^{-1}(x))^2} \mathrm{d}x \right) + (1+2\ln\ln n)\cdot m\cdot\Delta_{\max}.$$

This is because $\sum_{t=m+1}^{n} \frac{1}{t\ln t} \leq \int_{m}^{n} \frac{1}{t\ln t}\mathrm{d}t \leq \ln\ln n$ when $m > e$.

Theorem 2 (restated) *Consider a CMAB problem with an (α,β)-approximation oracle. If the bounded smoothness function $f(x) = \gamma\cdot x^{\omega}$ for some $\gamma > 0$ and $\omega \in (0,1]$, the regret of CUCB is at most:*

$$\frac{2\gamma}{2-\omega}\cdot(6m\ln n)^{\omega/2}\cdot n^{1-\omega/2} + \left(\frac{\pi^2}{3}+1\right)\cdot m\cdot\Delta_{\max}.$$

In Theorem 1, when Δ_{min}^i is extremely small, the regret would be approaching infinite. This is not applicable since the number of samples needed to be sufficient is exceeding the time horizon. In what follows, we prove a distribution-independent regret bound. The rough idea is, if $\Delta_{\min}^i \leq 1/\sqrt{n}$, it can only contribute $\sqrt{n}$ regret at time horizon n.

Proof. Following the proof of the main theorem, we only need to consider the bad arms that are played when they are under-sampled. Following the intuition, we need to quantify when Δ is too small. In particular, we measure the threshold for $\Delta_{\min}^i$ based on $N_{i,n}$, i.e., the counter of arm i at time horizon n. Let $\{n_i \mid i \in [m]\}$ be a set of possible counter values at time horizon n. Our analysis will then be conditioned on the event that $\{N_{i,n} = n_i\}$. The catch is, in our analysis for under-sampled super arms, that we only need counting based arguments will be still applicable under any condition.

For an arm in $\{i \mid \Delta_{\min}^i > 0\}$, we have

$$\mathbb{E}[\sum_{l\in[K_i]} N_{i,n}^{l,und}\cdot\Delta^{i,l} \mid \{N_{j,n} = n_j\}\,] = \sum_{t=m+1}^{n}\sum_{l\in[K_i]} \mathbb{E}[\mathbb{I}\{S_t = \mathcal{S}_{i,\mathrm{B}}^l, N_{i,t} > N_{i,t-1}, N_{i,t-1} \leq \ell_n(\Delta^{i,l}) \mid \{N_{j,n} = n_j\}\,\}]\cdot\Delta^{i,l}$$

Define $\Delta^*(n_i) = \left(\frac{6\gamma^{2/\omega}\ln n}{n_i}\right)^{\omega/2}$, i.e., $\ell_n(\Delta^*(n_i)) = n_i$. Now we consider two cases. Case (1): $\Delta_{\min}^i > \Delta^*(n_i)$. Following the same counting steps as in Eq.(28), we have

$$\mathbb{E}[\sum_{l\in[K_i]} N_{i,n}^{l,und}\cdot\Delta^{i,l} \mid \{N_{j,n} = n_j\}\,] \leq \frac{2}{2-\omega}\cdot\frac{6\cdot\gamma^{2/\omega}\ln n}{(\Delta_{\min}^i)^{\frac{2}{\omega}-1}} \leq \frac{2\gamma}{2-\omega}\cdot(6\ln n)^{\omega/2} n_i^{1-\omega/2}.$$

Case (2): $\Delta_{\min}^i \leq \Delta^*(n_i)$. Let $l^* = \min\{l \mid \Delta^{i,l} > \Delta^*(n_i)\}$. Notice that $\Delta^{i,l^*} \leq \left(\frac{6\gamma^{2/\omega}\ln n}{n_i}\right)^{\omega/2}$. Since the counter cannot go beyond n_i, we have

$$\begin{aligned}\mathbb{E}[\sum_{l\in[K_i]} N_{i,n}^{l,und}\cdot\Delta^{i,l} \mid \{N_{j,n}=n_j\}] =& \sum_{t=m+1}^{n}\sum_{l\in[K_i]} \mathbb{E}[\mathbb{I}\{S_t=\mathcal{S}_{i,\mathrm{B}}^{l}, N_{i,t}>N_{i,t-1}, N_{i,t-1}\le \ell_n(\Delta^{i,l}) \mid \{N_{j,n}=n_j\}\}]\cdot\Delta^{i,l}\\ \le&(\ell_n(\Delta^*(n_i))-\ell_n(\Delta^{i,l^*-1}))\cdot\Delta^*(n_i)+\sum_{j\in[l^*-1]}(\ell_n(\Delta^{i,j})-\ell_n(\Delta^{i,j-1}))\cdot\Delta^{i,j}\\ \le&\ell_n(\Delta^*(n_i))\cdot\Delta^*(n_i)+\int_{\Delta^*(n_i)}^{\Delta^{i,1}}\ell_n(x)\mathrm{d}x\\ \le&\frac{2\gamma}{2-\omega}\cdot(6\ln n)^{\omega/2}n_i^{1-\omega/2}. \end{aligned}\tag{29}$$

Therefore, Eq.(29) holds in both cases. We then have

$$\begin{aligned}\mathbb{E}\left[\sum_{\{i|\Delta_{\min}^i>0\}}\sum_{l\in[K_i]} N_{i,n}^{l,und}\cdot\Delta^{i,l} \mid \{N_{j,n}=n_j\}\right] &\le \frac{2\gamma}{2-\omega}\cdot(6\ln n)^{\omega/2}\cdot\sum_{\{i|\Delta_{\min}^i>0\}} n_i^{1-\omega/2}\\ &\le \frac{2\gamma}{2-\omega}\cdot(6m\ln n)^{\omega/2}\cdot n^{1-\omega/2}.\end{aligned}$$

The last inequality comes from Jesen's inequality and $\sum_i n_i \le n$. Since the final inequality does not depend on n_i, we can drop the conditional expectation above. The result then follows from Claim 1 and Eq.(18). □

B. CMAB with Clustered Arms

In many applications, multiple arms are clustered and are always played together. For example, in the PMC bandit problem, all arms (edges) incident to a node in L are always played together; in the influence maximization bandit problem, all arms (outgoing edges) from the same node are always played together. In this section, we show how to take advantage of such arm clusters to further improve the regret analysis.

We consider the following CMAB problem with clustered arms. Formally, each cluster $C\subseteq[m]$ contains a set of simple arms. Denote U as the set of all clusters. Notice that one arm may belong to multiple clusters. We assume $|U|<m$. In this setting, each super arm S is a union of several clusters: $S=\bigcup_{C\in g(S)} C$, where $g(S)$ is the set of clusters that forms S. When super arm S is played in round t, the outcomes of all arms in the clusters in S will be revealed.

We will use the same CUCB algorithm with a minor change to the initialization rounds: In the first $|U|$ rounds of initialization, for each cluster C, we play a super arm S such that $C\in g(S)$ and update variables $\hat{\mu}_i$ accordingly.

For a given cluster $C\subseteq[m]$, we sort all *bad* super arms whose cluster set contains C as $S_{C,B}^1, S_{C,B}^2,\cdots,S_{C,B}^{K_C}$ by increasing reward. Define

$$\Delta^{C,j}=\alpha\cdot\mathrm{opt}_{\boldsymbol{\mu}}-r_{\boldsymbol{\mu}}(S_{C,\mathrm{B}}^j),\tag{30}$$

$\Delta_{\max}^C=\Delta^{C,1}$ and $\Delta_{\min}^C=\Delta^{C,K_C}$. If C does not belong to any *bad* super arm, $K_C=0$ and set $\Delta_{\max}^C=\Delta_{\min}^C=0$. Furthermore, define $\Delta_{\max}=\max_{C\in U}\Delta_{\max}^C$.

Theorem 3. *Consider the CMAB problem with the set of clusters U of arms. In n rounds the (α,β)-approximation regret of the CUCB algorithm using an (α,β)-approximation oracle is at most*

$$\sum_{C\mid\Delta_{\min}^C>0}\left(\frac{6\ln n}{(f^{-1}(\Delta_{\min}^C))^2}\cdot\Delta_{\min}^C+\int_{\Delta_{\min}^C}^{\Delta_{\max}^C}\frac{6\ln n}{(f^{-1}(x))^2}\mathrm{d}x\right)+\left(\frac{\pi^2}{3}+1\right)\cdot m\cdot\Delta_{\max}.$$

Discussion. Comparing with the regret bound in Theorem 1, we are taking the summation over all clusters instead of all underlying arms. Since we assume $|U|<m$, intuitively, we could be better off. However, it is not clear how the $\Delta_{\min}$'s of the underlying arms and clusters are correlated with each other. When clusters do not

intersect with one another and thus form a partition of the underlying arms, it is straightforward to show that $\Delta^i_{\min} = \Delta^C_{\min}$ if the cluster C contains the arm i. In this case, the new regret bound of Theorem 3 is a strict improvement over Theorem 1. The two applications discussed in this paper, i.e., the bandit PMC problem and the bandit influence maximization problem, belong to this category and thus Theorem 3 could be applied and obtain improved regret bounds.

Proof. The proof of this theorem is almost identical to Theorem 1. In addition to T_i, our analysis requires T_C which is the number of time cluster C is selected to play. Let $T_{C,n}$ be the value of T_C at the end of round n, that is, $T_{C,n}$ is the number of times cluster C is played in the first n rounds. Let $T_{i,n}$ still be the value of T_i at the end of round n, that is, $T_{i,n}$ is the number of times arm i is played in the first n rounds. Since arm i might be contained in multiple clusters, here $T_{i,n}$ is larger than $T_{C,n}$ for any C containing i.

For the proof, we maintain counter N_C for each cluster C after the U initialization rounds. Let $N_{C,t}$ be the value of N_C after the t-th round and $N_{C,|U|} = 1$. Note that $\sum_C N_{C,|U|} = |U|$. $\{N_C\}$ is updated in the following way.

For a round $t > |U|$, let S_t be the super arm selected in round t. Round t is bad if the oracle selects a super arm $S_t \in \mathcal{S}_{\mathrm{B}}$, which is not an α-approximate super arm. If round t is bad, let $C = \mathrm{argmin}_{C\in g(S_t)} N_{C,t-1}$ and increment N_C by one, i.e., $N_{C,t} = N_{C,t-1} + 1$. In other words, we find the cluster C with smallest counter in $g(S_t)$ and increase its counter. If C is not unique, we pick an arbitrary cluster with the smallest counter in $g(S_t)$.

By definition $N_{C,t} \leq T_{C,t}$. The total number of bad rounds in the first n rounds is $\sum_C N_{C,n}$.

Each time N_C gets updated, one of the bad arm whose cluster set contains C is played. We further divide N_C into more counters as follows:

$$\forall l \in [K_C],\ N^l_{C,n} = \sum_{t=|U|+1}^{n} \mathbb{I}\{S_t = \mathcal{S}^l_{C,\mathrm{B}}, N_{C,t} > N_{C,t-1}\}.$$

Define $\ell_n(\Delta) = \frac{6\ln n}{(f^{-1}(\Delta))^2}$. When counter $N^l_{i,t}$ is increased at time t, i.e., $S_t = S^l_{C,B}$, we inspect the counter $N^l_{C,t-1}$. Notice that $N^l_{C,t-1}$ is the smallest time that all arms in S_t have been played. If $N_{c,t-1} > \ell_n(\Delta^{C,l})$, we call the bad arm $S^l_{C,B}$ is sufficiently sampled. Otherwise, it is under-sampled. We write as

$$\begin{aligned}
N^{l,suf}_{C,n} &= \sum_{t=m+1}^{n} \mathbb{I}\{S_t = \mathcal{S}^l_{C,\mathrm{B}}, N_{C,t} > N_{C,t-1}, N_{C,t-1} > \ell_n(\Delta^{C,l})\},\\
N^{l,und}_{C,n} &= \sum_{t=m+1}^{n} \mathbb{I}\{S_t = \mathcal{S}^l_{C,\mathrm{B}}, N_{C,t} > N_{C,t-1}, N_{C,t-1} \leq \ell_n(\Delta^{C,l})\}.
\end{aligned}$$

Then we have $N_{C,n} = 1 + \sum_{l\in[K_C]}(N^{l,suf}_{C,n} + N^{l,und}_{C,n})$. Using this notation, the total reward at time horizon n is at least

$$n \cdot \alpha \cdot \mathrm{opt}_{\boldsymbol{\mu}} - \mathbb{E}\left[\sum_{C\in U}\left(\Delta^{C,1} + \sum_{l\in[K_C]}(N^{l,suf}_{C,n} + N^{l,und}_{C,n}) \cdot \Delta^{C,l}\right)\right]. \tag{31}$$

Note that the total sampled time of underlying arms in one cluster will not be smaller than the total sampled time of that cluster. We claim that it is unlikely that a bad super arm is played when all the underlying arms are sufficiently sampled. In other words, for a bad super arm, if all its underlying arms are sufficiently sampled, it should not be played in the first place. More specifically, we have the following claim.

Claim 2. *For any time horizon $n > m$,*

$$\mathbb{E}\left[\sum_{C\in U}\sum_{l\in[K_C]} N^{l,suf}_{C,n}\right] \leq (1-\beta)n + \frac{\pi^2}{3}\cdot m \tag{32}$$

Proof. By the definition of $N^{l,suf}_{C,n}$, it is sufficient to show that for any $t > m$,

$$\begin{aligned}
&\mathbb{E}\left[\sum_{C\in[m],l\in[K_C]} \mathbb{I}\{S_t = \mathcal{S}^l_{C,\mathrm{B}}, N_{C,t} > N_{C,t-1}, N_{C,t-1} > \ell_n(\Delta^{C,l})\}\right] \\
\leq &\sum_{C\in[m],l\in[K_C]} \Pr\{S_t = \mathcal{S}^l_{C,\mathrm{B}}, N_{C,t} > N_{C,t-1}, \forall s \in \mathcal{S}^l_{C,\mathrm{B}}, T_{s,t-1} > \ell_n(\Delta^{C,l})\} \\
\leq &(1-\beta) + 2mt^{-2}
\end{aligned}$$

Define $\Lambda_{i,t} = \sqrt{\frac{3\ln t}{2T_{i,t-1}}}$ (a random variable since $T_{i,t-1}$ is a random variable) and $\Lambda_t = \max\{\Lambda_{i,t} \mid i \in S_t\}$. Define $\Lambda^{C,l} = \sqrt{\frac{3\ln t}{2\ell_n(\Delta^{C,l})}}$.

Let $\mathcal{N}_t$ indicate the event that the process is *nice* at time t. Let F_t indicate the event that the oracle fails to return an α-approximation with respect to the input vector at time t. For any particular $C \in U$ and $l \in [K_C]$, if $\left\{\mathcal{N}_t, \neg F_t, S_t = \mathcal{S}^l_{C,\mathrm{B}}, \forall s \in S_t, T_{s,t-1} > \ell_n(\Delta^{C,l})\right\}$ holds at time t, we have the following properties:

$$\begin{aligned}
r_{\boldsymbol{\mu}}(S_t) + f(2\Lambda^{C,l}) >& r_{\boldsymbol{\mu}}(S_t) + f(2\Lambda_t) && \text{strict monotonicity of } f(\cdot) \text{ and Eq.(8)} \\
\geq& r_{\bar{\boldsymbol{\mu}}_t}(S_t) && \text{bounded smoothness property and Eq.(7)} \\
\geq& \alpha \cdot \mathrm{opt}_{\bar{\boldsymbol{\mu}}_t} && \neg F_t \Rightarrow S_t \text{ is an } \alpha \text{ approximation w.r.t } \bar{\boldsymbol{\mu}}_t \\
\geq& \alpha \cdot r_{\bar{\boldsymbol{\mu}}_t}(S^*_{\boldsymbol{\mu}}) && \text{definition of } \mathrm{opt}_{\bar{\boldsymbol{\mu}}_t} \\
\geq& \alpha \cdot r_{\boldsymbol{\mu}}(S^*_{\boldsymbol{\mu}}) = \alpha \cdot \mathrm{opt}_{\boldsymbol{\mu}}. && \text{monotonicity of } r_{\boldsymbol{\mu}}(S) \text{ and Eq.(9)}
\end{aligned}$$

So we have

$$r_{\boldsymbol{\mu}}(S^l_{C,\mathrm{B}}) + f(2\Lambda^{C,l}) > \alpha \cdot \mathrm{opt}_{\boldsymbol{\mu}}. \tag{33}$$

Since $\ell_n(\Delta^{C,l}) = \frac{6\ln n}{(f^{-1}(\Delta^{C,l}))^2}$, we have $f(2\Lambda^{C,l}) \leq \Delta^{C,l}$. Therefore, Eq. (33) contradicts the definition of $\Delta^{C,l}$. In other words,

$$\begin{aligned}
&\forall C \in [m]\, \forall l \in [K_C],\ \Pr\left\{\mathcal{N}_t, \neg F_t, S_t = S^l_{C,\mathrm{B}}, N_{C,t} > N_{C,t-1}, \forall s \in S_t, N_{s,t-1} > \ell_n(\Delta^{C,l})\right\} = 0 \\
\Rightarrow &\Pr\left\{\mathcal{N}_t, \neg F_t, \exists C \in U, \exists l \in [K_C], S_t = S^l_{C,\mathrm{B}}, N_{C,t} > N_{C,t-1}, \forall s \in S_t, N_{s,t-1} > \ell_n(\Delta^{C,l})\right\} = 0 \\
\Rightarrow &\Pr\left\{\exists C \in U, \exists l \in [K_C], S_t = S^l_{C,\mathrm{B}}, N_{C,t} > N_{C,t-1}, \forall s \in S_t, N_{s,t-1} > \ell_n(\Delta^{C,l})\right\} \\
&\quad \leq \Pr[F_t \vee \neg\mathcal{N}_t] \leq (1-\beta) + 2mt^{-2} \\
\Rightarrow &\sum_{C\in U,l\in[K_C]} \Pr\left\{S_t = S^l_{C,\mathrm{B}}, N_{C,t} > N_{C,t-1}, \forall s \in S_t, N_{s,t-1} > \ell_n(\Delta^{C,l})\right\} \leq (1-\beta) + 2mt^{-2}
\end{aligned}$$

The first inequality comes from Lemma 3. Last inequality comes from the fact that events $\left\{S_t = S^l_{C,\mathrm{B}}, N_{C,t} > N_{C,t-1}, \forall s \in S_t, N_{s,t-1} > \ell_n(\Delta^{C,l})\right\}$ are disjoint since for the cluster C whose counter is updated at time t only one l satisfies $S_t = S^l_{C,\mathrm{B}}$. □

Now we consider the bad super arms that are under-sampled when played. To simplify the notation, define $\ell_n(\Delta^{C,0}) = 0$. For a cluster C,

$$
\begin{aligned}
&\sum_{l\in[K_C]} N_{C,n}^{l,und}\cdot\Delta^{C,l}\\
=&\sum_{t=|U|+1}^{n}\sum_{l\in[K_C]}\mathbb{I}\{S_t=\mathcal{S}_{C,\mathrm{B}}^{l},N_{C,t}>N_{C,t-1},N_{C,t-1}\le\ell_n(\Delta^{C,l})\}\cdot\Delta^{C,l}\\
=&\sum_{t=|U|+1}^{n}\sum_{l\in[K_C]}\sum_{j=1}^{l}\mathbb{I}\{S_t=\mathcal{S}_{C,\mathrm{B}}^{l},N_{C,t}>N_{C,t-1},N_{C,t-1}\in(\ell_n(\Delta^{C,j-1}),\ell_n(\Delta^{C,j})]\}\cdot\Delta^{C,l}\\
\le&\sum_{t=|U|+1}^{n}\sum_{l\in[K_C]}\sum_{j=1}^{l}\mathbb{I}\{S_t=\mathcal{S}_{C,\mathrm{B}}^{l},N_{C,t}>N_{C,t-1},N_{C,t-1}\in(\ell_n(\Delta^{C,j-1}),\ell_n(\Delta^{C,j})]\}\cdot\Delta^{C,j}\\
\le&\sum_{t=|U|+1}^{n}\sum_{l\in[K_C]}\sum_{j\in[\mathbf{K_C}]}\mathbb{I}\{S_t=\mathcal{S}_{C,\mathrm{B}}^{l},N_{C,t}>N_{C,t-1},N_{C,t-1}\in(\ell_n(\Delta^{C,j-1}),\ell_n(\Delta^{C,j})]\}\cdot\Delta^{C,j}\\
=&\sum_{t=|U|+1}^{n}\sum_{j\in[K_C]}\mathbb{I}\{S_t\in\mathcal{S}_{C,\mathrm{B}},N_{C,t}>N_{C,t-1},N_{C,t-1}\in(\ell_n(\Delta^{C,j-1}),\ell_n(\Delta^{C,j})]\}\cdot\Delta^{C,j}\\
=&\sum_{j\in[K_C]}\sum_{t=|U|+1}^{n}\mathbb{I}\{S_t\in\mathcal{S}_{C,\mathrm{B}},N_{C,t}>N_{C,t-1},N_{C,t-1}\in(\ell_n(\Delta^{C,j-1}),\ell_n(\Delta^{C,j})]\}\cdot\Delta^{C,j}\\
\le&\sum_{j\in[K_C]}(\ell_n(\Delta^{C,j})-\ell_n(\Delta^{C,j-1}))\cdot\Delta^{C,j}\\
=&\ell_n(\Delta^{C,K_i})\Delta^{C,K_i}+\sum_{j\in[K_C]}\ell_n(\Delta^{C,j})\cdot(\Delta^{C,j}-\Delta^{C,j+1})\\
\le&\ell_n(\Delta^{C,K_C})\Delta^{C,K_C}+\int_{\Delta^{C,K_C}}^{\Delta^{C,1}}\ell_n(x)\mathrm{d}x.
\end{aligned}
\tag{34}
$$

Last inequality comes from the fact that $\ell_n(x)$ is decreasing. Notice that by our definition, for clusters that $C_j\in[m]\setminus\{C\mid\Delta_{\min}^{C}>0\}$, the counter N_C remains one after the initialization. Since they do not contribute to any regret, we have $K_C=0$ for all these arms. Combining with Eq.(32) and Eq.(34), the overall regret of our algorithm is

$$
\begin{aligned}
&Reg_{\boldsymbol{\mu},\alpha,\beta}^{A}(n)\\
\le&n\cdot\alpha\cdot\beta\cdot\mathrm{opt}_{\boldsymbol{\mu}}-\left(\alpha\cdot n\cdot\mathrm{opt}_{\boldsymbol{\mu}}-\mathbb{E}\left(\sum_{C\in U}\left(\Delta^{C,1}+\sum_{l\in[K_C]}(N_{C,n}^{l,suf}+N_{C,n}^{l,und})\cdot\Delta^{C,l}\right)\right)\right)\Bigg]\\
\le&\Delta_{\max}\cdot\left(m+\mathbb{E}[\sum_{C\in U,l\in[K_C]}N_{C,n}^{l,suf}]\right)+\sum_{C\mid\Delta_{\min}^{C}>0}\left(\Delta^{C,1}+\ell_n(\Delta^{C,K_C})\Delta^{C,K_C}+\int_{\Delta^{C,K_C}}^{\Delta^{C,1}}\ell_n(x)\mathrm{d}x\right)-(1-\beta)\cdot n\cdot\alpha\cdot\mathrm{opt}_{\boldsymbol{\mu}}\\
\le&\left(\frac{\pi^2}{3}+1\right)\cdot m\cdot\Delta_{\max}+\sum_{C\mid\Delta_{\min}^{C}>0}\left(\ell_n(\Delta^{C,K_C})\Delta^{C,K_C}+\int_{\Delta^{C,K_C}}^{\Delta^{C,1}}\ell_n(x)\mathrm{d}x\right).
\end{aligned}
$$

The theorem follows directly. □

C. ε_t-Greedy algorithm

Unlike CUCB algorithm, ε_t-greedy algorithm exhibits the combination of exploration and exploitation more explicitly. In the t-th round, with probability ε_t the algorithm performs *exploration*, i.e., chooses an arm i uniformly at random, then select an arbitrary super arm $S\in\mathcal{S}$ containing i; with probability $1-\varepsilon_t$, the

```
For each arm $i$, maintain variable $\hat{\mu}_i$ as the mean of all outcomes $X_{i,*}$'s of arm $i$ observed so far.
$n \leftarrow 0$
while true do
    $t \leftarrow t+1$; $\varepsilon_t \leftarrow \min\{\frac{\gamma}{t}, 1\}$.
    With probability $\varepsilon_t$, choose an arm $i$ uniform at random, then play an arbitrary super arm $S \in \mathcal{S}$ containing
    $i$; with probability $1-\varepsilon_t$, get $S = \text{Oracle}(\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_m)$.
    Play $S$ and update all $\hat{\mu}_i$'s.
end while
```

Algorithm 1: ε_t-greedy algorithm with computation oracle

algorithm performs *exploitation*, i.e., uses the approximation oracle to choose a super arm. As t grows, the probability of performing exploration decreases so that the regret can be bounded. See Algorithm 1 for details. Note that if an arm i has never been played, $\hat{\mu}_i$ could be any arbitrary value.

The appeal of the ϵ_t-greedy algorithm is its simplicity and match with intuition. However, as shown in the following theorem, in order to have a theoretical guarantee on the regret bound, parameter γ needs to be set appropriately and it depends on $\Delta_{\min}$ and function $f(\cdot)$. In constrast, the CUCB algorithm does not rely on the knowledge of $\Delta_{\min}$ and $f(\cdot)$, and thus CUCB is applicable to more settings in this sense.

Theorem 4. *For any constant $c > 1$, define $\gamma = 3m \max\left\{\frac{20c}{3}, (c+1)\cdot(f^{-1}(\frac{\Delta_{\min}}{2}))^{-2}\right\}$. In n rounds the (α,β)-approximation regret of the ε_t-greedy algorithm using an (α,β)-approximation oracle is at most*

$$\left(\gamma \ln n + 3\cdot\zeta(c)\cdot m + \gamma^3\right)\Delta_{\max},$$

where $\zeta(c) = \sum_{t=1}^{\infty}\frac{1}{t^c}$ is the Riemann's zeta function.

Recall that the definition of $\Delta_{\min}$ is

$$\Delta_{\min} = \alpha \cdot \text{opt}_{\boldsymbol{\mu}} - \max\{r_{\boldsymbol{\mu}}(S) \mid S \in \mathcal{S}_{\text{B}}\}. \tag{35}$$

Proof. Let $R_{i,t}$ be the indicator for the event that i was chosen to *explore* in the t-th round and $N_{i,t}$ be the number of rounds that arm i is explored in the first t rounds. Set $\varphi = \frac{\gamma}{3m}$. For simplicity, we assume γ is integer. We have:

$$\mathbb{E}[N_{i,n}] = \sum_{t=1}^{n}\mathbb{E}[R_{i,t}] = \sum_{t=1}^{n}\frac{\varepsilon_t}{m} = \frac{\gamma-1}{m} + \sum_{t=\gamma}^{n}\frac{3\varphi}{t} > 3\varphi + \int_{\gamma}^{n}\frac{3\varphi}{x}\,\mathrm{d}x = 1 + 3\varphi\ln(n/\gamma) \tag{36}$$

When $n > \gamma^3$, (36) is at least $2\varphi\ln n + 1$. Now let $X_{i,t} = R_{i,t} - \mathbb{E}[R_{i,t}]$. We have $\mathbb{E}[X_{i,t}] = 0$, $|X_{i,t}| \le 1$, and

$$\sum_{t=1}^{n}\mathbb{E}\left[X_{i,t}^2\right] = \sum_{t=\gamma}^{n}\left(1 - \frac{3\varphi}{t}\right)\frac{3\varphi}{t} < 3\varphi\ln n.$$

By Bernstein inequality in Lemma 2, when $n > \gamma^3$, we have

$$\begin{aligned}
\Pr\left[\left|\sum_{t=1}^{n}X_{i,t}\right| > \varphi\ln n\right] &\le \exp\left\{-\frac{\varphi^2(\ln n)^2/2}{\sum_{t=1}^{n}\mathbb{E}[X_{i,t}^2] + k(\varphi\ln n)/3}\right\} \\
&\le \exp\left\{-\frac{\varphi^2(\ln n)^2/2}{3\varphi\ln n + \varphi\ln n/3}\right\} \\
&= e^{-\frac{3}{20}\varphi\ln n} = n^{-\frac{3}{20}\varphi} \le n^{-c}.
\end{aligned}$$

In other words, $\Pr[N_{i,t} \le \varphi\ln t + 1] \le t^{-c}$. By union bound, $\Pr[\exists i \in [m], N_{i,t} \le \varphi\ln t + 1] \le mt^{-c}$. Let P_t to be the indicator of the event that in the t-th round, all the arms have been played for at least $\varphi\ln t + 1$ times. So, $\Pr[P_t = 0] \le mt^{-c}$. Set $\ell_t = \frac{(c+1)\ln t}{\left(f^{-1}\left(\frac{\Delta_{\min}}{2}\right)\right)^2} \le \varphi\ln t$. Note $P_t = 1$ indicates for every arm i, $T_{i,t} \ge N_{i,t} \ge \varphi\ln t + 1 > \ell_t$.

Let I_t be the event that we choose a bad arm $S_t \in \mathcal{S}_{\mathrm{B}}$ in the t-th round. Let Y_t be the event that the action taken in the t-th round is exploitation (not exploration). Let F_t be the event that the oracle failed to produce an α-approximate answer in an exploitation round t. We have $\mathbb{E}[F_t \mid Y_t] \leq 1-\beta$.

We have,

$$\begin{aligned}
\sum_{t=\gamma^3+1}^{n} \mathbb{I}\{I_t\} &= \sum_{t=\gamma^3+1}^{n} (\mathbb{I}\{I_t, \neg Y_t\} + \mathbb{I}\{I_t, Y_t\}) \\
&= \sum_{t=\gamma^3+1}^{n} \varepsilon_t \cdot \mathbb{I}\{I_t \mid \neg Y_t\} + \sum_{t=\gamma^3+1}^{n} \mathbb{I}\{I_t, Y_t\} \\
&\leq \gamma \ln n + \sum_{t=\gamma^3+1}^{n} \mathbb{I}\{I_t, Y_t\}
\end{aligned}$$

Consider the second term.

$$\begin{aligned}
&\sum_{t=\gamma^3+1}^{n} \mathbb{I}\{I_t, Y_t\} \\
\leq &\sum_{t=\gamma^3+1}^{n} (\mathbb{I}\{F_t, Y_t\} + \mathbb{I}\{\neg F_t, I_t, Y_t\}) \\
\leq &(1-\beta)(n-\gamma^3) + \sum_{t=\gamma^3+1}^{n} (\mathbb{I}\{\neg F_t, I_t, \neg P_t, Y_t\} + \mathbb{I}\{\neg F_t, I_t, P_t, Y_t\}) \\
\leq &(1-\beta)n + \sum_{t=\gamma^3+1}^{n} \left(mt^{-c} \cdot \mathbb{I}\{I_t \mid \neg F_t, \neg P_t, Y_t\} + \mathbb{I}\{\neg F_t, I_t, P_t, Y_t\}\right) \\
\leq &(1-\beta)n + \zeta(c) \cdot m + \sum_{t=\gamma^3+1}^{n} \mathbb{I}\{\neg F_t, I_t, P_t, Y_t\} \\
= &(1-\beta)n + \zeta(c) \cdot m + \sum_{t=\gamma^3+1}^{n} \mathbb{I}\{\neg F_t, Y_t, S_t \in \mathcal{S}_{\mathrm{B}}, \forall i \in [m],\, T_{i,t-1} > \ell_t\}
\end{aligned}$$

We claim that

$$\Pr[\{\neg F_t, Y_t, S_t \in \mathcal{S}_{\mathrm{B}}, \forall i \in [m],\, T_{i,t-1} > \ell_t\}] \leq 2 \cdot m \cdot t^{-c}.$$

We now prove this claim. Same as in the proof of Theorem 1, let $T_{i,n}$ be the number of times arm i is played in the first n rounds; let $\hat{\mu}_{i,s}$ be the value of $\hat{\mu}_i$ after arm i is played s times, that is, $\hat{\mu}_{i,s} = (\sum_{j=1}^{s} X_{i,j})/s$. Then, the value of variable $\hat{\mu}_i$ at the end of round n is $\hat{\mu}_{i,T_{i,n}}$. By Chernoff bound in Lemma 1, for any $i \in [m]$,

$$\Pr\left[|\hat{\mu}_{i,T_{i,t-1}} - \mu_i| \geq \sqrt{\frac{2\ln t + \ln\ln t}{2T_{i,t-1}}}\right] \leq t \cdot 2e^{-2\ln t - \ln\ln t} \leq 2(t\ln t)^{-1}. \tag{37}$$

Define $\Delta_{i,t} = \sqrt{\frac{2\ln t + \ln\ln t}{2T_{i,t-1}}}$. Define $E_t = \{\forall i \in [m],\, |\hat{\mu}_{i,T_{i,t-1}} - \mu_i| \leq \Delta_{i,t}\}$. By union bound, $\Pr[\neg E_t] \leq 2 \cdot m \cdot t^{-c}$. Let $\Delta = \sqrt{\frac{2\ln t + \ln\ln t}{2\ell_t}}$. Notice that when $\forall i \in [m], T_{i,t-1} > \ell_t$, we have $\Delta > \Delta^t \overset{\text{def}}{=} \max\{\Delta_{i,t} \mid i \in [m]\}$.

Let $\hat{\boldsymbol{\mu}}_t = (\hat{\mu}_{1,T_{1,t-1}}, \ldots, \hat{\mu}_{m,T_{m,t-1}})$ be the random vector representing the estimated expectation vector at round t before calling the oracle. Then, when $\{E_t, \neg F_t, Y_t, S_t \in \mathcal{S}_{\mathrm{B}}, \forall i \in [m], T_{i,t-1} > \ell_t\}$ holds, we have the following properties:

$$
\begin{aligned}
r_{\boldsymbol{\mu}}(S_t) + 2f(\Delta) &> r_{\boldsymbol{\mu}}(S_t) + (1+\alpha)f(\Delta^t) && \text{monotonicity of } f(\cdot)\\
&\geq r_{\hat{\boldsymbol{\mu}}_t}(S_t) + \alpha f(\Delta^t) && \text{bounded smoothness property with } E_t\\
&\geq \alpha \cdot \mathrm{opt}_{\hat{\boldsymbol{\mu}}_t} + \alpha f(\Delta^t) && \neg F_t \Rightarrow S_t \text{ is an } \alpha \text{ approximation w.r.t } \hat{\boldsymbol{\mu}}_t\\
&\geq \big(r_{\hat{\boldsymbol{\mu}}_t}(S^*_{\boldsymbol{\mu}}) + f(\Delta^t)\big) \cdot \alpha && \text{definition of } \mathrm{opt}_{\hat{\boldsymbol{\mu}}_t}\\
&\geq \alpha \cdot r_{\boldsymbol{\mu}}(S^*_{\boldsymbol{\mu}}) = \alpha \cdot \mathrm{opt}_{\boldsymbol{\mu}}. && \text{bounded smoothness property with } E_t
\end{aligned}
$$

These above inequalities imply that when E_t holds, we have

$$r_{\boldsymbol{\mu}}(S^j) + 2f(\Delta) > \alpha \cdot \mathrm{opt}_{\boldsymbol{\mu}}. \tag{38}$$

Since $\ell_t = \frac{2\ln t + \ln\ln t}{\left(f^{-1}\left(\frac{\Delta_{\min}}{2}\right)\right)^2}$, we have

$$2f(\Delta) = 2f\left(f^{-1}\left(\frac{\Delta_{\min}}{2}\right)\right) = \Delta_{\min}.$$

With $2f(\Delta) = \Delta_{\min}$, Eq. (38) is in conflict with the definition of $\Delta_{\min}$ in Eq. (35). In other words,

$$
\begin{aligned}
&\Pr\left[\{E_t, \neg F_t, Y_t, S_t \in \mathcal{S}_{\mathrm{B}}, \forall i \in [m], T_{i,t-1} > \ell_t\}\right] = 0 \Rightarrow\\
&\Pr\left[\{\neg F_t, Y_t, S_t = S^j, \forall i \in S^j \cup S^*_{\boldsymbol{\mu}}, T_{i,t-1} = s_i\}\right] \leq \Pr[\neg E_t] \leq 2 \cdot m \cdot (t \ln t)^{-1}.
\end{aligned}
$$

Thus,

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{n} \mathbb{I}\{I_t\}\right]\\
\leq&\gamma^3 + \gamma \ln n + (1-\beta)n + \zeta(c) \cdot m + \sum_{t=\gamma^3+1}^{n} 2 \cdot m(t \ln t)^{-1}\\
\leq&\gamma^3 + \gamma \ln n + (1-\beta)n + \zeta(c) \cdot m + 2m \ln\ln n
\end{aligned}
$$

That means, the regret is at most:

$$
\begin{aligned}
Reg^{A}_{\boldsymbol{\mu},\alpha,\beta}(n) &\leq n \cdot \alpha \cdot \beta \cdot \mathrm{opt}_{\boldsymbol{\mu}} - \left(n \cdot \alpha \cdot \mathrm{opt}_{\boldsymbol{\mu}} - \Delta_{\max} \cdot \mathbb{E}\left[\sum_{t=1}^{n} \mathbb{I}\{I_t\}\right]\right)\\
&\leq \big(\gamma^3 + \gamma \ln n + (1-\beta)n + 3 \cdot \zeta(c) \cdot m\big)\,\Delta_{\max} - (1-\beta) \cdot n \cdot \alpha \cdot \mathrm{opt}_{\boldsymbol{\mu}}\\
&\leq \big(\gamma \ln n + 3 \cdot \zeta(c) \cdot m + \gamma^3\big)\,\Delta_{\max}.
\end{aligned}
$$

□

References

Garivier, A. and Cappé, O. The KL-UCB algorithm for bounded stochastic bandits and beyond. In *COLT*, 2011.