## Appendix A. Additional Experiments Results

### A.1. Additional Explanation on the experiment settings

$\mu_k$ is sampled from an even distribution on the interval $[1, 11]$. The utility perturbation $\epsilon$ is set follow the same Gaussian distribution $\mathcal{N}\left(0, \sigma^2\right)$ for all arms across all settings, with $\sigma = 0.5$. We modified the returned utility function in both the Orch and MP-MA-SE algorithms to align with our problem setting and compare their performance with ours. Simulations were conducted on both versions of our algorithm, differing only in the estimator used for the capacity confidence interval. For each setting, we run 20 simulations and the resulting regrets are averaged.

### A.2. Impact of Total Capacity

In figure 2, we set the interval that $m_k$ is evenly sampled from $[10, 15], [10, 20], [10, 25], [10, 30]$ respectively. We observe that as the capacities of the arms increase, the regret is larger at the same time slot. This observation does not contradict the regret bounds presented in Theorem 10 and Theorem 14 in our setting. The primary reason is that IEs with only one play generate higher regret as the actual capacities increase, and such IEs are unavoidable in all four algorithms when the capacity confidence intervals are not well-learned. However, this impact is only evident during the early time slots and does not incur a long-term regret increase that scales with the number of time slots. In all settings, our algorithms significantly outperform Orch and MP-SE-SA. Additionally, the improvement from using the new estimator is substantial, leading to much faster convergence of capacity confidence intervals.

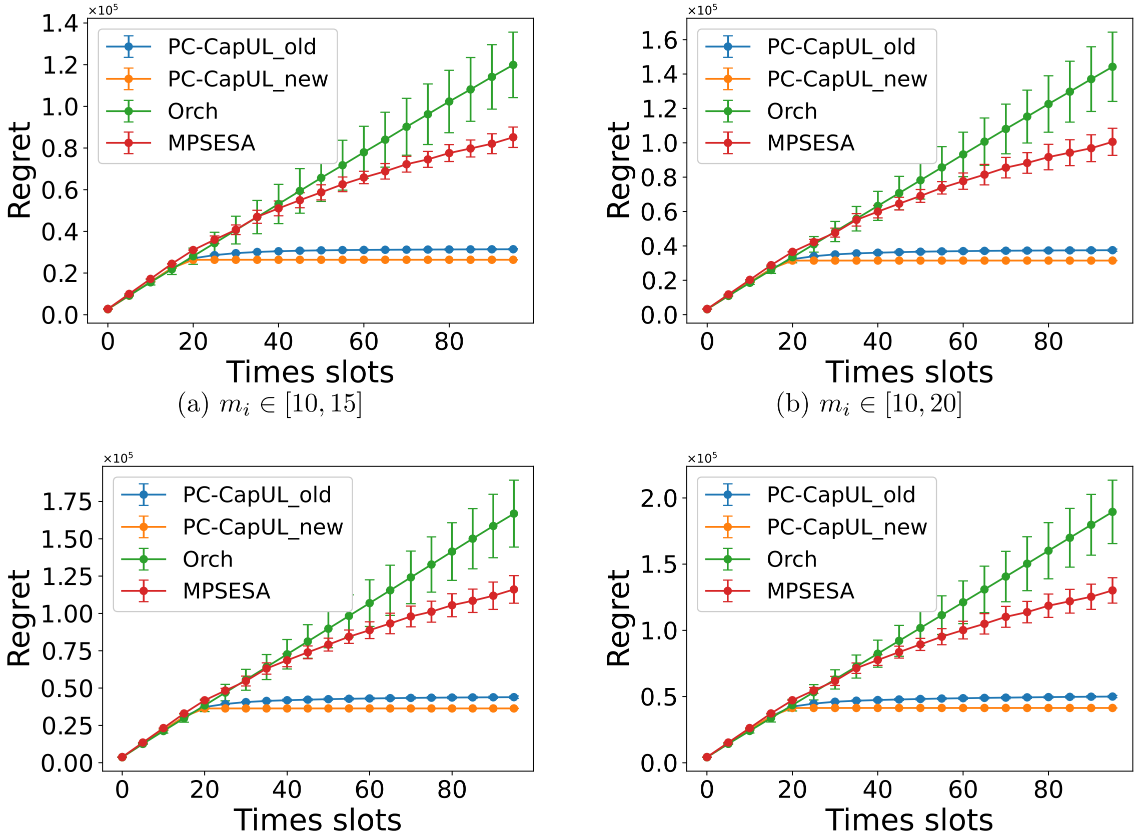
### A.3. Impact of Number of Plays

In figure 3, we fix $M$ as $\sum_{k=1}^{K} m_k$ and set the ratio $N/M$ as $1, 1.1, 1.2, 1.4$ respectively. We find that as $N$ varies, our algorithms outperform the Orch and the MP-SE-SA in all four settings. The main reason is that a greater number of plays allows our algorithms to perform more UEs simultaneously, thereby reducing the number of time slots required for the capacity confidence intervals to converge. However, the increase in the number of plays has little impact on the performance of Orch, as its UEs are restricted by a conservative strategy designed for scenarios where $N < M$.

Additional experiments are also conducted to illustrate the impact of extremely large $N$, which is set to be 2 to 64 times greater than $M$. Specifically, we fix $m_k$ such that $\sum_k m_k = 263$, and vary $N$ from 600 to 19200. The numerical results are presented in the table below for a detailed comparison.

Table 2: Performance comparison over different $N$ and $T$ values

| **N** (×600) | **1** | **2** | **4** | **8** | **16** | **32** |
|---|---|---|---|---|---|---|
| T=300 | 33424.66 | 33589.91 | 33798.40 | 34341.67 | 35239.21 | 37269.60 |
| T=350 | 33435.15 | 33593.29 | 33809.00 | 34354.02 | 35244.12 | 37274.29 |
| T=400 | 33442.39 | 33593.52 | 33812.25 | 34357.53 | 35244.68 | 37275.12 |



(a) $m_i \in [10, 15]$

(b) $m_i \in [10, 20]$

(c) $m_i \in [10, 25]$

(d) $m_i \in [10, 30]$

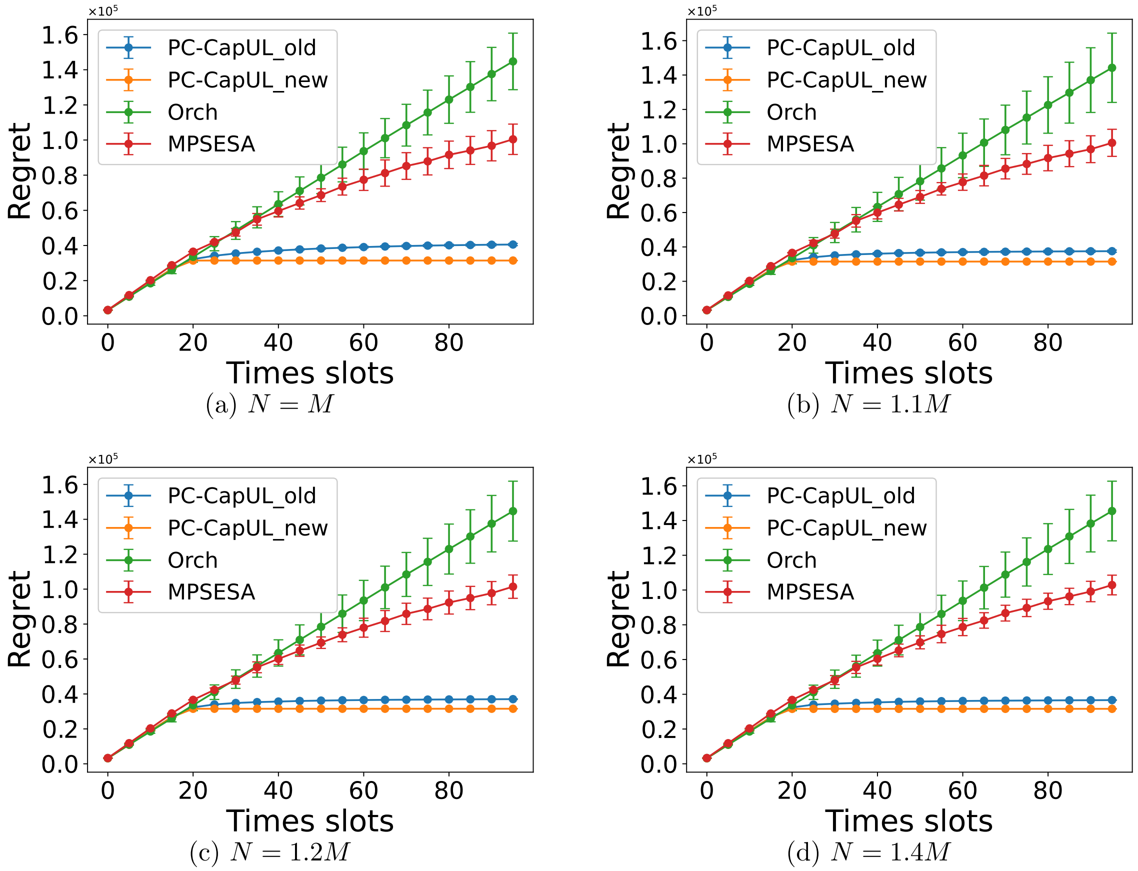
Figure 2: Impact of capacities of Arms.

For a fixed $N$, when $T$ increases from 300 to 400, the regret grows only slightly, indicating that the arm capacities estimation is nearly finished.

For a fixed $T$, when $N$ increases 31-fold from 600 to 19200, the regret increases by only about 0.12 times, much slower than predicted by our theoretical upper bound. This suggests that PC-CapUL's dependence on $N$ in the regret bound is sublinear rather than linear, highlighting its strong performance even for very large $N$.

### A.4. Impact of movement cost

In figure 4, we set the movement cost $c = 0.2, 0.1, 0.01$ respectively. We find that as $c$ decreases, the regrets of all four algorithms decrease. It is reasonable that with a smaller $c$, the cost of UEs decreases across all four algorithms, resulting in lower regret if other parameters remain unchanged. However, this change in movement cost has little impact on the comparative performance of the four algorithms. The primary reason is that changing $c$ mainly influences the regret generated by UEs, and UEs are relatively rare compared to IEs in all four algorithms. When $N$ is not much larger than $M$, the regret generated by UEs is typically smaller than that of IEs.



(a) $N = M$

(b) $N = 1.1M$

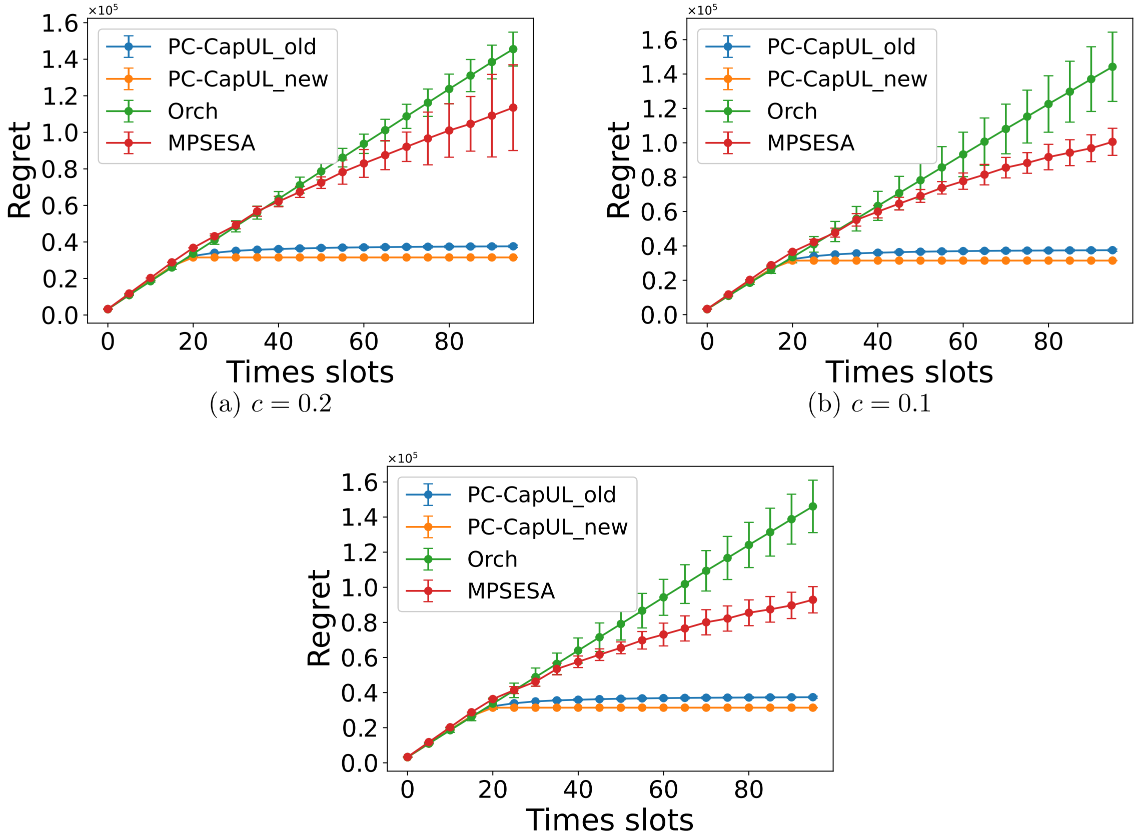
(c) $N = 1.2M$

(d) $N = 1.4M$

Figure 3: Impact of number of plays

### A.5. Compare of the old and new estimators

In figure 5, we set $K = 1$, $M = m_1 = 15$, $N = 30$ , and do UEs and IEs in an alternating way to explore the capacity. We first set the estimators of LCB and UCB of the capacity as formula (5) and (6) in the main paper, and record their values as new-LCB and new-UCB, as shown in the figure 5. Next, we set the estimators as those used in Wang et al. (2022a), denoting them as old-LCB and old-UCB. For both estimator settings, we run simulations 20 times and average the recorded LCB and UCB values. As shown in figure 5, it is clear that the new estimator converges far more rapidly than the old one, even though both estimators eventually converge to the correct capacity after sufficient explorations.



(a) $c = 0.2$

(b) $c = 0.1$

(c) $c = 0.01$

Figure 4: Impact of movement cost

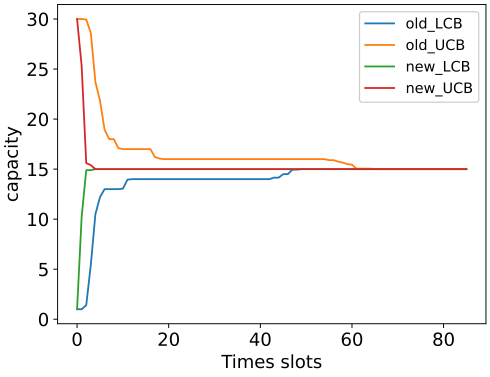


Figure 5: Compare of the old and new estimators

# Appendix B. Technical Proofs

## B.1. Sample Complexity Proof

**Proof** (Theorem 2 )

Consider there is an arm with capacity $m_k$ and unit utility value $\mu$. Assume that there are only two possible values for $m_k$: $\{m, m+1\}$ where $m$ is a positive integer, and the perturbation on the arm follows $\mathcal{N}\left(0, \sigma^2\right)$. Let $T$ be number of the explorations performed on this arm.

For any strategy $\pi$ that calculates the capacity after several explorations, we consider the probability of misjudging the capacity, i.e., the probabilities:

$$\begin{aligned} &\mathbb{P}_1\left[\hat{m}=m+1\right], \\ &\mathbb{P}_2\left[\hat{m}=m\right]. \end{aligned}$$

where $\hat{m}$ is the estimator given by the strategy $\pi$, and $\mathbb{P}_1$,$\mathbb{P}_2$ are the probability measures defined on the whole $T$ explorations where the real capacities are $m$ and $m+1$, respectively.

Since there are only two possible values of $m_k$, we have $\{\hat{m}=m+1\} = \{\hat{m}=m\}^C$, meaning that these two events are complementary to each other. This satisfies the condition of Theorem 14.2 in Lattimore and Szepesvári (2020), and we have:

$$\begin{aligned} &\mathbb{P}_1\left[\hat{m}=m+1\right]+\mathbb{P}_2\left[\hat{m}=m\right] \\ \geq &\frac{1}{2}\exp\left(-KL\left(\mathbb{P}_1,\mathbb{P}_2\right)\right). \end{aligned}$$

As for the KL divergence, we use the result obtained in equation (7), which will be derived in the proof of Theorem 9. Let $N\left(T\right)$ be the number of actions assigned by $\pi$ satisfying that $a_t \geq m+1$, and then we have:

$$KL\left(\mathbb{P}_1,\mathbb{P}_2\right) = \mathbb{E}_1\left[N\left(T\right)\right]\frac{\mu^2}{2\sigma^2} \leq T\frac{\mu^2}{2\sigma^2}.$$

If $\pi$ works well for probability at least $\delta$, then we have:

$$\mathbb{P}_1\left[\hat{m}=m+1\right]+\mathbb{P}_2\left[\hat{m}=m\right] \leq 2\delta.$$

Consequently we get:

$$\begin{aligned} &2\delta \\ \geq &\mathbb{P}_1\left[\hat{m}=m+1\right]+\mathbb{P}_2\left[\hat{m}=m\right] \\ \geq &\frac{1}{2}\exp\left(-KL\left(\mathbb{P}_1,\mathbb{P}_2\right)\right) \\ \geq &\frac{1}{2}\exp\left(-T\frac{\mu^2}{2\sigma^2}\right). \end{aligned}$$

By rearranging the terms we get:

$$T \geq \frac{2\sigma^2}{\mu^2}\log\left(\frac{1}{4\delta}\right).$$

■

**Proof** (Lemma 4)

The learning process of the confidence intervals of $\hat{\mu}_{k,t}$ $\hat{\upsilon}_{k,t}$ and $m_k$ resembles a "chicken-and-egg" problem, where the estimators of $\hat{\mu}_{k,t}$ and $\hat{\upsilon}_{k,t}$ depend on $m^l_{k,t-1}$ and $m^u_{k,t-1}$, and the confidence intervals of $m_k$ are updated based on $\hat{\mu}_{k,t}$ and $\hat{\upsilon}_{k,t}$. We first assume that for all $t \in [T]$, $m_k \in [m^l_{k,t-1}, m^u_{k,t-1}]$.

Consider the IEs on the arm $k$ as actions in an one-arm linear bandit. Denote $\epsilon_{k,i}$ as the sub-Gaussian noise on arm $k$ in round $i$. Let

$$\begin{aligned}
\hat{V}_{k,t} &:= \sum_{i=1}^{t} a_{k,i}^2 \cdot \mathbb{1}(a_{k,i} \le m^l_{k,i-1}), \\
S_{k,t} &:= \sum_{i=1}^{t} a_{k,i}\epsilon_{k,i} \cdot \mathbb{1}(a_{k,i} \le m^l_{k,i-1}), \\
M_t(x) &= \exp\left(x \cdot S_{k,t} - \frac{\sigma^2}{2}\hat{V}_{k,t}x^2\right).
\end{aligned}$$

According to lemma 20.2 Lattimore and Szepesvári (2020), it is straightforward to verify that $M_t(x)$ is an $\mathbb{F}$-adapted non-negative supermartingale with $M_0(x) = 1$.

We then set $x$ to follow the distribution of $\mathcal{N}\left(0, \frac{1}{\sigma^2}\right)$. According to lemma 20.3 Lattimore and Szepesvári (2020), $\bar{M}_t = \int_{\mathbb{R}} M_t(x) \cdot \frac{\sigma}{\sqrt{2\pi}} \exp\left(-\frac{\sigma^2}{2}x^2\right) dx$ is an $\mathbb{F}$-adapted nonnegative supermartingale with $\bar{M}_0 = 1$.

According to Theorem 3.9 Lattimore and Szepesvári (2020), we have:

$$\mathbb{P}\left(\sup_{t\in\mathbb{N}} \bar{M}_t \ge \frac{1}{\delta}\right) \le \delta.$$

It is straightforward to derive that $\bar{M}_t = \exp\left(\frac{S^2_{k,t}}{2\sigma^2(V_{k,t}+1)}\right) \cdot \frac{1}{\sqrt{V_{k,t}+1}}$. By rearranging the terms and setting $\delta$ as $\delta/2$, we get:

$$\left|\frac{S_{k,t}}{\hat{V}_{k,t}}\right| \ge \sigma\sqrt{\frac{2\hat{V}_{k,t}+2}{\hat{V}^2_{k,t}}\log\frac{2}{\delta} + \frac{(\hat{V}_{k,t}+1)\log(\hat{V}_{k,t}+1)}{\hat{V}^2_{k,t}}} = \sigma\phi\left(\hat{V}_{k,t}, \delta\right).$$

holds with probability less than $\delta/2$ for all $t$. Since $\hat{\mu}_{k,t} - \mu_k = \frac{S_{k,t}}{\hat{V}_{k,t}}$, we finally reach the confidence interval of the estimator $\hat{\mu}_{k,t}$:

$$\hat{\mu}_{k,t} \in \left[\mu_k - \sigma\phi(\hat{V}_{k,t}, \delta), \mu_k + \sigma\phi(\hat{V}_{k,t}, \delta)\right],$$

and this confidence interval holds with probability greater than $1 - \delta/2$.

We then consider the UEs on the arm $k$ as actions in an one-arm linear bandit. Similarly we get the confidence interval for $m_k\mu_k$ as :

$$\hat{\upsilon}_{k,t} \in \left[m_k\mu_k - \sigma\phi\left(\hat{\imath}_{k,t}, \delta\right), m_k\mu_k + \sigma\phi\left(\hat{\imath}_{k,t}, \delta\right)\right],$$

and this confidence interval holds with probability greater than $1 - \delta/2$.

Using the endpoints of the confidence interval of $m_k$ and $m_k\mu_k$, we get $m_k$'s confidence interval $[m^l_{k,t}, m^u_{k,t}]$ as:

$$\begin{aligned}
m^l_{k,t} &= \max\left(\left\lceil \frac{\hat{\upsilon}_{k,t} - \sigma\phi\left(\hat{\iota}_{k,t}, \delta\right)}{\hat{\mu}_{k,t} + \sigma\phi(\hat{V}_{k,t}, \delta)} \right\rceil, 1\right), \\
m^u_{k,t} &= \min\left(\left\lfloor \frac{\hat{\upsilon}_{k,t} + \sigma\phi\left(\hat{\iota}_{k,t}, \delta\right)}{\hat{\mu}_{k,t} - \sigma\phi(\hat{V}_{k,t}, \delta)} \right\rfloor, N\right).
\end{aligned}$$

This confidence interval $[m^l_{k,t}, m^u_{k,t}]$ is correct for all $t \in [T]$ with probability at least $1 - \delta$.

As for the assumption that $m_k \in [m^l_{k,t-1}, m^u_{k,t-1}]$ for all $t \in [T]$, it follows naturally as a corollary from the confidence interval of $\hat{\mu}_{k,t}$,$\hat{\upsilon}_{k,t}$, provided that the initializations of $m^l_{k,t}$ and $m^u_{k,t}$ are correct. Given that $[1, N - K + 1]$ is clearly a valid initialization for $[m^l_{k,t-1}, m^u_{k,t-1}]$, all three confidence intervals hold for all $t \in [T]$ with probability at least $1 - \delta$. ■

**Proof** (Theorem 5)

We then consider the scenario where there is only one arm $k$, with its $m_k$ and $\mu_k$ unknown. Our objective is to determine the correct $m_k$ by reducing $m^u_{k,t} - m^l_{k,t}$ . A sufficient condition for the confidence interval to converge is:

$$\frac{\hat{\upsilon}_{k,t} + \sigma\phi\left(\hat{\iota}_{k,t}, \delta\right)}{\hat{\mu}_{k,t} - \sigma\phi(\hat{V}_{k,t}, \delta)} - \frac{\hat{\upsilon}_{k,t} - \sigma\phi\left(\hat{\iota}_{k,t}, \delta\right)}{\hat{\mu}_{k,t} + \sigma\phi(\hat{V}_{k,t}, \delta)} < 1,$$

with the assumption that $\hat{\mu}_{k,t} - \sigma\phi(\hat{V}_{k,t}, \delta) > 0$.

Replacing the empirical values $\hat{\upsilon}_{k,t}$ and $\hat{\mu}_{k,t}$ with the endpoints of their confidence interval, we derive another sufficient condition:

$$\frac{m_k\mu_k + 2\sigma\phi\left(\hat{\iota}_{k,t}, \delta\right)}{\mu_k - 2\sigma\phi(\hat{V}_{k,t}, \delta)} - \frac{m_k\mu_k - 2\sigma\phi\left(\hat{\iota}_{k,t}, \delta\right)}{\mu_k + 2\sigma\phi(\hat{V}_{k,t}, \delta)} < 1,$$

with the assumption that $\mu_k - 2\sigma\phi(\hat{V}_{k,t}, \delta) > 0$. We further assume that $\mu_k > 4\sigma\phi(\hat{V}_{k,t}, \delta)$, and the left-hand side of the inequality above can be bounded as follows:

$$\begin{aligned}
&\frac{m_k\mu_k + 2\sigma\phi\left(\hat{\iota}_{k,t}, \delta\right)}{\mu_k - 2\sigma\phi(\hat{V}_{k,t}, \delta)} - \frac{m_k\mu_k - 2\sigma\phi\left(\hat{\iota}_{k,t}, \delta\right)}{\mu_k + 2\sigma\phi(\hat{V}_{k,t}, \delta)} \\
=&\left(\frac{m_k\mu_k + 2\sigma\phi\left(\hat{\iota}_{k,t}, \delta\right)}{\mu_k - 2\sigma\phi(\hat{V}_{k,t}, \delta)} - m_k\right) + \left(m_k - \frac{m_k\mu_k - 2\sigma\phi\left(\hat{\iota}_{k,t}, \delta\right)}{\mu_k + 2\sigma\phi(\hat{V}_{k,t}, \delta)}\right) \\
\leq&\frac{4\sigma\phi\left(\hat{\iota}_{k,t}, \delta\right) + 4m_k\sigma\phi(\hat{V}_{k,t}, \delta)}{\mu_k} + \frac{2\sigma\phi\left(\hat{\iota}_{k,t}, \delta\right) + 2m_k\sigma\phi(\hat{V}_{k,t}, \delta)}{\mu_k} \\
=&6\frac{\sigma\phi\left(\hat{\iota}_{k,t}, \delta\right) + m_k\sigma\phi(\hat{V}_{k,t}, \delta)}{\mu_k}.
\end{aligned}$$

A sufficient condition for $m^l_{k,t} = m^u_{k,t}$ is then given as:

$$\sigma\phi\left(\hat{\iota}_{k,t}, \delta\right) \leq \frac{1}{12}\mu_k,$$
$$\sigma\phi(\hat{V}_{k,t}, \delta) \leq \frac{1}{12m_k}\mu_k.$$

Note that $\frac{1}{12m_k} \leq \frac{1}{4}$, so the assumption $\mu_k > 4\sigma\phi(\hat{V}_{k,t}, \delta)$ holds if the sufficient condition above is satisfied.

Solving the inequalities, we obtain:

$$\hat{\iota}_{k,t} \geq \frac{144\sigma^2}{\mu_k^2} \cdot 8 \cdot \log\left(\frac{2}{\delta}\right),$$
$$\hat{V}_{k,t} \geq \frac{144\sigma^2 m_k^2}{\mu_k^2} \cdot 8 \cdot \log\left(\frac{2}{\delta}\right).$$

We then bound the number of UEs and IEs required for $m^l_{k,t} \geq m_k/2$. A sufficient condition for satisfying the above inequalities is:

$$\hat{\iota}_{k,t} \geq \frac{64\sigma^2}{m_k^2\mu_k^2} \cdot 8 \cdot \log\left(\frac{2}{\delta}\right),$$
$$\hat{V}_{k,t} \geq \frac{16\sigma^2}{\mu_k^2} \cdot 8 \cdot \log\left(\frac{2}{\delta}\right).$$

Since $m_k \geq 1$ and $\hat{V}_{k,t} \geq t/2$, after at most $\frac{1024\sigma^2}{\mu_k^2}\log\left(\frac{2}{\delta}\right)$ time slots, we have $m^l_{k,t} \geq m_k/2$. For subsequent IEs, $a_{k,t} \geq m_k/2$. To ensure $\hat{V}_{k,t} \geq \frac{144\sigma^2 m_k^2}{\mu_k^2} \cdot 8 \cdot \log\left(\frac{2}{\delta}\right)$, at most $\frac{4608\sigma^2}{\mu_k^2}\log\left(\frac{2}{\delta}\right)$ additional IEs are required. Therefore, after at most $\frac{9216\sigma^2}{\mu_k^2}\log\left(\frac{2}{\delta}\right)$ additional time slots, the capacity $m_k$ is determined. Consequently, the maximum number of time slots required is $\frac{10240\sigma^2}{\mu_k^2}\log\left(\frac{2}{\delta}\right)$, aligning with the lower bound of the sample complexity.

Noting that in the first two explorations, we assign 1 and $N$ plays to the arm respectively, a constant 2 should be added to the upper bound. This proof is then complete. ■

### B.2. Regret Lower Bound Proof

**Proof** (Theorem 7)

To avoid unnecessary mathematical complexities and simplify the proof, we focus on the case where $M/K$ and $K/4$ are both integers. We first construct two instances of the problem as follows:

- Instance $E_1$: each arm whose index is an odd number has $\left(\frac{M}{K} - 1\right)$ units of capacity and each of the remaining arms has $\left(\frac{M}{K} + 1\right)$ units of capacity. The per unit reward mean is fixed to $\mu$, i.e., $\mu_1 = \ldots = \mu_K = \mu$, and variance is fixed to $\sigma$, i.e., $\sigma_1 = \ldots = \sigma_K = \sigma$. Formally,

| | arm 1 | arm 2 | | arm $K-1$ | arm $K$ |
|---|---|---|---|---|---|
| Instance $E_1$: | $M/K - 1$ | $M/K + 1$ | $\cdots$ | $M/K - 1$ | $M/K + 1$ |
| | $\mu, \sigma$ | $\mu, \sigma$ | | $\mu, \sigma$ | $\mu, \sigma$ |

- Instance $E_2$: each arm whose index is an even number has $\left(\frac{M}{K}-1\right)$ units of capacity and each of the remaining arms has $\left(\frac{M}{K}+1\right)$ units of capacity. The per unit reward mean is fixed to $\mu$, i.e., $\mu_1 = \ldots = \mu_K = \mu$, and variance is fixed to $\sigma$, i.e., $\sigma_1 = \ldots = \sigma_K = \sigma$. Formally,

$$\begin{array}{cccccc} & \text{arm } 1 & \text{arm } 2 & & \text{arm } K-1 & \text{arm } K \\ \text{Instance } E_2\text{:} & M/K+1 & M/K-1 & \cdots & M/K+1 & M/K-1 \\ & \mu,\sigma & \mu,\sigma & & \mu,\sigma & \mu,\sigma \end{array}$$

For an arbitrary learning algorithm or strategy $\pi$, let $Reg^1\left(T,\pi\right)$ and $Reg^2\left(T,\pi\right)$ represent $\pi$'s regrets in instance $E_1$ and $E_2$ respectively. Let $T_1$ denote the number of time slots during which at least $\frac{K}{4}$ arms with odd indices are assigned exactly $\left(\frac{M}{K}-1\right)$ plays. Define $B$ as the event where $T_1 \geq \frac{1}{2}T$:

$$B = \left\{T_1 \geq \frac{1}{2}T\right\}.$$

We use event $B$ to bound the expectation of the regret in $E_1$ as follows:

$$\begin{aligned} &\mathbb{E}_{E_1}\left[Reg^1\left(T,\pi\right)\right] \\ =&\mathbb{E}_{E_1}\left[Reg^1\left(T,\pi\right)\mathbb{1}\left\{B\right\}\right] + \mathbb{E}_{E_1}\left[Reg^1\left(T,\pi\right)\mathbb{1}\left\{B^C\right\}\right] \\ \geq& 0 + \frac{TK}{8}\min\left(\mu-c,c\right)\mathbb{P}_{E_1}\left(B^C\right). \end{aligned}$$

Similarly we have

$$\mathbb{E}_{E_2}\left[Reg^2\left(T,\pi\right)\right] \geq \frac{TK}{8}\cdot 2\left(\mu-c\right)\mathbb{P}_{E_2}\left(B\right).$$

Note that Theorem 14.2 in Lattimore and Szepesvári (2020) indicates:

$$\mathbb{P}_{E_1}\left(B^C\right) + \mathbb{P}_{E_2}\left(B\right) \geq \frac{1}{2}\exp\left(-KL\left(\mathbb{P}_{E_1},\mathbb{P}_{E_2}\right)\right).$$

Then, the sum of the regrets of $\pi$ in the two instances can be lower-bounded as:

$$\begin{aligned} &\mathbb{E}_{E_1}\left[Reg^1\left(T,\pi\right)\right] + \mathbb{E}_{E_2}\left[Reg^2\left(T,\pi\right)\right] \\ \geq&\frac{TK}{8}\min\left(\mu-c,c\right)\left(\mathbb{P}_{E_1}\left(B^C\right)+\mathbb{P}_{E_2}\left(B\right)\right) \\ \geq&\frac{TK}{16}\min\left(\mu-c,c\right)\exp\left(-KL\left(\mathbb{P}_{E_1},\mathbb{P}_{E_2}\right)\right). \end{aligned}$$

Note that the probability measure $\mathbb{P}_{E_1}$ is defined over the entire learning process spanning $T$ time slots,i.e.

$$\mathbb{P}_{E_1}\left[\boldsymbol{a}_1,\boldsymbol{x}_1,...,\boldsymbol{a}_T,\boldsymbol{x}_T\right] = \prod_{t=1}^{T}\pi_t\left(\boldsymbol{a}_t|\boldsymbol{a}_1,\boldsymbol{x}_1,...,\boldsymbol{a}_{t-1},\boldsymbol{x}_{t-1}\right)P_{E_1,\boldsymbol{a}_t}\left(\boldsymbol{x}_t\right).$$

Here, $\boldsymbol{a}_t$ is the action chosen at the time slot $t$ and the vector $\boldsymbol{x}_t$ is the resulting reward on the $K$ arms after playing $\boldsymbol{a}_t$. $\pi_t$ is the probability measure of the action $\boldsymbol{a}_t$ based on the observation of the past $t-1$ pairs of actions and rewards. $P_{E_1,\boldsymbol{a}_t}$ is the probability measure of the reward vector $\boldsymbol{x}_t$ for the fixed action $\boldsymbol{a}_t$ in instance $E_1$. Regarding the calculation of the KL-divergence, it can be decomposed into $T$ parts:

$$\begin{aligned}
&KL\left(\mathbb{P}_{E_1}, \mathbb{P}_{E_2}\right)\\
=&\mathbb{E}_{E_1}\left[\log\left(\frac{d\mathbb{P}_{E_1}}{d\mathbb{P}_{E_2}}\right)\right]\\
=&\mathbb{E}_{E_1}\left[\sum_{t=1}^{T}\log\frac{P_{E_1,\boldsymbol{a}_t}\left(\boldsymbol{x}_t\right)}{P_{E_2,\boldsymbol{a}_t}\left(\boldsymbol{x}_t\right)}\right]\\
=&\sum_{t=1}^{T}\mathbb{E}_{E_1}\left[\log\frac{P_{E_1,\boldsymbol{a}_t}\left(\boldsymbol{x}_t\right)}{P_{E_2,\boldsymbol{a}_t}\left(\boldsymbol{x}_t\right)}\right]\\
=&\sum_{t=1}^{T}\mathbb{E}_{E_1}\left[\mathbb{E}_{E_1}\left[\log\frac{P_{E_1,\boldsymbol{a}_t}\left(\boldsymbol{x}_t\right)}{P_{E_2,\boldsymbol{a}_t}\left(\boldsymbol{x}_t\right)}\middle|\boldsymbol{a}_t\right]\right]\\
=&\sum_{t=1}^{T}\mathbb{E}_{E_1}\left[KL\left(P_{E_1,\boldsymbol{a}_t}, P_{E_2,\boldsymbol{a}_t}\right)\right],
\end{aligned}$$

where in the last equality we use the fact that under $\mathbb{P}_{E_1}\left(\cdot|\boldsymbol{a}_t\right)$, the distribution of $\boldsymbol{x}_t$ is $P_{E_1,\boldsymbol{a}_t}$.

Since the measure $P_{E_1,\boldsymbol{a}_t}$ is a product of $K$ independent probability measures, we can decompose the KL divergence as follows:

$$KL\left(P_{E_1,\boldsymbol{a}_t}, P_{E_2,\boldsymbol{a}_t}\right) = \sum_{k=1}^{K} KL\left(P_{E_1,a_{k,t}}, P_{E_2,a_{k,t}}\right).$$

Here, $P_{E_1,a_{k,t}}$ and $P_{E_2,a_{k,t}}$ follow normal distributions:

$$P_{E_1,a_{k,t}} \sim \mathcal{N}\left(\min\left(a_{k,t}, m_k^{(1)}\right)\mu - a_{k,t}c, \sigma^2\right),$$

$$P_{E_2,a_{k,t}} \sim \mathcal{N}\left(\min\left(a_{k,t}, m_k^{(2)}\right)\mu - a_{k,t}c, \sigma^2\right),$$

and $m_k^{(1)}$ and $m_k^{(2)}$ denote the capacities of arm $k$ in the $E_1$ and $E_2$, respectively. The KL-divergence of two Gaussian distributions is given by the following formula:

**Lemma 18** *For each $i \in \{1,2\}$, let $\mu_i \in \mathbb{R}, \sigma_i^2 > 0$ and $P_i = \mathcal{N}\left(\mu_i, \sigma_i^2\right)$. Then we have:*

$$KL\left(P_1, P_2\right) = \frac{1}{2}\left(\log\left(\frac{\sigma_2^2}{\sigma_1^2}\right) + \frac{\sigma_1^2}{\sigma_2^2} - 1\right) + \frac{\left(\mu_1 - \mu_2\right)^2}{2\sigma_2^2}.$$

Applying lemma 18, we have:

$$KL\left(P_{E_1,a_{1,t}}, P_{E_2,a_{1,t}}\right) = \frac{\left(\min\left(a_{1,t}, m_k^{(1)}\right)\mu - \min\left(a_{1,t}, m_k^{(2)}\right)\mu\right)^2}{2\sigma^2}.$$

We aim to find the action $a_{1,t}$ that maximizes $KL\left(P_{E_1,a_{1,t}}, P_{E_2,a_{1,t}}\right)$ at time slot $t$ on the first arm. It is straightforward to observe that $a_{1,t}$ should be no less than $m_1^{(2)} = \frac{M}{K}+1$ to achieve the maximum $KL\left(P_{E_1,a_{1,t}}, P_{E_2,a_{1,t}}\right)$. The same principle applies to other arms $k$ with odd indices. Similarly, to maximize $KL\left(P_{E_1,a_{2,t}}, P_{E_2,a_{2,t}}\right)$, the action $a_{2,t}$ for the second arm should satisfy $a_{2,t} \geq m_2^{(1)} = \frac{M}{K}+1$. The same is true for other arms $k$ with even indices. Therefore, we conclude that:

$$\begin{aligned}KL\left(P_{E_1,a_{1,t}}, P_{E_2,a_{1,t}}\right) \leq \frac{2\mu^2}{\sigma^2},\\ KL\left(P_{E_1,a_{2,t}}, P_{E_2,a_{2,t}}\right) \leq \frac{2\mu^2}{\sigma^2}.\end{aligned}$$

It is worth noting that $a_{1,t}, a_{2,t}, ..., a_{K,t}$ may not all be simultaneously feasible in the real world. However, this poses no issue since our focus is solely on the upper bound of the KL-divergence.

Note that $\mathbb{E}[X] \leq \max[X]$. We obtain that:

$$\begin{aligned}&KL\left(\mathbb{P}_{E_1}, \mathbb{P}_{E_2}\right)\\ =&\sum_{t=1}^{T} \mathbb{E}_{E_1}\left[KL\left(P_{E_1,\boldsymbol{a}_t}, P_{E_2,\boldsymbol{a}_t}\right)\right]\\ \leq& T\cdot \max_{\mathbf{a}\in\mathcal{A}}\left[KL\left(P_{E_1,\boldsymbol{a}}, P_{E_2,\boldsymbol{a}}\right)\right]\\ =&T\cdot \max_{\mathbf{a}\in\mathcal{A}}\left[\sum_{k=1}^{K} KL\left(P_{E_1,a_k}, P_{E_2,a_k}\right)\right]\\ \leq& T\cdot\sum_{k=1}^{K}\max_{a_k\in[N]}\left[KL\left(P_{E_1,a_k}, P_{E_2,a_k}\right)\right]\\ \leq& T\cdot\sum_{k=1}^{K}\frac{2\mu^2}{\sigma^2}\\ =&TK\frac{2\mu^2}{\sigma^2}.\end{aligned}$$

Furthermore, by letting $c = \frac{1}{2}\mu$, we obtain that:

$$\begin{aligned}&\mathbb{E}_{E_1}\left[Reg^1(T,\pi)\right] + \mathbb{E}_{E_2}\left[Reg^2(T,\pi)\right]\\ \geq&\frac{TK}{16}\min(\mu-c,c)\exp\left(-KL\left(\mathbb{P}_{E_1}, \mathbb{P}_{E_2}\right)\right)\\ =&\frac{TK}{32}\mu\exp\left(-KL\left(\mathbb{P}_{E_1}, \mathbb{P}_{E_2}\right)\right)\\ \geq&\frac{TK}{32}\mu\exp\left(-2TK\frac{\mu^2}{\sigma^2}\right).\end{aligned}$$

Let $\mu = \sigma/\sqrt{2TK}$. Then we obtain that:

$$\max\left(\mathbb{E}_{E_1}\left[Reg^1(T,\pi)\right], \mathbb{E}_{E_2}\left[Reg^2(T,\pi)\right]\right) \geq \frac{\sigma}{64e\sqrt{2}}\sqrt{TK}.$$

This proof is then complete. ■

**Proof** (Theorem 9)

Here we only consider the set of algorithms which are consistent over the class of MP-MAB $\mathcal{E}$ described in section 3. Additionally, for simplicity, we assume that the perturbation of the returned utility follows a Gaussian distribution $\mathcal{N}\left(0,\sigma^2\right)$, where $\sigma^2 \leq 1/2$ .

**Definition 19** *A policy $\pi$ is defined as consistent over a class of bandits $\mathcal{E}'$ if, for all $E \in \mathcal{E}'$ and $p > 0$, it holds that :*

$$\lim_{T\to\infty} \frac{Reg\,(T)}{T^p} = 0.$$

First, we choose a consistent policy $\pi$. Let $E_1 \in \mathcal{E}$ be an instance, where the arm $k$ has $m_k$ units of capacity, each with utility $\mu_k$. Next, we consider the number of time slots $TB_k\,(T)$ during which arm $k$ is assigned more than $m_k$ plays by $\pi$ in $T$ time slots, i.e.,

$$TB_k\,(T) := \sum_{t=1}^{T} \mathbb{1}\left\{a_{k,t} \geq m_k + 1\right\}.$$

For a fixed $k \in [K]$, let $E_2 \in \mathcal{E}$ be another instance, where for $j \neq k$, there are $m_j$ units of capacity with unit utility $\mu_j$ on the arm $j$. On the arm $k$ in $E_2$, there are $m_k + 1$ units of capacity with unit utility $\mu_k$. Let $B$ be the event that $TB_k \leq \frac{T}{2}$:

$$B := \left\{TB_k \leq \frac{T}{2}\right\}.$$

Let $Reg^1\,(T,\pi)$,$Reg^2\,(T,\pi)$ denote the regret of policy $\pi$ in instances $E_1$ and $E_2$, respectively. By a similar analysis as in the previous subsection, we obtain that:

$$\begin{aligned}
&\mathbb{E}_{E_1}\left[Reg^1\,(T,\pi)\right] \\
=&\mathbb{E}_{E_1}\left[Reg^1\,(T,\pi)\,\mathbb{1}\left\{B\right\}\right] + \mathbb{E}_{E_1}\left[Reg^1\,(T,\pi)\,\mathbb{1}\left\{B^C\right\}\right] \\
\geq& 0 + \frac{T}{2}c\mathbb{P}_{E_1}\left(B^C\right).
\end{aligned}$$

Similarly, we obtain that:

$$\mathbb{E}_{E_2}\left[Reg^2\,(T,\pi)\right] \geq \frac{T}{2}\left(\mu_k - c\right)\mathbb{P}_{E_2}\left(B\right).$$

Then, the sum of the expected regrets of $\pi$ in the two instances can be lower-bounded as:

$$\begin{aligned}
&\mathbb{E}_{E_1}\left[Reg^1\,(T,\pi)\right] + \mathbb{E}_{E_2}\left[Reg^2\,(T,\pi)\right] \\
\geq&\frac{T}{2}\min\left(\mu_k - c, c\right)\left(\mathbb{P}_{E_1}\left(B^C\right) + \mathbb{P}_{E_2}\left(B\right)\right) \\
\geq&\frac{T}{4}\min\left(\mu_k - c, c\right)\exp\left(-KL\left(\mathbb{P}_{E_1},\mathbb{P}_{E_2}\right)\right).
\end{aligned}$$

As for the KL-divergence, we can decompose it across time slots and arms, as shown in the previous subsection:

$$\begin{aligned}
&KL\left(\mathbb{P}_{E_1}, \mathbb{P}_{E_2}\right)\\
=&\sum_{t=1}^{T} \mathbb{E}_{E_1}\left[KL\left(P_{E_1,\boldsymbol{a}_t}, P_{E_2,\boldsymbol{a}_t}\right)\right]\\
=&\sum_{t=1}^{T} \mathbb{E}_{E_1}\left[\sum_{i=1}^{K} KL\left(P_{E_1,a_{i,t}}, P_{E_2,a_{i,t}}\right)\right].
\end{aligned}$$

Note that $E_1$ and $E_2$ only differ in arm $k$. Therefore, the above summation can be simplified to:

$$\begin{aligned}
&\sum_{t=1}^{T} \mathbb{E}_{E_1}\left[\sum_{i=1}^{K} KL\left(P_{E_1,a_{i,t}}, P_{E_2,a_{i,t}}\right)\right]\\
=&\sum_{t=1}^{T} \mathbb{E}_{E_1}\left[KL\left(P_{E_1,a_{k,t}}, P_{E_2,a_{k,t}}\right)\right]\\
=&\sum_{t=1}^{T} \mathbb{E}_{E_1}\left[KL\left(P_{E_1,a_{k,t}}, P_{E_2,a_{k,t}}\right)\mathbb{1}\left\{a_{k,t} \geq m_k + 1\right\}\right]\\
+&\sum_{t=1}^{T} \mathbb{E}_{E_1}\left[KL\left(P_{E_1,a_{k,t}}, P_{E_2,a_{k,t}}\right)\mathbb{1}\left\{a_{k,t} \leq m_k\right\}\right]\\
=&\sum_{t=1}^{T} \mathbb{E}_{E_1}\left[KL\left(P_{E_1,a_{k,t}}, P_{E_2,a_{k,t}}\right)\mathbb{1}\left\{a_{k,t} \geq m_k + 1\right\}\right] + 0.
\end{aligned}$$

According to lemma 18, when $a_{k,t} \geq m_k + 1$, we obtain that:

$$KL\left(P_{E_1,a_{k,t}}, P_{E_2,a_{k,t}}\right) = \frac{\mu_k^2}{2\sigma^2}.$$

Therefore, we obtain that :

$$\begin{aligned}
&\sum_{t=1}^{T} \mathbb{E}_{E_1}\left[KL\left(P_{E_1,a_{k,t}}, P_{E_2,a_{k,t}}\right)\mathbb{1}\left\{a_{k,t} \geq m_k + 1\right\}\right]\\
=&\sum_{t=1}^{T} \mathbb{E}_{E_1}\left[\mathbb{1}\left\{a_{k,t} \geq m_k + 1\right\}\right]\frac{\mu_k^2}{2\sigma^2}\\
=&\mathbb{E}_{E_1}\left[\sum_{t=1}^{T} \mathbb{1}\left\{a_{k,t} \geq m_k + 1\right\}\right]\frac{\mu_k^2}{2\sigma^2}\\
=&\mathbb{E}_{E_1}\left[TB_k\left(T\right)\right]\frac{\mu_k^2}{2\sigma^2}.
\end{aligned}$$

Consequently, we calculate the KL-divergence as :

$$KL\left(\mathbb{P}_{E_1}, \mathbb{P}_{E_2}\right) = \mathbb{E}_{E_1}\left[TB_k\left(T\right)\right]\frac{\mu_k^2}{2\sigma^2}. \tag{7}$$

Combining (7) with the lower bound of the sum of the expected regrets, we obtain that:

$$\mathbb{E}_{E_1}\left[Reg^1\left(T,\pi\right)\right]+\mathbb{E}_{E_2}\left[Reg^2\left(T,\pi\right)\right]\geq\frac{T}{4}\min\left(\mu_k-c,c\right)\exp\left(-\mathbb{E}_{E_1}\left[TB_k\left(T\right)\right]\frac{\mu_k^2}{2\sigma^2}\right).$$

Rearranging and taking the limit inferior on $T$ leads to:

$$\begin{aligned}\liminf_{T\to\infty}\frac{\mathbb{E}_{E_1}\left[TB_k\left(T\right)\right]}{\log\left(T\right)}\geq&\frac{2\sigma^2}{\mu_k^2}\liminf_{T\to\infty}\frac{\log\left(\frac{T\min(\mu_k-c,c)}{4\left(\mathbb{E}_{E_1}\left[Reg^1(T,\pi)\right]+\mathbb{E}_{E_2}\left[Reg^2(T,\pi)\right]\right)}\right)}{\log\left(T\right)}\\=&\frac{2\sigma^2}{\mu_k^2}\left(1-\limsup_{T\to\infty}\frac{\log\left(\mathbb{E}_{E_1}\left[Reg^1\left(T,\pi\right)\right]+\mathbb{E}_{E_2}\left[Reg^2\left(T,\pi\right)\right]\right)}{\log\left(T\right)}\right).\end{aligned}$$

Since the policy $\pi$ is consistent, for any $p>0$ , there is a constant $C_p$ such that for sufficiently large $T$: $\mathbb{E}_{E_1}\left[Reg^1\left(T,\pi\right)\right]+\mathbb{E}_{E_2}\left[Reg^2\left(T,\pi\right)\right]\leq C_pT^p$, which implies that:

$$\begin{aligned}&\limsup_{T\to\infty}\frac{\log\left(\mathbb{E}_{E_1}\left[Reg^1\left(T,\pi\right)\right]+\mathbb{E}_{E_2}\left[Reg^2\left(T,\pi\right)\right]\right)}{\log\left(T\right)}\\\leq&\limsup_{T\to\infty}\frac{p\log\left(T\right)+\log\left(C_p\right)}{\log\left(T\right)}\\=&p.\end{aligned}$$

Noting that $p$ can be arbitrarily small, we obtain that:

$$\limsup_{T\to\infty}\frac{\log\left(\mathbb{E}_{E_1}\left[Reg^1\left(T,\pi\right)\right]+\mathbb{E}_{E_2}\left[Reg^2\left(T,\pi\right)\right]\right)}{\log\left(T\right)}=0.$$

Consequently,

$$\liminf_{T\to\infty}\frac{\mathbb{E}_{E_1}\left[TB_k\left(T\right)\right]}{\log\left(T\right)}\geq\frac{2\sigma^2}{\mu_k^2}.$$

It is noteworthy that:

$$\begin{aligned}&\mathbb{E}_{E_1}\left[Reg_k^1\left(T,\pi\right)\right]\\=&\mathbb{E}_{E_1}\left[\sum_{t=1}^{T}\left[\left(m_k\mu_k-cm_k\right)-\left(\min\left\{a_{k,t},m_k\right\}\cdot\mu_k-c\cdot a_{k,t}\right)\right]\right]\\\geq&\mathbb{E}_{E_1}\left[\sum_{t=1}^{T}\left[\left(m_k\mu_k-cm_k\right)-\left(\min\left\{a_{k,t},m_k\right\}\cdot\mu_k-c\cdot a_{k,t}\right)\right]\mathbb{1}\left\{a_{k,t}\geq m_k+1\right\}\right]\\\geq&\mathbb{E}_{E_1}\left[\sum_{t=1}^{T}c\cdot\mathbb{1}\left\{a_{k,t}\geq m_k+1\right\}\right]\\=&c\cdot\mathbb{E}_{E_1}\left[TB_k\left(T\right)\right].\end{aligned}$$

Taking the limit inferior on $T$ leads to:

$$\begin{aligned}&\liminf_{T\to\infty}\frac{\mathbb{E}_{E_1}\left[Reg_k^1\left(T,\pi\right)\right]}{\log\left(T\right)}\\ \geq& c\cdot\liminf_{T\to\infty}\frac{\mathbb{E}_{E_1}\left[TB_k\left(T\right)\right]}{\log\left(T\right)}\\ \geq& c\cdot\frac{2\sigma^2}{\mu_k^2}.\end{aligned}$$

The proof is complete. ■

### B.3. Regret Upper Bound Proof

**Proof** (Theorem 12)

The expectation of $Reg_k\left(T\right)$ can be separated by the event $A_k$:

$$\begin{aligned}&\mathbb{E}\left[Reg_k\left(T\right)\right]\\ =&\mathbb{E}\left[Reg_k\left(T\right)\mathbb{1}\{\,A_k\,\}\right]+\mathbb{E}\left[Reg_k\left(T\right)\mathbb{1}\{\,A_k^C\,\}\right]\\ \leq&\mathbb{E}\left[Reg_k\left(T\right)\mathbb{1}\{\,A_k\,\}\right]+\mathbb{P}\left(A_k^C\right)\max\left(\mathbb{E}\left[Reg_k\left(T\right)\right]\right).\end{aligned}$$

$\max\left(\mathbb{E}\left[Reg_k\left(T\right)\right]\right)$ can be bounded by $T$ multiplied by the maximum per-time-slot regret on the arm $k$, which is generated either by an IE with only one play or a UE with all $N$ plays. Let $Regmax_k$ represent the maximal per-time-slot regret on arm $k$, so we have $Regmax_k \leq \max\left(m_k\mu_k, Nc\right)$. Therefore, the second term can be bounded by $\delta T\cdot Regmax_k$.

The first term can be split into two parts: the regret caused by IEs and the regret caused by UEs. When event $A_k$ occurs, according to Lemma 4, the confidence intervals of $m_k$ are correct. We will then examine the convergence of $m_{k,t}^l$ and $m_{k,t}^u$ to bound the regret caused by IEs and UEs, respectively.

As for the regret caused by IEs, it is known that the $m_{k,t}^l$ increases until $m_{k,t}^l = m_k$, and this process will terminate within a finite number of time slots, as shown in the sample complexity results. So we then consider the number of time slots required for $m_{k,t}^l \geq \lambda+1$, conditioned on $m_{k,t}^l \geq \lambda$ for $\lambda \leq m_k - 1$. A sufficient condition for $m_{k,t}^l \geq \lambda+1$ is $m_{k,t}^l > \lambda$:

$$\frac{m_k\mu_k - 2\sigma\phi\left(\hat{\iota}_{k,t},\delta\right)}{\mu_k + 2\sigma\phi(\hat{V}_{k,t},\delta)} > \lambda.$$

By rearranging the terms, we obtain that:

$$\left(m_k-\lambda\right)\mu_k > 2\lambda\sigma\phi(\hat{V}_{k,t},\delta) + 2\sigma\phi\left(\hat{\iota}_{k,t},\delta\right).$$

A sufficient condition for this is :

$$\begin{aligned}\phi\left(\hat{\iota}_{k,t},\delta\right) &< \frac{\mu_k}{4\sigma}\left(m_k-\lambda\right),\\ \phi(\hat{V}_{k,t},\delta) &< \frac{\mu_k}{4\sigma}\frac{m_k-\lambda}{\lambda}.\end{aligned}$$

By solving the inequalities, we obtain that:

$$\begin{aligned}
\hat{\iota}_{k,t} &> \frac{(4\sigma)^2}{\mu_k^2\left(m_k-\lambda\right)^2}\cdot 8\cdot\log\left(\frac{2}{\delta}\right),\\
\hat{V}_{k,t} &> \frac{(4\sigma)^2}{\mu_k^2\left(m_k-\lambda\right)^2}\cdot 8\cdot\log\left(\frac{2}{\delta}\right)\lambda^2.
\end{aligned}$$

Let $g_1\left(\lambda\right) := \frac{(4\sigma)^2}{\mu_k^2(m_k-\lambda)^2}\cdot 8\cdot\log\left(\frac{2}{\delta}\right)\lambda^2$. If $m_{k,t}^l \geq \lambda$, there should be an additional $\frac{g_1(\lambda)-g_1(\lambda-1)}{\lambda^2}$ IEs. Note that $\frac{g_1(\lambda)-g_1(\lambda-1)}{\lambda^2} \leq 128\frac{\sigma^2}{\mu_k^2}\log\left(\frac{2}{\delta}\right)\cdot 2\frac{\lambda}{m_k-\lambda}\frac{m_k}{(m_k-\lambda)(m_k-\lambda+1)}\frac{1}{\lambda^2}$. Similarly we can calculate the additional UEs required for $\hat{\iota}_{k,t} > \frac{(4\sigma)^2}{\mu_k^2(m_k-\lambda)^2}\cdot 8\cdot\log\left(\frac{2}{\delta}\right)$, and it is clear that more IEs are required than UEs in this case. Note that if $m_{k,t}^l \geq \lambda$, each IE will generate a regret of at most $\left(m_k-\lambda\right)\left(\mu_k-c\right)$. By summing the regret caused by IEs from $\lambda = 1$ to $\lambda = m_k - 1$, we can upper-bound the regret caused by IEs in the entire learning process as follows:

$$\begin{aligned}
&128\frac{\sigma^2\left(\mu_k-c\right)}{\mu_k^2}\log\left(\frac{2}{\delta}\right)\cdot\sum_{\lambda=1}^{m_k-1}\frac{2m_k}{\lambda\left(m_k-\lambda\right)^2}\\
=&128\frac{\sigma^2\left(\mu_k-c\right)}{\mu_k^2}\log\left(\frac{2}{\delta}\right)\cdot\left(\sum_{\lambda=1}^{m_k-1}\frac{m_k}{\lambda\left(m_k-\lambda\right)^2}+\frac{m_k}{\lambda^2\left(m_k-\lambda\right)}\right)\\
=&128\frac{\sigma^2\left(\mu_k-c\right)}{\mu_k^2}\log\left(\frac{2}{\delta}\right)\cdot\sum_{\lambda=1}^{m_k-1}\frac{m_k^2}{\lambda^2\left(m_k-\lambda\right)^2}\\
=&128\frac{\sigma^2\left(\mu_k-c\right)}{\mu_k^2}\log\left(\frac{2}{\delta}\right)\cdot\left(\sum_{\lambda=1}^{\left\lfloor\frac{m_k}{2}\right\rfloor}\frac{m_k^2}{\lambda^2\left(m_k-\lambda\right)^2}+\sum_{\lambda=\left\lfloor\frac{m_k}{2}\right\rfloor+1}^{m_k-1}\frac{m_k^2}{\lambda^2\left(m_k-\lambda\right)^2}\right)\\
\leq&128\frac{\sigma^2\left(\mu_k-c\right)}{\mu_k^2}\log\left(\frac{2}{\delta}\right)\cdot\left(\sum_{\lambda=1}^{\left\lfloor\frac{m_k}{2}\right\rfloor}\frac{4}{\lambda^2}+\sum_{\lambda=\left\lfloor\frac{m_k}{2}\right\rfloor+1}^{m_k-1}\frac{4}{\left(m_k-\lambda\right)^2}\right)\\
\leq&128\frac{\sigma^2\left(\mu_k-c\right)}{\mu_k^2}\log\left(\frac{2}{\delta}\right)\cdot 8\cdot\frac{\pi^2}{6}\\
=&\frac{512\pi^2\sigma^2\left(\mu_k-c\right)}{3\mu_k^2}\log\left(\frac{2}{\delta}\right),
\end{aligned}$$

where in the last inequality we use $\sum_{x=1}^{N}\frac{1}{x^2}\leq\frac{\pi^2}{6}$ for all $N\in\mathbb{N}_+$.

As for the regret cause by UEs, we first consider the number of time slots required for $m_{k,t}^l > m_k/2$, with a more detailed analysis than in the proof of sample complexity.

Based on the analysis of the regret caused by IEs, the number of IEs required for $m_{k,t}^l > m_k/2$ can be bounded by:

$$
\begin{aligned}
&128\frac{\sigma^2}{\mu_k^2}\log\left(\frac{2}{\delta}\right)\cdot\sum_{\lambda=1}^{\left\lfloor\frac{m_k}{2}\right\rfloor}\frac{2m_k}{\lambda\left(m_k-\lambda\right)^2\left(m_k-\lambda+1\right)}\\
\leq&128\frac{\sigma^2}{\mu_k^2}\log\left(\frac{2}{\delta}\right)\cdot\frac{2}{m_k}\cdot\sum_{\lambda=1}^{\left\lfloor\frac{m_k}{2}\right\rfloor}\frac{2m_k}{\lambda\left(m_k-\lambda\right)^2}\\
\leq&128\frac{\sigma^2}{\mu_k^2}\log\left(\frac{2}{\delta}\right)\cdot\frac{2}{m_k}\cdot\sum_{\lambda=1}^{m_k-1}\frac{2m_k}{\lambda\left(m_k-\lambda\right)^2}\\
\leq&\frac{1024\pi^2\sigma^2}{3\mu_k^2}\log\left(\frac{2}{\delta}\right)\cdot\frac{1}{m_k}.
\end{aligned}
$$

To ensure that $\hat{V}_{k,t} > \frac{64\sigma^2}{\mu_k^2}\cdot 8\cdot\log\left(\frac{2}{\delta}\right)$, at most $\frac{256\sigma^2}{\mu_k^2}\cdot 8\cdot\log\left(\frac{2}{\delta}\right)\frac{1}{m_k^2}$ additional IEs are required. Note that $\hat{V}_{k,t} > \frac{64\sigma^2}{\mu_k^2}\cdot 8\cdot\log\left(\frac{2}{\delta}\right)$ is a sufficient condition for $\phi(\hat{V}_{k,t},\delta) < \frac{\mu_k}{8\sigma}$.

Next, we consider the time slots required for $m_{k,t}^u < 2m_k$. A sufficient condition for $m_{k,t}^u < 2m_k$ is :

$$
\frac{m_k\mu_k + 2\sigma\phi\left(\hat{\imath}_{k,t},\delta\right)}{\mu_k - 2\sigma\phi(\hat{V}_{k,t},\delta)} < 2m_k.
$$

By rearranging the terms, we get another sufficient condition:

$$
\begin{aligned}
\phi\left(\hat{\imath}_{k,t},\delta\right) &< \frac{m_k\mu_k}{4\sigma},\\
\phi(\hat{V}_{k,t},\delta) &< \frac{\mu_k}{8\sigma}.
\end{aligned}
$$

Solving the above inequality, we obtain that:

$$
\begin{aligned}
\hat{\imath}_{k,t} &> \frac{16\sigma^2}{m_k^2\mu_k^2}\cdot 8\cdot\log\left(\frac{2}{\delta}\right),\\
\hat{V}_{k,t} &> \frac{64\sigma^2}{\mu_k^2}\cdot 8\cdot\log\left(\frac{2}{\delta}\right).
\end{aligned}
$$

Note that the number of required IEs exceeds the number of required UEs for $m_{k,t}^u < 2m_k$.

Consequently, we find that after at most $\frac{1024\pi^2\sigma^2}{3\mu_k^2}\log\left(\frac{2}{\delta}\right)\cdot\frac{1}{m_k} + \frac{256\sigma^2}{\mu_k^2}\cdot 8\cdot\log\left(\frac{2}{\delta}\right)\frac{1}{m_k^2}$ alternating IEs and UEs, $m_{k,t}^u < 2m_k$ and $m_{k,t}^l > m_k/2$. The regret of these UEs can be bounded by $(N-m_k)\,c$. Although these UEs may incur significant costs, their number is limited, as it is inversely related to the capacity $m_k$.

Next, we consider the number of time slots required for $m_{k,t}^u \leq m_k+\lambda-1$, conditioned on $m_{k,t}^u \leq m_k+\lambda$ for $\lambda \leq m_k-1$:

$$
\frac{m_k\mu_k + 2\sigma\phi\left(\hat{\imath}_{k,t},\delta\right)}{\mu_k - 2\sigma\phi(\hat{V}_{k,t},\delta)} < m_k+\lambda.
$$

By rearranging the terms, we obtain that:

$$2\sigma\phi\left(\hat{\iota}_{k,t},\delta\right)+2\left(m_k+\lambda\right)\sigma\phi(\hat{V}_{k,t},\delta)<\lambda\mu_k.$$

A sufficient condition for the inequality above is that:

$$\begin{aligned}\phi\left(\hat{\iota}_{k,t},\delta\right)&<\frac{\mu_k}{4\sigma}\cdot\lambda,\\ \phi(\hat{V}_{k,t},\delta)&<\frac{\mu_k}{4\sigma}\cdot\frac{\lambda}{m_k+\lambda}.\end{aligned}$$

By solving the inequalities above, we obtain that:

$$\begin{aligned}\hat{\iota}_{k,t}&>\frac{\left(4\sigma\right)^2}{\mu_k^2\lambda^2}\cdot 8\cdot\log\left(\frac{2}{\delta}\right),\\ \hat{V}_{k,t}&>\frac{\left(4\sigma\right)^2}{\mu_k^2\lambda^2}\cdot 8\cdot\log\left(\frac{2}{\delta}\right)\left(m_k+\lambda\right)^2.\end{aligned}$$

Let $g_2\left(\lambda\right):=\frac{\left(4\sigma\right)^2}{\mu_k^2\lambda^2}\cdot 8\cdot\log\left(\frac{2}{\delta}\right)\left(m_k+\lambda\right)^2$. If $m_{k,t}^u\le m_k+\lambda$, noting that $m_{k,t}^l>m_k/2$, we can upper-bound the number of additional IEs as $\left(g_2\left(\lambda\right)-g_2\left(\lambda+1\right)\right)\cdot\frac{4}{m_k^2}$, which can be further bounded by:

$$\begin{aligned}&\left(g_2\left(\lambda\right)-g_2\left(\lambda+1\right)\right)\cdot\frac{4}{m_k^2}\\ \le&\frac{\left(4\sigma\right)^2}{\mu_k^2}\cdot 8\cdot\log\left(\frac{2}{\delta}\right)\cdot 8\left(\frac{1}{\lambda}+\frac{1}{m_k}\right)\frac{1}{\lambda\left(\lambda+1\right)}\\ \le&\frac{\left(4\sigma\right)^2}{\mu_k^2}\cdot 8\cdot\log\left(\frac{2}{\delta}\right)\cdot 16\frac{1}{\lambda^2\left(\lambda+1\right)}.\end{aligned}$$

Compared to the number of additional UEs required for $\hat{\iota}_{k,t}>\frac{\left(4\sigma\right)^2}{\mu_k^2\lambda^2}\cdot 8\cdot\log\left(\frac{2}{\delta}\right)$, the number of additional IEs is greater. Noting that the regret caused by a UE is at most $\lambda c$ given that $m_{k,t}^u\le m_k+\lambda$, we can upper-bound the regret caused by UEs when $m_{k,t}^u<2m_k$ as:

$$\begin{aligned}&2048\frac{c\sigma^2}{\mu_k^2}\log\left(\frac{2}{\delta}\right)\sum_{\lambda=1}^{m_k-1}\frac{1}{\lambda\left(\lambda+1\right)}\\ =&2048\frac{c\sigma^2}{\mu_k^2}\log\left(\frac{2}{\delta}\right)\left(\sum_{\lambda=1}^{m_k-1}\frac{1}{\lambda}-\frac{1}{\lambda+1}\right)\\ \le&2048\frac{c\sigma^2}{\mu_k^2}\log\left(\frac{2}{\delta}\right).\end{aligned}$$

Setting $\delta = 2/T$ and summing up the regret together, we can upper-bound the regret of exploring one arm as:

$$\begin{aligned}
\mathbb{E}[Reg_k\left(T\right)] \leq & 2048\frac{c\sigma^2}{\mu_k^2}\log\left(T\right) + \frac{512\pi^2\sigma^2\left(\mu_k - c\right)}{3\mu_k^2}\log\left(T\right) \\
& + \left(\frac{1024\pi^2\sigma^2}{3\mu_k^2}\log\left(T\right)\cdot\frac{1}{m_k} + \frac{256\sigma^2}{\mu_k^2}\cdot 8\cdot\log\left(T\right)\frac{1}{m_k^2}\right)\left(Nc\right) \\
& + 2\max\left(Nc, m_k\mu_k\right) \\
= & O\left(\frac{Nc\sigma^2 + \left(\mu_k - c\right)\sigma^2}{\mu_k^2}\log\left(T\right)\right).
\end{aligned}$$

■

**Proof** (Theorem 14)

We can recalculate $Reg\left(T\right)$ as the sum of the regrets on each arm individually:

$$\begin{aligned}
&Reg\left(T\right)\\
=&\sum_{t=1}^{T}\left(f\left(\mathbf{a}^{*}\right)-f\left(\mathbf{a}_{t}\right)\right)\\
=&\sum_{t=1}^{T}\left(\left(\sum_{k=1}^{K}\left(m_{k}\mu_{k}-cm_{k}\right)\right)-\left(\sum_{k=1}^{K}\left(\min\{a_{k,t},m_{k}\}\cdot\mu_{k}-c\cdot a_{k,t}\right)\right)\right)\\
=&\sum_{t=1}^{T}\left(\sum_{k=1}^{K}\left(m_{k}\mu_{k}-cm_{k}-\min\left\{a_{k,t},m_{k}\right\}\cdot\mu_{k}+c\cdot a_{k,t}\right)\right)\\
=&\sum_{k=1}^{K}\left(\sum_{t=1}^{T}\left(m_{k}\mu_{k}-cm_{k}-\min\left\{a_{k,t},m_{k}\right\}\cdot\mu_{k}+c\cdot a_{k,t}\right)\right)\\
=&\sum_{k=1}^{K}Reg_{k}\left(T\right),
\end{aligned}$$

where $Reg_{k}\left(T\right):=\sum_{t=1}^{T}\left(m_{k}\mu_{k}-cm_{k}-\min\left\{a_{k,t},m_{k}\right\}\cdot\mu_{k}+c\cdot a_{k,t}\right).$

Unlike $Reg_k$ in the sample regret, the regret on the arm $k$ in real MP-MAB setting involves compulsory IEs due to the limited number of plays. As a result, summing the sample regret upper bounds over all $K$ arms may not provide a reasonable upper bound for the regret in the real MP-MAB setting.

However, a similar approach can be applied to the partition of the regret $Reg_k$ based on the event $A$. The expectation of $Reg_k\left(T\right)$ can be separated by the event $A$:

$$\begin{aligned}
&\mathbb{E}\left[Reg_{k}\left(T\right)\right]\\
=&\mathbb{E}\left[Reg_{k}\left(T\right)\mathbb{1}\{\,A\,\}\right]+\mathbb{E}\left[Reg_{k}\left(T\right)\mathbb{1}\{\,A^{C}\,\}\right]\\
\leq&\mathbb{E}\left[Reg_{k}\left(T\right)\mathbb{1}\{\,A\,\}\right]+\mathbb{P}\left(A^{C}\right)\max\left(\mathbb{E}\left[Reg_{k}\left(T\right)\right]\right),
\end{aligned}$$

where the second term can be upper-bounded by $\left(K\delta\right)T\cdot Regmax_{k}$. We will bound the first term by analyzing the convergence of the confidence intervals of $m_k$.

It is shown in the proof of sample regret that the number of IEs and UEs should be balanced when studying the capacity. Applying too many UEs in the early time slots on an arm can be costly and hinder the progress of learning the capacities of other arms. Similarly, too many IEs in the first few time slots can result in significant regret.

Consider the alternating IE and UE strategy in which the optimal sample regret is achieved. If extra IEs are inserted in the the learning process on the arm $k$, the convergence of the confidence upper bound of $m_{k,t}^{u}$ will actually be accelerated at the time slots when UEs are applied on the arm $k$. In other words, when focusing solely on the regret caused by the UEs, the additional IEs do not lead to an increase in this portion of the regret. Therefore, the regret caused by the UEs on arm $k$ can be bounded in the same way as shown in the sample regret. The main challenge lies in limiting the number of costly IEs on the arms, which can be frequent in our setting, as arms are often required to be played with IEs due to a lack of plays.

Since we expect that the number of UEs should be fewer than the number of IEs on the same arm, and balancing the two is crucial for achieving the optimal sample regret, the number of UEs becomes more decisive when learning the capacity. Consequently, it is natural to apply UEs to arms whose capacities are not well learned. Given that the regret caused by UEs can be bounded similarly to the sample regret, we now aim to use the regret caused by a single IE on a arm as a criterion for evaluating the learning progress of that arm.

The regret caused by a single IE on the arm $k$ can be upper-bounded as :

$$\begin{aligned}
&\left(m_k - \frac{\hat{\upsilon}_{k,t} - \sigma\phi\left(\hat{\iota}_{k,t},\delta\right)}{\hat{\mu}_{k,t} + \sigma\phi(\hat{V}_{k,t},\delta)}\right)\mu_k \\
\leq &\left(m_k - \frac{\hat{\upsilon}_{k,t} - \sigma\phi\left(\hat{\iota}_{k,t},\delta\right)}{\hat{\mu}_{k,t} + \sigma\phi(\hat{V}_{k,t},\delta)}\right)\left(\hat{\mu}_{k,t} + \sigma\phi(\hat{V}_{k,t},\delta)\right) \\
= &m_k\hat{\mu}_{k,t} + \sigma m_k\phi(\hat{V}_{k,t},\delta) - \hat{\upsilon}_{k,t} + \sigma\phi\left(\hat{\iota}_{k,t},\delta\right) \\
\leq &m_k\mu_k + 2\sigma m_k\phi(\hat{V}_{k,t},\delta) - m_k\mu_k + 2\sigma\phi\left(\hat{\iota}_{k,t},\delta\right) \\
= &2\sigma m_k\phi(\hat{V}_{k,t},\delta) + 2\sigma\phi\left(\hat{\iota}_{k,t},\delta\right).
\end{aligned}$$

This regret upper bound further demonstrates that when the numbers of UEs and IEs are balanced, the regret caused by a single IE will decrease rapidly. Since we have no knowledge of the true capacity $m_k$, we can use $\phi(\hat{V}_{k,t},\delta) + \phi\left(\hat{\iota}_{k,t},\delta\right)$ as an alternative criterion. Both criteria serve the same purpose of balancing the numbers of UEs and IEs on a particular arm. Additionally, $\phi(\hat{V}_{k,t},\delta)$ and $\phi\left(\hat{\iota}_{k,t},\delta\right)$ measure the width of the confidence intervals of $\hat{\mu}_{k,t}$'s and $\hat{\upsilon}_{k,t}$, reflecting the extent of capacity learning on the arm $k$. Therefore, besides requiring that an arm is not played with UEs in two consecutive time slots, we also require that UEs be applied first to arms with greater $\phi(\hat{V}_{k,t},\delta) + \phi\left(\hat{\iota}_{k,t},\delta\right)$, provided these arms are not forced to be played with an IE based on the first condition.

Note that $\phi(\hat{V}_{k,t},\delta) \leq \phi\left(\hat{\iota}_{k,t},\delta\right)$ for $t \geq K+1$. It can be observed that for any two arms $i, j$, a sufficient condition for $\phi(\hat{V}_{i,t},\delta) + \phi(\hat{\iota}_{i,t},\delta) \geq \phi(\hat{V}_{j,t},\delta) + \phi(\hat{\iota}_{j,t},\delta)$ is $\phi\left(\hat{\iota}_{i,t},\delta\right) \geq 2\phi\left(\hat{\iota}_{j,t},\delta\right)$. Solving the inequality above, we get a sufficient condition as $8\hat{\iota}_{i,t} \leq \hat{\iota}_{j,t}$. This implies that during the learning process of the algorithm, no arm is assigned with more than eight times as many UEs as any other arm at any time slot $t$. From this, we directly obtain the following lemma.

**Lemma 20** *For arbitrary arm $k$ and arbitrary positive integer $\lambda$, a sufficient condition for having at least $\lambda$ UEs on arm $k$ is that*

$$t \geq 8\lambda K.$$

According to the result in sample regret, at most additional $128\frac{\sigma^2}{\mu_k^2}\log\left(\frac{2}{\delta}\right)\cdot 2\frac{m_k}{(m_k-\lambda)^2(m_k-\lambda+1)\lambda}$ UEs are required for $m^l_{k,t} \geq \lambda + 1$, conditioned on $m^l_{k,t} \geq \lambda$. Let $x_i$ denote the number of IEs applied on arm $k$ when $m^l_{k,t} = i$:

$$x_i := \sum_{t=1}^{T} \mathbb{1}\left(a_{k,t} = i\right).$$

For any integer $\lambda \in [1, m_k - 1]$, the number of UEs required for $m_{k,t}^{l} \geq \lambda + 1$ is at most $\sum_{i=1}^{\lambda} 128\frac{\sigma^2}{\mu_k^2}\log\left(\frac{2}{\delta}\right) \cdot 2\frac{m_k}{(m_k-i)^2(m_k-i+1)i}$. Then, according to Lemma 20, the number of IEs on arm $k$ is at most:

$$\sum_{i=1}^{\lambda} 128\frac{\sigma^2}{\mu_k^2}\log\left(\frac{2}{\delta}\right) \cdot 2\frac{m_k}{\left(m_k - i\right)^2\left(m_k - i + 1\right)i} \cdot 8K.$$

So we have the following conditions on $x_i$: for all integer $\lambda \in [1, m_k - 1]$:

$$\sum_{i=1}^{\lambda} x_i \leq \sum_{i=1}^{\lambda} 128\frac{\sigma^2}{\mu_k^2}\log\left(\frac{2}{\delta}\right) \cdot 2\frac{m_k}{\left(m_k - i\right)^2\left(m_k - i + 1\right)i} \cdot 8K. \tag{8}$$

The regret caused by these IEs can be expressed as:

$$\sum_{i=1}^{m_k-1} x_i\left(m_k - i\right)\left(\mu_k - c\right). \tag{9}$$

It is evident that the maximum value of the expression (9) is achieved when the inequalities (8) hold with equalities for all integer $\lambda \in [1, m_k - 1]$. Consequently, the summation of the expression (9) can be bounded using the same method demonstrated in the sample regret analysis, as follows:

$$\frac{512\pi^2\sigma^2\left(\mu_k - c\right)}{3\mu_k^2}\log\left(\frac{2}{\delta}\right) \cdot 8K.$$

Setting $\delta = 2/T$ and noting that the regrets caused by UEs can be bound in the same way as the sample regret, we derive the final form of the regret upper bound on the arm $k$ as:

$$\begin{aligned}\mathbb{E}[Reg_k\left(T\right)] \leq & 2048\frac{c\sigma^2}{\mu_k^2}\log\left(T\right) + \frac{512\pi^2\sigma^2\left(\mu_k - c\right)}{3\mu_k^2}\log\left(T\right) \cdot 8K \\ & + \left(\frac{1024\pi^2\sigma^2}{3\mu_k^2}\log\left(T\right) \cdot \frac{1}{m_k} + \frac{256\sigma^2}{\mu_k^2} \cdot 8 \cdot \log\left(T\right)\frac{1}{m_k^2}\right)\left(Nc\right) \\ & + 2K\max\left(Nc, m_k\mu_k\right) \\ = & O\left(\frac{\frac{N}{m_k}c\sigma^2 + K\left(\mu_k - c\right)\sigma^2}{\mu_k^2}\log\left(T\right)\right)\end{aligned}$$

Summing up the inequalities above for all $k$, we can upper-bound the regret of the Algorithm 2 as:

$$\begin{aligned}
Reg\left(T\right) \leq & \sum_{k=1}^{K}\left(2048\frac{c\sigma^2}{\mu_k^2}\log\left(T\right)+\frac{512\pi^2\sigma^2\left(\mu_k-c\right)}{3\mu_k^2}\log\left(T\right)\cdot 8K\right.\\
& +\left(\frac{1024\pi^2\sigma^2}{3\mu_k^2}\log\left(T\right)\cdot\frac{1}{m_k}+\frac{256\sigma^2}{\mu_k^2}\cdot 8\cdot\log\left(T\right)\frac{1}{m_k^2}\right)\left(Nc\right)\\
& \left.+\,2K\max\left(Nc,m_k\mu_k\right)\right)\\
= & O\left(\sum_{k=1}^{K}\left(\frac{N}{m_k}c+K\left(\mu_k-c\right)\right)\frac{\sigma^2}{\mu_k^2}\log\left(T\right)\right)
\end{aligned}$$

■

**Proof** (Theorem 16)

According to Theorem 14, for arbitrary $\Delta > 0$, we have:

$$
\begin{aligned}
Reg\left(T\right) \leq & \sum_{k=1}^{K} \left( 2048 \frac{c\sigma^2}{\mu_k^2} \log\left(T\right) + \frac{512\pi^2\sigma^2\left(\mu_k - c\right)}{3\mu_k^2} \log\left(T\right) \cdot 8K \right. \\
& + \left( \frac{1024\pi^2\sigma^2}{3\mu_k^2} \log\left(T\right) \cdot \frac{1}{m_k} + \frac{256\sigma^2}{\mu_k^2} \cdot 8 \cdot \log\left(T\right) \frac{1}{m_k^2} \right) \left(Nc\right) \\
& \left. + 2 \max\left(Nc, m_k\mu_k\right) \right) \\
\leq & \sum_{m_k\mu_k \geq \delta}^{K} \left( 2048 \frac{c\sigma^2}{\mu_k^2} \log\left(T\right) + \frac{512\pi^2\sigma^2\left(\mu_k - c\right)}{3\mu_k^2} \log\left(T\right) \cdot 8K \right. \\
& + \left( \frac{1024\pi^2\sigma^2}{3\mu_k^2} \log\left(T\right) \cdot \frac{1}{m_k} + \frac{256\sigma^2}{\mu_k^2} \cdot 8 \cdot \log\left(T\right) \frac{1}{m_k^2} \right) \left(Nc\right) \\
& \left. + 2K \max\left(Nc, m_k\mu_k\right) \right) \\
& + \sum_{m_k\mu_k < \delta}^{K} \left( T(\mu_k - c) m_k \right) \\
\leq & \sigma^2 \sum_{k=1}^{K} 2048 \cdot \frac{m_k}{\Delta} \log(T) + K\sigma^2 \sum_{k=1}^{K} \frac{4096\pi^2}{3} \cdot \frac{m_k}{\Delta} \log(T) \\
& + N\sigma^2 \sum_{k=1}^{K} \frac{1024\pi^2}{3} \cdot \frac{1}{\Delta} \log(T) + N\sigma^2 \sum_{k=1}^{K} \frac{2048}{m_k} \cdot \frac{1}{\Delta} \log(T) \\
& + \sum_{k=1}^{K} T\Delta + \sum_{k=1}^{K} 2K \max\left(\mu_k m_k, Nc\right) \\
\leq & \sigma^2 \frac{\log(T)}{\Delta} \left( 2048M + \frac{4096\pi^2}{3} KM + \frac{1024\pi^2}{3} NK + 2048NK \right) \\
& + TK\Delta + \sum_{k=1}^{K} 2K \max\left(\mu_k m_k, Nc\right) \\
\leq & \sigma \sqrt{\left( 2048M + \frac{4096\pi^2}{3} KM + \frac{1024\pi^2}{3} NK + 2048NK \right) K \left(T \log\left(T\right)\right)} + \sum_{k=1}^{K} 2K \max\left(\mu_k m_k, Nc\right)
\end{aligned}
$$

The final step is letting $\Delta = \sigma \sqrt{\frac{\left(2048M + \frac{4096\pi^2}{3} KM + \frac{1024\pi^2}{3} NK + 2048NK\right)}{TK} \log(T)}$. ■