Finite Budget Analysis of Multi-armed Bandit Problems

Yingce Xia¹, Tao Qin², Wenkui Ding³, Haifang Li², Xu-Dong Zhang³, Nenghai Yu¹, Tie-Yan Liu²

¹University of Science and Technology of China  ²Microsoft Research  ³Tsinghua University  ⁴University of Chinese Academic of Sciences

¹yingce.xia@gmail.com, ²ynh@ustc.edu.cn, ³taoqin.tie-yan.liu@microsoft.com, ⁴dingwenkui@gmail.com, ⁵zhangxd@tsinghua.edu.cn; ⁶lihaifang@amss.ac.cn

Abstract

In the budgeted multi-armed bandit (MAB) problem, a player receives a random reward and needs to pay a cost after pulling an arm, and he cannot pull any more arm after running out of budget. In this paper, we give an extensive study of the upper confidence bound based algorithms and a greedy algorithm for budgeted MABs. We perform theoretical analysis on the proposed algorithms, and show that they all enjoy sublinear regret bounds with respect to the budget $B$. Furthermore, by carefully choosing the parameters, they can even achieve log linear regret bounds. We also prove that the asymptotic lower bound for budgeted Bernoulli bandits is $\Omega(\ln B)$. Our proof technique can be used to improve the theoretical results for fractional KUBE [27] and Budgeted Thompson Sampling [31].

Keywords: Budgeted Multi-Armed Bandits, UCB Algorithms, Regret Analysis

1. Introduction

Multi-armed bandits (MAB) correspond to a typical sequential decision problem, in which a player receives a random reward by playing one of $K$ arms from a slot machine at each round and wants to maximize his cumulated reward. Many real world applications can be modeled as MAB problems, such as auction mechanism design [25], search advertising [28], UGC mechanism design [18] and personalized recommendation [23]. Many algorithms have been designed for MAB problems and studied from both theoretical and empirical perspectives, like UCB1, $\epsilon$-GREEDY [6], UCB-V [4], LinRel [5], DMED [19], and KL-UCB [17]. A survey on MAB can be found in [12].

Most of the aforementioned works assume that playing an arm is costless, however, in many real applications including the real-time bidding problem in ad exchange [13], the bid optimization problem in sponsored search [10], the on-spot instance bidding problem in Amazon EC2 [9], and the cloud service provider selection problem in IaaS [2], one needs to pay some cost to play an arm and the number of plays is constrained by a budget. To model these applications, a new kind of MAB problems, called budgeted MAB, have been proposed and studied in recent years, in which the play of an arm is associated with both a random reward and a cost. According to different settings of budgeted MAB, the cost could be either deterministic or random, either discrete or continuous.

In the literature, a few algorithms have been developed for some particular settings of the budgeted MAB problem. The setting of deterministic cost was studied in [27], and two algorithms named KUBE and fractional KUBE were proposed, which learn the probability of pulling an arm by solving an integer programming problem. It has been proven that these two algorithms can lead to a regret bound of $O(\ln B)$. The setting of random discrete cost was studied in [14], and two upper confidence bound (UCB) based algorithms with specifically designed (complex) indexes were proposed and the log linear regret bounds were derived for the two algorithms.

The above algorithms for the budgeted MAB problem could only address some (but not all) settings of budgeted MAB. Given the tight connection between budgeted MAB and standard MAB problems, an interesting question to ask is whether some extensions of the algorithms originally designed for standard MAB (without budget constraint) could be good enough to fulfill the budgeted MAB tasks, and perhaps in a more general way. [31] shows that a simple extension of the Thompson sampling algorithm, which is designed for standard MAB, works quite well for budgeted MAB with very general settings. Inspired by that work, we are interested in whether we can handle budgeted MAB by extending other algorithms designed for standard MAB.

In order to answer the question, we study the following natural extensions of the UCB and $\epsilon$-GREEDY algorithms [6] in this paper (these extensions do not need to know the budget $B$ in advance). We first propose four basic algorithms for budgeted MABs: (1) i-UCB, which replaces the average reward of an arm in the exploitation term of UCB1 by the average reward-to-cost ratio; (2) c-UCB, which further incorporates the average cost of an arm into the exploration term of UCB1 (c-UCB can be regarded as an adaptive version of i-UCB); (3) m-UCB, which mixes the upper confidence bound of reward and the lower confidence bound of cost; (4) b-GREEDY, which replaces the average reward in $\epsilon$-GREEDY with the average reward-to-cost ratio.

We conduct theoretical analysis on these algorithms, and show that they all enjoy sublinear regret bounds with respect to $B$. By carefully setting the hyper parameter in each algorithm, we show that the regret bounds can be further improved to be log linear. Although the basic idea of regret analysis for the
algorithms for budgeted MAB is similar to standard MAB, i.e., we only need to bound the number of plays of all the suboptimal arms (the arms whose expected-reward-to-expected-cost ratios are not the maximum), there are two challenges to address comparing with the standard MAB (without budgets). First, there are two random factors, rewards and costs, that can influence the choice of arm simultaneously, which will bring difficulties when decomposing the probabilities that suboptimal arms are pulled. Second, the stopping time (i.e., the time/round that a pulling algorithm runs out of budget and stops) is a random variable related to the costs of each arm, due to which the independence of the costs at different rounds is destroyed and it brings difficulties when applying concentration inequalities.

To address the first challenge, we introduce the δ-gap (5), with which we can separate the terms w.r.t the ratio of rewards and costs into terms related to rewards only and costs only. To address the second one, we make a decomposition like (8): before round \(2B/\mu_{\min}^1\) (\(\mu_{\min}\) is the minimum expected cost among all the arms), we only consider the event about pulling a suboptimal arm; after round \(2B/\mu_{\min}^1\), we only consider the event that there is remaining budget at each round. Combining the two techniques with those for standard MAB, we can derive the regret bounds for the aforementioned four algorithms.

Furthermore, we give a lower bound to the budgeted Bernoulli MAB (the reward and cost of a budgeted Bernoulli bandit is either 0 or 1), and show that our proposed algorithms can match the lower bound (by carefully setting the hyper parameters).

In addition to the theoretical analysis, we also conduct empirical study on these algorithms. We simulate two bandits, one with 10 arms and the other one with 50 arms. For each bandit, we consider two sub cases with different distributions for rewards and costs: the Bernoulli distribution in the first sub case and the beta distribution in the second one. The simulation results show that our extensions work surprisingly well for both the bandits and the sub cases, and can achieve comparable or better performances than existing algorithms.

The Budget-UCB algorithm proposed in [30] can be seen as a combined version of c-UCB and m-UCB. With our proposed δ-gap, the regret of Budget-UCB can be improved.

Besides the literature mentioned above, there also exist some works studying the MAB problems with multiple budget constraints. For example, in [7], the bandits with knapsacks setting (shortly, BwK) is studied: the total number of plays is constrained by a predefined number \(T\) and the total cost of the plays is constrained by a monetary budget \(B\) (the problem). [7] studies the distribution-free regret bound for BwK while [16] provides distribution-dependent bound. In [1], the above setting is extended to arms with concave rewards. Furthermore, in [8] contextual bands are considered. At the first glance, these settings seem to be more general than the budgeted MAB problem defined in the previous subsection. We would like to point out that their algorithms and theoretical analysis cannot be directly applied here, because in their settings the total number of plays \((T)\) is critical to their algorithms and analysis, however, such a number \(T\) does not exist in the setting under our investigation.

2. Problem Setup

A budgeted MAB problem can be described as follows. A player is facing a slot machine with \(K\) arms \((K \geq 2)\). At time/round \(t\), he pulls an arm \(i \in [K]\), (for ease of reference, let \([K]\) denote the set \(\{1, 2, \cdots, K\}\)) receives a random reward \(r_i(t)\), and pays a cost \(c_i(t)\) until his budget, \(B\), runs out. Both the rewards and costs are supported in \([0, 1]\). There can be different settings depending on the values of the costs, which could be either deterministic or random, either discrete or continuous. In this work, we mainly focus on the setting with random costs, since deterministic costs can be regarded as a special case of random ones. When analyzing the regret bounds of our proposed algorithms, the costs could be either discrete or continuous. Following the common practice in standard MAB, we assume the independence between arms and rounds: the rewards and costs of an arm are independent of any other arm, and the rewards (and costs) of arm \(i\) at different round \(t\) are independently drawn from the same distribution with expectation \(\mu_i^f\) (and \(\mu_i^c\)). We do not need to assume that the rewards of an arm are independent of its costs. Without loss of generality, we assume \(0 < \mu_i^f, \mu_i^c < 1\) for any \(i \in [K]\) and \(\mu_i^f > \mu_i^c\) for any \(i \neq 1\). We name arm 1 as the optimal arm and the others as suboptimal arms. The goal of the player is to maximize his total expected reward before the budget runs out.

We introduce some notations that will be frequently used throughout the text. Let \(I_i\) denote the arm chosen at round \(t\). \(1\{\cdot\}\) is the indicator function. Let \(r_i(t)\) and \(c_i(t)\) denote the reward and cost of arm \(i \in [K]\) at round \(t\). Note \(r_i(t)\) and \(c_i(t)\) are visible to the player only if arm \(i\) is pulled at round \(t\). Let \(n_{i,t}, \tau_{i,t}\) and \(\bar{c}_{i,t}\) denote the number of pulling time, the average reward, and the average cost of arm \(i\) before (excluding) round \(t\) respectively. Mathematically, they are

\[
\tau_{i,t} = \frac{1}{n_{i,t}} \sum_{s=1}^{t-1} r_i(s) 1[I_s = i], \quad \bar{c}_{i,t} = \frac{1}{n_{i,t}} \sum_{s=1}^{t-1} c_i(s) 1[I_s = i].
\]

We often use pseudo regret to evaluate an algorithm (denoted as \(R^i\) for algorithm “a”), which is defined as follows:

\[
R^i = R^i - \mathbb{E} \sum_{s=1}^{t} r_i(s) 1[B_s \geq 0],
\]

where \(R^i\) is the expected reward of the optimal policy (the policy that can obtain the maximum expected reward when the reward distribution and the cost distribution of each arm is known in advance), \(B_s\) is the remaining budget at round \(s\), (i.e., \(B_t = B - \sum_{s=1}^{t} r_i(s)\)), and the expectation is taken with respect to the randomness of the algorithm, the received rewards/costs, and the stopping time. Maximizing the expected reward is equivalent to minimizing the pseudo regret.

Please note that it could be very complex to obtain the optimal policy for the budgeted MAB problem (under the condition that the reward and cost distributions of each arm are known). Even for its degenerated case, where the reward and cost of each arm are deterministic, the problem is known to be NP-hard (actually in this case the problem becomes an unbounded knapsack problem [24]). Therefore, generally speaking, it is hard to calculate \(R^i\) in an exact manner.
However, we find that it is much easier to approximate the optimal policy and to upper bound \( R^* \). Specifically, when the reward and cost per pulling are supported in \([0, 1]\) and \( B \) is large, always pulling the optimal arm could be very close to the optimal policy\(^1\). The results are summarized in Lemma 1, together with upper bounds on \( R^* \). The proof of Lemma 1 can be found in the full version of [31].

**Lemma 1.** When the reward and cost per pulling are supported in \([0, 1]\), we have \( R^* = (\bar{\mu}_i^{c}/\mu_i^{c})(B+1) \), and the suboptimality of always pulling arm 1 (as compared to the optimal policy) is at most \( 2\bar{\mu}_1^{c}/\mu_1^{c} \).

### 3. Algorithms

We present two kinds of algorithms for budgeted multi-armed bandit problems. As pointed in Section 2, although the optimal policy is quite complex for budgeted MAB, always pulling the optimal arm can bring almost the same expected reward as the optimal policy. Therefore, our proposed algorithms target at pulling the optimal arm as frequently as possible, with some tradeoff between exploration (on the less pulled arms) and exploitation (on the empirical best arms).

#### 3.1. Extensions of UCB Based Algorithm

UCB algorithms [6, 4, 33] are widely used and well studied for standard MAB. The main idea of UCB is to play the arm with the largest index, where the index of an arm is the sum of two terms: an exploitation term (the average reward \( \bar{r}_{i,t} \) of the arm), and an exploration term (the uncertainty of the average reward as an estimator of the true expected reward).

When we have a cost for each arm and a budget constraint, it is straightforward to replace the average reward in the exploitation term of UCB1 by the average reward-to-cost ratio, (i.e., \( \bar{r}_{i,t}/c_{i,t} \)). By doing so, we can get the so-called i-UCB algorithm, in which the exploration term is defined as

\[
\bar{E}_{i,t} = \alpha \sqrt{\ln(t-1)/n_{i,t}},
\]

where \( \alpha \) is a positive hyper parameter. Please note that such a hyper parameter also exists in [23], which can make the algorithm flexible. Here, “i” stands for independence, since the exploration term is independent to the average reward and cost. The index\(^2\) for i-UCB is the \( D_{i,t} \) in (4).

Note that the exploration term of i-UCB only considers the pulling time of an arm (i.e., a less frequently pulled arm is more likely to be explored), and does not take costs into consideration. By considering the average cost of an arm in exploration, we get another algorithm called c-UCB. Here, “c” stands for “cost”. In c-UCB, an arm that was less frequently pulled and/or has smaller costs is more likely to be explored. The index for c-UCB is the \( D_{i,t} \) in (4).

Furthermore, by mixing the exploitation and exploration terms together, we can get the m-UCB algorithm, where “m” stands for mixed. As can be seen, m-UCB plays arms according to the ratio of the upper confidence bound of reward to the lower confidence bound of cost. Here the upper confidence bound of reward is capped by 1 and the lower confidence bound of cost is floored by zero, so as to ensure that they are supported in \([0, 1]\). The index for m-UCB is the \( D_{i,t} \) in (4).

\[
D_{i,t} = \bar{r}_{i,t}/\bar{c}_{i,t} + \bar{E}_{i,t}, \quad D_{i,t} = \bar{r}_{i,t}/\bar{c}_{i,t} + E_{i,t}, \quad D_{i,t} = \min[\bar{r}_{i,t} + E_{i,t}, 1]/\max[\bar{c}_{i,t} - E_{i,t}, 0].
\]

The details of UCB based algorithms are summarized in Algorithm 1.

#### 3.2. Extension of a Greedy Algorithm

The \( \epsilon \)-GREEDY algorithm [6] is a simple and well known algorithm for standard MAB, which plays with probability 1 - \( \epsilon \) the arm with the largest average reward, and with probability \( \epsilon \) a randomly chosen arm. It can be easily extended to budgeted MAB by replacing the average reward of each arm by its average reward-to-cost ratio. For ease of reference, we call the extended algorithm b-GREEDY. Algorithm 2 shows the details of the b-GREEDY algorithm.

#### 4. Regret Analysis

##### 4.1. Upper Bounds of the Regrets

For any suboptimal arm \( i \geq 2 \), the weighted ratio gap \( \Delta_i \), asymmetric \( \delta \)-gap \( \delta(\gamma) \) with \( \gamma \geq 0 \) and the symmetric \( \varphi \)-gap...
\( \psi \) play important roles through this work, which are defined as follows:

\[
\Delta_i = \mu_i^* - \mu_i; \quad \delta_i(y) = \frac{\Delta_i}{\sqrt{\theta_i/|\psi_i| + 1}}; \quad \phi_i = \frac{\mu_i^* \Delta_i}{\mu_i^* + \mu_i + \mu_i^*'.}
\]

(5)

One can verify that

\[
\frac{\mu_i^*}{\mu_i} = \frac{\mu_i^* + \delta_i(y)}{\phi_i} - \frac{\mu_i^* - \phi_i}{\phi_i + \phi_i}. \tag{6}
\]

(6) shows two kinds of gaps between arm \( i \) and arm \( i' \): (i) to make a suboptimal arm \( i \) an optimal arm, we can increase the \( \mu_i^* \) by \( \delta_i(y) \), while decreasing \( \mu_i^* \) by \( \gamma \delta_i(y) \). When \( \gamma = 1 \), the asymmetric gap becomes symmetric. (ii) To make arm 1 and a suboptimal arm \( i \) the same expected reward to expected cost ratio, we can increase the \( \mu_i^* \) while decreasing the \( \mu_i^* \). We will see that these gaps are highly correlated to the regret bounds of the algorithms later.

To present the regret bounds of the proposed algorithms, we need to introduce some other notations: (1) \( \tau_B = [2\beta/|\psi_{\text{min}}|^2] \), where \( |\psi_{\text{min}}| = \min_{i \in [K]} |\psi_i|; \) (2) define four indices as follows:

\[
\begin{align*}
oc_i &= \frac{1}{2}(\mu_i^* - |\psi_i|^2)/(\mu_i^* + 1)^2; \\
oco_i &= (\mu_i^* + \mu_i + \mu_i^*')^2; \\
\alpha_o &= \mu_i^* + 1; \\
\alpha_i &= \min_{i \in [K]} \{\alpha_i, \alpha_o\}.
\end{align*}
\]

The subscripts "i.e.,m," stand for "i-UCB, c-UCB, m-UCB, and b-GREEDY" algorithms respectively. We will use them throughout this work when the context is clear. The regret bounds of the proposed algorithms are summarized as follows:

**Theorem 2.** For any algorithm \( a \in \{i, c, m, b\} \):

(I) If \( \alpha_o > 0 \), the regret \( R^a \) can be upper bounded as follows:

\[
R^a \leq \sum_{t=1}^{T} \ell_t \ln(\tau_B + O(1)),
\]

where \( \ell_t = (4r^2(\mu_i^*')^2)/\Delta_i, \ell_{i,j} = (\mu_i^* + \alpha + 1)^2/\Delta_i, \ell_{o,i} = (\mu_i^* + 1)^2/\Delta_i, \ell_{o,i} = (\mu_i^* + 1)^2/\Delta_i, \) and \( O(1) \) is only related to \( \mu_i^*, \mu_i \) for any \( i \in [K] \).

(II) If \( \alpha_o < 1 \), the regret \( R^a \) can be upper bounded as follows:

\[
R^a \leq \sum_{t=1}^{T} \ell_t \ln(\tau_B + O(1)),
\]

where \( \ell_t = (4r^2(\mu_i^*')^2)/\Delta_i, \ell_{i,j} = (\mu_i^* + \alpha + 1)^2/\Delta_i, \ell_{o,i} = (\mu_i^* + 1)^2/\Delta_i, \ell_{o,i} = (\mu_i^* + 1)^2/\Delta_i, \) and \( O(1) \) is only related to \( \mu_i^*, \mu_i \) for any \( i \in [K] \).

(III) If \( \alpha_o = 1 \) for any \( a \in \{i, c, m\} \), the regret \( R^a \) is \( O(\ln B) \); if \( \alpha_o = 1 \), the regret \( R^b \) of b-GREEDY is \( O(\ln B) \).

More detailed theoretical regrets are summarized in Appendix A. In general, the regret bounds of the proposed algorithms are sublinear in terms of \( \tau_B \) (i.e., \( O(1/\tau_B) \)) and thus \( B \). By carefully hyper parameters, making \( \alpha_o > 1 \) (by setting \( \alpha \) large enough), all the four algorithms can achieve log linear regret bounds. Corollary 3 provides a group of \( \alpha \)'s that can achieve \( O(\ln B) \) regrets:

**Corollary 3.** If the parameters \( \lambda \) and \( D \), which satisfy \( \lambda \leq |\psi_{\text{min}}| \) and \( D \leq \min_{i \in [K]} \Delta_i \), are known in advance, by setting \( \alpha \) as \( 2(1 + 1/\lambda, 2(1 + 1/\lambda, 2 \text{and max}[10, 16/\psi_{\text{min}}]) \) for i-UCB, c-UCB, m-UCB, and b-GREEDY respectively, their regrets are

\[
R^a \leq \sum_{t=1}^{T} \ell_t \ln(\tau_B + O(1)). \tag{7}
\]

For ease of reference, denote the second \( \lambda \) of (11) as \( \lambda_B \).

(S3) Bound \( \sum_{i=1}^{K} \hat{\psi} \): Note that given \( t \geq K + 1 \), we have \( n_{ij} \geq 1 \) and further, \( E_{ij}^t > 0 \). If \( \tau_i \neq 1 \), the ratio \( (\hat{\tau}_i, \hat{E}_{ij}^t) \) is infinity. Then \( \psi(\hat{\psi}, \hat{\psi}) = 0 \). Therefore, we only need to consider the case of \( \hat{\tau}_i > 0 \). Since

\[
\frac{\hat{\tau}_i}{\hat{\psi}_i} = \frac{\mu_i^*}{\mu_i^* + \mu_i^*'} + \frac{\hat{\tau}_i - \mu_i^*}{\mu_i^*}, \tag{12}
\]

where \( \hat{\psi}_i = (\mu_i^*')^2(1 + 1/\lambda, 2(1 + 1/\lambda, 2 \text{and max}[10, 16/\psi_{\text{min}}])) \).
if \( \neg \mathcal{A}_i \) holds, at least one of the following equations is true:

\[
(\tau_{ij} - \mu_i') \leq \frac{\mu_i'}{\mu_i' + \mu_{ij}'} \leq \frac{\mu_i E_{ij}'}{\mu_i' + \mu_{ij}'} \leq - \frac{\mu_i}{\mu_i' + \mu_{ij}'} E_{ij}'
\]

As a result, we have \( \mathbb{P}[\neg \mathcal{A}_i] = \mathbb{P}[\neg \mathcal{A}_i, \varepsilon_{ij} > 0] \), and

\[
\mathbb{P}[\neg \mathcal{A}_i, \varepsilon_{ij} > 0] \geq \mathbb{P}\left(\tau_{ij} - \mu_i' \leq \frac{\mu_i E_{ij}'}{\mu_i' + \mu_{ij}'} \leq \frac{\mu_i}{\mu_i' + \mu_{ij}'} E_{ij}' \right) \]

\[
+ \mathbb{P}\left(\tau_{ij} - \mu_i' \leq - \frac{\mu_i}{\mu_i' + \mu_{ij}'} E_{ij}' \right)
\]

\[
\leq \mathbb{P}\left(\tau_{ij} - \mu_i' \leq \frac{\mu_i E_{ij}'}{\mu_i' + \mu_{ij}'} \right) + \mathbb{P}\left(\tau_{ij} - \mu_i' \leq - \frac{\mu_i}{\mu_i' + \mu_{ij}'} E_{ij}' \right).
\]

(13)

To make a clearer expression, temporarily denote \( X_i \) as reward from the \( t \)-th pulling of arm \( i \). Note that \( X_i \)'s are independent for different \( t \)'s. By the peeling technique, we have

\[
\mathbb{P}[\tau_{ij} \leq \mu_i' - \frac{\alpha \mu_i'}{\mu_i' + \mu_{ij}'} \left( \ln(t - 1) \right)\]

\[
\leq \mathbb{P}\left( \sum_{s=t+1}^{t+1} (X_i - \mu_i') \leq - \frac{\alpha \mu_i'}{\mu_i' + \mu_{ij}'} \sqrt{\ln(t - 1)} \right)
\]

\[
\leq \sum_{s=t+1}^{t+1} \mathbb{P}\left( \sum_{s=t+1}^{t+1} (X_i - \mu_i') \leq - \frac{\alpha \mu_i'}{\mu_i' + \mu_{ij}'} \sqrt{\ln(t - 1)} \right)
\]

\[
\leq 0 \sum_{s=t+1}^{t+1} (t - 1)^{\alpha} \leq \left( \log(t - 1) + 1 \right)(t - 1)^{-\alpha},
\]

in which the “\( \leq \)” marked with \( \hat{=} \) is obtained by Hoeffding’s maximal inequality (See Fact 2 in Appendix B). The second term of (13) can be similarly obtained as above. Thus, (13) is upper bounded by \( 2(\log(t - 1) + 1)(t - 1)^{-\alpha} \).

Define \( \varphi_B(x) = \sum_{i=1}^{t} 2(\log(t - 1) + 1)(t - 1)^{-\alpha} \) for any \( x > 0 \). Please note that when \( x \in (0, 1) \), \( \varphi_B(x) \) is order of \( \mathcal{P}[B] \); when \( x = 1 \), \( \varphi_B(x) \) is order of \( O(\ln^2 t \gamma) \); when \( x > 1 \), \( \varphi_B(x) \) can be bounded by a term irrelevant to \( B \). We can obtain that

\[
\sum_{i=1}^{t} \mathbb{P}[\neg \mathcal{A}_i] \leq \varphi_B(\alpha).
\]

(14)

\section{Lower Bound for Budgeted Bernoulli MABs}

Budgeted Bernoulli MAB is a kind of MAB that the reward and cost is either 0 or 1, i.e., \( r_i(t), c_i(t) \in \{0, 1\} \) \( \forall i \in [k], \forall t \geq 1 \). The lower bounds related to Bernoulli MAB have been widely studied for standard MAB \([1, 2, 3]\) literature. We give a lower bound for budgeted Bernoulli MAB. Like the standard bandits, the lower bound for the budgeted MAB also depends on the KL-divergence of the distributions between the optimal arm and suboptimal arms. Mathematically, define the KL-divergence of two Bernoulli distributions with success probabilities \( x \) and \( y \) as follows:

\[
kl(x, y) = x \ln \frac{x}{y} + (1 - x) \ln \frac{1 - x}{1 - y}, \quad \forall x, y \in (0, 1).
\]

(15)

To make a suboptimal arm \( j \in [k] \) an “optimal” \( \delta \)-gap for arm \( i \), since the KL-divergence defined in (17) is minimized. We can get that for any \( j \geq 2, \gamma^* \) and \( \gamma^*_j \) always exists, and \( \gamma^*_j \) is strictly positive.

The number of pulling rounds of arm \( i \in [k] \) in the first \( B \) rounds as \( \hat{T}^B \). The lower bound of \( \hat{T}^B \) for Bernoulli budgeted bandits is shown in the following theorem:

\textbf{Theorem 5.} For Bernoulli budgeted MABs, if the rewards are independent to the costs of each arm, for any pulling policy such that \( B \to \infty \), for any \( a \in (0, 1) \), \( \sum_{i=1}^{K} \mathbb{E}[T_i^B] = o(B^a) \), we have that for any \( \epsilon > 0, i \geq 2, \lim_{B \to \infty} \mathbb{P}[T_i^B \geq \frac{\epsilon}{(1 - \epsilon) \ln B}] = 1 \), and accordingly, \( \lim_{B \to \infty} \mathbb{E}[\tau^B_i] / \ln(B) \geq \gamma^*_j \).

We can easily bridge the \( \hat{T}^B \) to the regret. That is:

\textbf{Corollary 6.} For Bernoulli budgeted MABs, if the rewards are independent to the costs of each arm, the regret of any pulling policy, which satisfies that as \( B \to \infty \), \( \forall a \in (0, 1), \sum_{i=1}^{K} \mathbb{E}[T_i^B] = o(B^a) \), is at least \( \Omega(\sum_{i=1}^{K} \mathbb{E}[T_i^B]) \).

The basic idea of the proof of the theorem is as follows. First, we note that in budgeted MAB, the pulling time is at least \( B \), since the cost of each arm per pulling is at most 1. Second, to remove the randomness of the stopping time, we consider the pulling time of a suboptimal arm \( j \) during the first \( B \) rounds, which is a lower bound of \( \hat{T}^B \). Third, we extend the change-of-measure technique in \([12, 22]\) into budgeted MAB. Finally, with some calculations, we obtain the lower bound in Theorem 5. Corollary 6 can be obtained by some simple derivations. We leave the proof of Theorem 5 and Corollary 6 at Appendix D.

We can see that m-UCB can achieve an asymptotic \( O(\ln B) \) regret bound without any additional information of the bandit. Recall that Theorem 5 claims that the lower bound for budgeted Bernoulli MAB in terms of \( B \) is \( \Omega(\ln B) \). Therefore, by choosing the hyper parameters carefully, the regret bounds of all the four proposed algorithms can match the lower bound of budgeted Bernoulli MAB in terms of \( B \).
4.4. Discussions

4.4.1. Special Cases Discussion

By choosing the parameters carefully, our proposed algorithms can achieve regret $O(C_L \ln B)$. Meanwhile, denote the lower bound as $\Omega(C_L \ln B)$. It is difficult for our proposed UCB algorithms and the greedy algorithm to match the lower bound perfectly (i.e., $C_L > C_\delta$), which is a common drawback for conventional MABs (without budget constraints). However, we can still see that $C_L$ and $C_\delta$ share many similar trends. We discuss the m-UCB with $\alpha = 2$ in the following examples and the other algorithms can be similarly discussed.

(Example 1) The regrets w.r.t $\Delta_i$ for any $i \geq 2$: We introduce a similar setting in [21] to discuss the relations between the upper bounds of our algorithms and the lower bound. Assume that there exists an $l_0 \in (0, 0.5)$ s.t. $\mu_i^*, \mu_i^- \in [l_0, 1 - l_0]$ and $\Delta_i < l_0/2$ for any $i \in [K]$. One can verify that $\gamma = 1$ satisfies the constraint in (17) and $\delta_i(1) < l_0/2$. According to the fact that for any $p, q \in (0, 1)$,

$$2(p - q)^2 \leq kl(p, q) \leq \frac{(p - q)^2}{q(1 - q)},$$

we can get that $L^*_i$ is upper bounded by

$$\frac{\hat{\delta}_i^2(1)}{\gamma(\mu_i^* + \delta_i - \delta(1)) + (\mu_i^- - \delta_i)(1 - \mu_i^* + \delta_i(1))} \leq \frac{4\Delta_i^2}{\gamma^2}$$

Thus, $C_L = \sum_{i=2}^{K}(l_0^2/\Delta_i)$. We also know that $C_u = \sum_{i=2}^{K}(l_0^2/\Delta_i)$. Therefore, both the upper bound and the lower bound of the regret are linear to $\sum_{i=2}^{K}(1/\Delta_i) \ln B$.

(Example 2) The regrets w.r.t to the magnitude of the expected costs. Consider the $K$-armed Bernoulli bandits with expected rewards $\mu_i^{K}$ and expected costs $\theta_i^{K}$ for any $i$ is fixed and $\theta \in (1/\sqrt{B}, 1)$. Since $\gamma \approx \infty$ satisfies the constraint in (17), we have that $L^*_i$ is upper bounded by

$$\frac{\hat{\delta}_i^2(\infty)}{(\theta \mu_i^* - \delta(\infty))(1 - \theta \mu_i^* + \delta(\infty))} \leq \frac{\theta \mu_i^2 \Delta_i}{\mu_i^2(1 - \mu_i^2)}$$

That is, $C_L = K/\theta$. On the other hand, we can verify that the upper bound is of order $O(K/\sqrt{\theta^*} \ln(B/\theta^*))$. The upper and lower bound are not perfectly matched in terms of $\theta$, we can still get that we can see that when the magnitudes of the expected costs of all the arms become smaller, the regret becomes larger.

4.4.2. Proof Technique for Existing Algorithms

Our proof technique (and the proposed $\delta$-gap (5)) can also be used to analyze multiple existing algorithms:

(A) For the standard MAB [6] in which there is no cost of each arm and the pulling time is bounded by a given $T$, the $\Delta_i$ in [6] is exactly the $\delta_i(0)$ ($i \geq 2$) of our analysis. Thus, our proof technique can cover that of standard multi-armed bandit.

(B) For the budgeted MAB with fixed cost of each arm (studied in [27]), using our proposed method, we can obtain a better result than Theorem 2 in [27] (and Theorem 7 in its arXiv version). See Appendix E for details.

(C) For the Budgeted Thompson sampling algorithm proposed in [31], with our asymmetric $\delta$-gap, we can get that when the reward and cost are both Bernoulli for each arm, BTS can achieve an improved regret:

**Theorem 7.** For Bernoulli budgeted bandits, there exists an $\epsilon_0 \in (0, 1)$ such that for any $\epsilon \in (0, \epsilon_0)$, the regret of BTS are upper bounded by

$$\sum_{i=1}^{K} \frac{(1 + \epsilon)\Delta_i}{L_i} \ln T_B + X(B) \sum_{i=1}^{K} \Delta_i + O(K\epsilon^2),$$

where $L_i = \sup_{\beta} \min\{kl(\mu_i^*, \mu_i^- + \delta_i(\gamma)), kl(\mu_i^*, \mu_i^- - \gamma\delta_i(\gamma))\}$.

Please note that the regret in Theorem 7 outperforms the result in [31] and is closer to the lower bound. The proof of Theorem 7 is left in Appendix F.

5. EMPIRICAL EVALUATIONS

The theoretical results in the previous section can guarantee the asymptotic performances of the algorithms when the budget is very large. In this section, we empirically evaluate the practical performances of the algorithms when the budget is limited. Temporally, we do not verify the performances of the variants of the UCB algorithms.

Specifically, we first simulate a 10-armed Bernoulli bandit, in which the rewards and costs of each arm follow Bernoulli distributions, and a 10-armed Beta bandit, in which the rewards and costs of each arm follow Beta distributions. The parameters of the Bernoulli distribution and beta distribution are randomly chosen from $(0, 1)$. Then we simulate a 50-armed Bernoulli bandit and a 50-armed Beta bandit, to see whether the number of arms will lead to different results.

When implementing the algorithms introduced in the previous sections, the hyper parameters of the i-UCB, c-UCB and m-UCB are all chosen as 0.25. The hyper parameter in bGREEDY is chosen as 4. For the comparison purpose, we implement several baseline algorithms. The first is the budgeted Thompson sampling algorithm (BTS) proposed in [31]. The second baseline is the UCB-BV1 algorithm [14] designed for random discrete costs. The third baseline is the e-first algorithm [26], designed for deterministic costs. The fourth baseline is a variant of the PD-BwK algorithm [7], which we call vPD-BwK. At each time, this baseline algorithm plays the arm with the maximum ratio $\min(\mu_i^2, rad(\nu_i, 1)) \frac{n}{\max(\mu_i^2, rad(\nu_i, 1))}$, where $rad(\nu, N) = \sqrt{\frac{3}{N}} + \frac{20}{N}$ and $\text{Crnd} = x \log(BK)$. Note that we do not use the original PD-BwK algorithm because it requires a hard time constraint $T$ that indicates how many times one can play arms. However, in our setting, there is no such time constraint and therefore the original PD-BwK algorithm cannot be directly applied. For the hyper parameters of the baselines, we choose $\epsilon = 0.1$ for e-first and $x = 0.25$ for vPD-BwK to keep consistency with [31].

To run vPD-BwK and the e-first algorithm, the budget $B$ needs to be known in advance. In our experiments, we try a set of budgets: $\{100, 200, 500, 1K, 2K, 5K, 10K, 20K, 30K, 40K, 50K\}$. For each budget, we repeat
the experiments for 500 runs and report the average performance. All the other algorithms do not need to know \( B \) in advance. Therefore, by setting \( B = 50K \), we can get the performances of these algorithms for all budgets below 50K. Similarly, we repeat the experiments for 500 runs for these algorithms and report the average performance.

Figure 1 shows the performance of each algorithm for the 10-armed bandits. We do not include the regret curve of UCB-BV1 in Figure 1(a) and 1(b), because its empirical regrets are much worse than all the other algorithms and will affect the visualization of the results (its regret at \( B = 50K \) are 1687.7 for the Bernoulli bandit and 2151.8 for the Beta bandit). As can be seen from the figures, no matter for the Bernoulli bandit or the Beta bandit, the four extended algorithms introduced in this paper outperform the baselines (BTS, UCB-BV1, \( \epsilon \)-first, and vPD-BwK). Among them, i-UCB performs the best for the Bernoulli bandits while c-UCB performs best for the beta bandits.

In addition to the average regrets shown in Figure 1(a) and 1(b), we also visualize the per-run performance of each algorithm in Figure 1(c) and 1(d), in which rank 1 means an algorithm performs the best and rank 2 means it performs the second best. From Figure 1(c), we see that i-UCB performs the best in 252 runs out of the total 500 runs and b-GREEDY performs the best in 140 runs. Overall, the four extended algorithms introduced in this paper take the top 4 positions in most runs, e.g., i-UCB is ranked in top 4 positions for more than 400 runs. Thus, we can come to the conclusion that the four simple extended algorithms outperform the baseline algorithms with a high probability.

Recall that Corollary 3 shows that the four algorithms can achieve log linear regret when \( B \rightarrow \infty \) if carefully choosing the hyper parameters. One might also be interested in the empirical regrets for limited budget \( B \) when the \( \alpha \) is chosen in the way provided in Corollary 3. Our experiments show that such an approach of setting the hyper parameters is not good and leads to larger regret when \( B \) is limited. Taking \( B = 50000 \) and the Bernoulli bandit as an example, the regret of the four algorithms is 1806.3, 1784.8, 1693.9, 2003.3 respectively, which are much worse the results shown in Figure 1(a).

The results when the 50-armed bandits are shown in Figure 2. Again, we observe that the simple extended algorithms introduced in this paper perform better than the existing algorithms. In particular, comparing with the 10-armed bandits, the 50-armed bandits have more arms and therefore need more explorations. As a result, the regrets for the 50-armed bandits are larger than those for the 10-armed bandits.

Finally, we investigate the sensitivity of the proposed algorithms w.r.t the hyper parameter \( \alpha \). Figure 3 shows the results of a 50-armed beta bandit. We can see that b-GREEDY and i-UCB are quite stable to \( \alpha \). We need to set a relatively small \( \alpha \) for m-UCB and c-UCB; otherwise their performance will drop a lot.
6. Better Regret Bound for Budget-UCB

The Budget-UCB algorithm proposed in [30] can be seen as a combined version of our c-UCB and m-UCB and specializes the $\alpha$ as $\sqrt{\frac{2}{\lambda}}$. Mathematically, the index for Budget-UCB $i^*$ for any $i \in [K], t > K$.

$$\delta_{t+1} = \frac{\bar{r}_{i}^t}{c_{i}^t} + \frac{E_{i}^t}{c_{i}^t} \min\{\bar{r}_{i}^t + E_{i}^t, 1\} \frac{\max\{c_{i}^t - E_{i}^t, 0\}}{}$$  \hspace{1cm} (22)

With the $\delta_{t}(1)$ gap, we can get an improved regret bound compared to that in [30]: (The proof is in Appendix G.)

**Corollary 8.** By setting $\alpha = \sqrt{\frac{2}{\lambda}}$ in Eqn.(22), the regret of Budget-UCB is upper bounded by

$$\sum_{t=1}^{T} \frac{3 + 2\sqrt{2} \lambda}{\Delta_i} (\frac{\mu_i^*}{\mu_i^* - \bar{r}_i^t})^2 \ln T_B + (\lambda(B) + 3) \frac{2\mu_i^*}{\mu_i^* - \bar{r}_i^t}.$$  \hspace{1cm} (23)

7. CONCLUSIONS

In this work we have studied the budgeted MAB problems. We show that simple extensions of the algorithms designed for standard MAB work surprisingly well for budgeted MAB: they enjoy sublinear regret bounds with respect to the budget and perform comparably or even better than the baseline methods in our simulations. We also give a lower bound for the budgeted Bernoulli MAB.

There are many directions to explore in the future. First, in addition to the simulations, it would make more sense to test the performances of the extended algorithms in real-world applications. Second, in addition to the distribution-dependent regret bounds, it would be interesting to study distribution-free bounds as well. Third, in addition to extending UCB1 and $\epsilon$-GREEDY, we will consider extending other algorithms designed for standard MAB to the budgeted settings.

---

**References**


Appendix

Appendix A. Detailed Theoretical Results

To present the detailed regret bounds of the proposed algorithms, we need to first introduce some other notations:

1. \( J_{\alpha}^{\text{UCB}}(\alpha) = (\alpha \mu_{\max})^2 / \mu_1^2 + 1)^2 \)
2. \( J_{\beta}^{\text{UCB}}(\beta) = ((\alpha \mu_{\max})^2 / (2 \mu_1^2 + 1)^2) \)
3. \( \varphi_{\beta}(x) = \sum_{i=1}^{K} 2 \left( \log_2(t - 1) + 1 \right) \left( 1 / (x - 1)^2 \right) \forall x > 0 \)
4. \( \tilde{\varphi}_{\beta}(x) = \frac{e^{1 - x / \alpha K + 1}}{x} \sum_{i=1}^{K} \tilde{\varphi}(x > 0) \)
5. \( \psi_{\beta}(x) = (x - 1)^i 1 \{ x \leq 1 \} + 1 \{ x > 1 \} \)
6. \( \tilde{\psi}_{\beta}(x) = \alpha + 4 (1 / \kappa(\alpha))^2 \psi + \frac{16(\kappa(\alpha))^3}{\varphi_{\beta}} \) where \( \kappa(\alpha) = \frac{e^{1 - x / \alpha K + 1}}{x} \)

Please note that the functions \( \varphi_{\beta}(\cdot) \), \( \tilde{\varphi}_{\beta}(\cdot) \) and \( \psi_{\beta}(\cdot) \) are all sublinear in terms of \( B \). The \( X(B) \) is a refined version of the one defined in Lemma 4. When \( B \to \infty \), \( X(B) \to 0 \). \( X(B) \) can also be bounded by a term related to \( \mu_{\min}^c \) only.

**Theorem 9.** All the four algorithms enjoy sublinear regret bounds. Specifically,

\[
\mathcal{R} \leq \sum_{i=1}^{K} 4 \alpha^2 \mu_{\min}^c \ln \tau_B + \sum_{i=1}^{K} \Delta_i \left( \varphi_{\beta}(J_{\alpha}^{\text{UCB}}(\alpha)) + \frac{7(\mu_{\max}^c + 1)^2}{\Delta_i^2} \psi_{\beta}(J_{\beta}^{\text{UCB}}(\beta)) + 1 + X(B) \right) + 2 \mu_1^c / \mu_1^2 \\
\mathcal{R} \leq \sum_{i=1}^{K} \left( \alpha + 1 \right)^2 \ln \tau_B + \left( \varphi_{\beta}(\alpha) + 3 + X(B) \right) \sum_{i=1}^{K} \Delta_i + 2 \mu_1^c / \mu_1^2 \\
\mathcal{R} \leq \sum_{i=1}^{K} \frac{\left( \alpha + 1 \right) \mu_{\min}^c + 1}{\Delta_i} \ln \tau_B + \left( \varphi_{\beta}(\alpha) + 3 + X(B) \right) \sum_{i=1}^{K} \Delta_i + 2 \mu_1^c / \mu_1^2 \\
\mathcal{R} \leq \sum_{i=1}^{K} \Delta_i \left( \alpha \ln \tau_B + 4 \varphi_{\beta}(\alpha) + \frac{16}{\varphi_{\beta}} \psi_{\beta}(\varphi_{\beta}) + X(B) + g_i(\alpha) \right) + 2 \mu_1^c / \mu_1^2 \\
\]

In general, the regret bounds of the four proposed algorithms are sublinear in terms of \( B \). By choosing the hyper parameters carefully, all the four algorithms can achieve \( O(\ln B) \) regret bounds, which are shown in Corollary 10.

**Corollary 10.** If the parameters \( \lambda \) and \( D \), which satisfy \( \lambda \leq \mu_{\min}^c \) and \( D \leq \min_{i \geq 2} \Delta_i \), are known in advance, by setting \( \alpha \) as \( 2(1 + \frac{1}{\lambda}), 2(1 + \frac{1}{\lambda}) \), and max\{10, \frac{10}{\lambda x B^2} \} + 1 \) for \( i \)-UCB, \( c \)-UCB, \( m \)-UCB and \( b \)-GREEDY respectively, their regrets are

\[
\mathcal{R} \lesssim \sum_{i=1}^{K} \frac{16 \mu_{\min}^c \left( 1 + \frac{2}{\lambda} \right)^2 \ln \tau_B + \sum_{i=1}^{K} \Delta_i \left( 4 + X(B) + \frac{7(\mu_{\max}^c + 1)^2}{\Delta_i^2} \right) + 2 \mu_1^c / \mu_1^2 }{\Delta_i} \\
\mathcal{R} \lesssim \sum_{i=1}^{K} \left( \frac{\alpha_{\text{opt}}^c + 1}{\Delta_i} \right)^2 \ln \tau_B + (X(B) + 6) \sum_{i=1}^{K} \Delta_i + 2 \mu_1^c / \mu_1^2 \\
\mathcal{R} \lesssim \sum_{i=1}^{K} \frac{9(\mu_{\min}^c + 1)^2}{\Delta_i} \ln \tau_B + (X(B) + 6) \sum_{i=1}^{K} \Delta_i + 2 \mu_1^c / \mu_1^2 \\
\mathcal{R} \lesssim \left( \max \{ 10, \frac{10}{\lambda x B^2} \} + 1 \right) \left( \sum_{i=1}^{K} \Delta_i \right) \ln \tau_B + C, \\
\]

where \( C \) is related to \( X(B) \) and the expected reward/cost of each arm.

Appendix B. Important Facts

**Fact 1 (Chernoff-Hoeffding bound, Fact 1 of [6]).** Let \( X_1, \cdots, X_n \) be random variables with common range \( [0, 1] \) and such that \( E[X_i | X_{i-1}, \cdots, X_1] = \mu \). Let \( S_n = X_1 + X_2 + \cdots + X_n \). Then for all \( a \geq 0 \),

\[
P[S_n \geq n\mu + a] \leq e^{-\frac{a^2}{2}} \\
P[S_n \leq n\mu - a] \leq e^{-\frac{a^2}{2}} \\
\]
Appendix C. Detailed analysis of the regret upper bounds

Appendix C.1. Detailed analysis of i-UCB

Throughout this subsection, (A) let $T_i$ denote the number of pulling rounds of arm $i$ from round 1 to round $T_k$; (B) denote $\delta_i(1)$ as $\delta_i t_i \geq 2$. By Eqn. (S), we know $\frac{\delta_i^2}{\mu_i^2} \geq \frac{\mu_i^2}{\mu_i^2} \geq \frac{\mu_i^2}{\mu_i^2}$. (C) temporarily omit the “$\alpha$” from $E_{i_0}$, i.e., denote $E_{i_0}$ as $E_{i_0}$ for any $i \in [K]$.

(S1) Decompose $E(T_i) \forall i \geq 2$: For any suboptimal arm $i$, given $\eta \in (0, 1)$, we have

$$T_i \leq 1 + \sum_{t \in S_i} 1\{k \in [t]\} + \sum_{t \in S_i} 1\{i, n_{i_j} \geq U_i(\eta)\}$$

$$\leq 1 + U_i(\eta) + \sum_{t \in S_i} \left(\prod_{\substack{\eta \in S_i \cap \{i\}}} \left(\frac{\mu_j}{\mu_i} + \mu_j \leq \mu_i \right) \right) \leq 1 + U_i(\eta) + \sum_{t \in S_i} \left(\prod_{\substack{\eta \in S_i \cap \{i\}}} \left(\frac{\mu_j}{\mu_i} + \mu_j \leq \mu_i \right) \right)$$

$$\leq 1 + U_i(\eta) + \sum_{t \in S_i} \left(\prod_{\substack{\eta \in S_i \cap \{i\}}} \left(\frac{\mu_j}{\mu_i} + \mu_j \leq \mu_i \right) \right)$$

in which $U_i(\eta) = \frac{\eta^2 \ln \frac{t \cdot \ln \eta}{\lambda}}{\lambda^2}$, and $\eta \in (0, 1)$ is a parameter to be determined later. Accordingly,

$$\mathbb{E}(T_i) \leq 1 + U_i(\eta) + \sum_{t \in S_i} \left(\prod_{\substack{\eta \in S_i \cap \{i\}}} \left(\frac{\mu_j}{\mu_i} + \mu_j \leq \mu_i \right) \right)$$

We will bound the sums of term $a$ and $b$ in the following two steps.

(S2-1) **Bound the sum of term $a$:** First we have $\mathbb{P}\left[\frac{\mu_i}{\mu_i} - E_{i_0} \geq 0\right] = 0$. Thus we only need to focus on the case of $\frac{\mu_i}{\mu_i} - E_{i_0} \geq 0$. Given an $E_{i_0}$, one can find a proper $d_i(t) \in (0, \mu_i^2]$ s.t. $\frac{\mu_i}{\mu_i} - \mu_i^2 \geq 0$, whose solution is $d_i(t) = (\mu_i^2 - E_{i_0})/(\mu_i^2 + 1 - E_{i_0})$.

Defined $T_i^{UCB}(a) = \left(\frac{\mu_i}{\mu_i} + 1\right)^2$. Defined that $T_i^{UCB}(a) = \frac{1}{2} \left(\frac{\alpha_i}{\mu_i^2 + 1}\right)^2$. Accordingly, we have

$$\mathbb{P}\left[\frac{\mu_i}{\mu_i} - E_{i_0} \geq 0\right] = \mathbb{P}\left[\frac{\mu_i}{\mu_i} + E_{i_0} \leq \frac{\mu_i}{\mu_i} \leq \frac{\mu_i}{\mu_i} \right]$$

$$= \mathbb{P}\left[\frac{\mu_i}{\mu_i} - \mu_i^2 \leq 0\right]$$

Recall we have defined that $\forall i \in [K]$, $n_{i_j} = \sum_{s=1}^{l_i} 1\{s \in S_i\}, \tau_{i_j} = \frac{t \cdot \ln \eta}{\lambda}$.
To make a clearer expression, temporarily denote \( \bar{\tau}_{i,t} \) as \( \bar{X}_{i,m_i}, \forall i \in [K] \). And \( X_{i,t} \) denotes the \( i \)-th observation of the reward of the \( i \)-th arm. \( \bar{X}_{i,t} = \frac{1}{t} \sum_{s=1}^{t} X_{i,s} \) We can get that

\[
\mathbb{P}\left[ \bar{\tau}_{i,t} \leq \mu_i' - \frac{\alpha \mu_i'}{n_i} + \frac{\ln(t-1)}{n_i} \right] \leq \mathbb{P}\left[ \exists s \in \{1, 2, \ldots, t-1\} \text{ s.t. } X_{i,s} \leq \mu_i' - \frac{\alpha \mu_i'}{n_i} + \frac{\ln(t-1)}{n_i} \right] \leq \mathbb{P}\left[ \exists s \in \{1, 2, \ldots, t-1\} \text{ s.t. } \sum_{s=1}^{t} (X_{i,s} - \mu_i') \leq -\frac{\alpha \mu_i'}{n_i} + \frac{\ln(t-1)}{n_i} \right] \leq \mathbb{P}\left[ \exists s \in \{1, 2, \ldots, t-1\} \text{ s.t. } \sum_{s=1}^{t} \left( \frac{1}{2} \right)^{i-1} (t-1) < s \leq \left( \frac{1}{2} \right)^{i-1} (t-1) \text{ s.t. } \sum_{s=1}^{t} (X_{i,s} - \mu_i') \leq -\frac{\alpha \mu_i'}{n_i} + \frac{\ln(t-1)}{n_i} \right] \leq \alpha \sum_{j=0}^{n-1} \left( \frac{1}{t-1} \right)^{j} e^{j \ln(t-1)} \leq \left( \log_2(t-1) + 1 \right) \left( \frac{1}{t-1} \right) e^{j \ln(t-1)}.
\]

where the “\( \leq \)” marked with \( \alpha \) is obtained by Hoeffding’s maximal inequality [11] (in Page 30). Similarly, we also have

\[
\mathbb{P}\left[ \bar{\tau}_{i,t} \geq \mu_i' + \frac{\alpha \mu_i'}{n_i} + \frac{\ln(t-1)}{n_i} \right] \leq \left( \log_2(t-1) + 1 \right) \left( \frac{1}{t-1} \right) e^{j \ln(t-1)}.
\]

Consequently, we have obtained that term \( a \) is bounded by \( 2 \left( \log_2(t-1) + 1 \right) \left( \frac{1}{t-1} \right) e^{j \ln(t-1)} \). Therefore, the sum term \( a \) is upper bounded by \( \varphi_B \left( \tau_a^{UCB} \right) \), which is sublinear in \( B \).

(S2-2) **Bound the sum of term b:** Similar to the analysis on term \( a \), we can find a proper \( d_b(t) \) s.t.

\[
\frac{\mu_i' + \delta_i - d_b(t)}{\mu_i' - \delta_i} = \frac{\mu_i' + \delta_i - \frac{\ln(t-1)}{n_i}}{\mu_i' - \delta_i},
\]

the solution of which is

\[
d_b(t) = \left( \frac{\alpha (\mu_i' - \delta_i)}{\mu_i'} \right) \left( \frac{\frac{1}{t-1} \ln(t-1)}{n_i} \right) + \alpha \left( \frac{1}{t-1} \right) \ln(t-1).
\]

For any \( \eta \in (0, 1) \), if \( n_{i,t} \geq U_i^j(\eta) \) and \( t \in [K + 1, \ldots, \tau_B] \), we can verify that \( 0 < d_b(t) \leq 1 - \eta \delta_i \). We have that

\[
\mathbb{P}\left[ \frac{\bar{\tau}_{i,t} + \mathcal{E}_{i,t}}{\tau_a} \geq \frac{\mu_i' + \delta_i}{\mu_i' - \delta_i}, n_{i,t} \geq U_i^j(\eta) \right] \leq \mathbb{P}\left[ \frac{\bar{\tau}_{i,t} + \mathcal{E}_{i,t}}{\tau_a} \geq \frac{\mu_i' + \delta_i - d_b(t)}{\mu_i' - \delta_i}, n_{i,t} \geq U_i^j(\eta) \right] \leq \mathbb{P}\left[ \bar{\tau}_{i,t} \geq \mu_i' + \delta_i - d_b(t), n_{i,t} \geq U_i^j(\eta) \right] + \mathbb{P}\left[ \bar{\tau}_{i,t} \leq \mu_i' - \delta_i + d_b(t), n_{i,t} \geq U_i^j(\eta) \right] \leq \mathbb{P}\left[ \bar{\tau}_{i,t} \geq \mu_i' + \eta \delta_i, n_{i,t} \geq U_i^j(\eta) \right] + \mathbb{P}\left[ \bar{\tau}_{i,t} \leq \mu_i' - \eta \delta_i, n_{i,t} \geq U_i^j(\eta) \right].
\]

For the simplicity of use, define

\[
f_b(t, \alpha, \eta) = 2 \frac{\alpha^2 \eta^2 \delta_i^2}{(\frac{1}{t-1} \ln(t-1))} \left( \frac{\mu_i' - \eta \delta_i}{\mu_i' - \delta_i} \right)^2.
\]

Again, temporarily denote \( \bar{\tau}_{i,t} \) as \( \bar{X}_{i,m_i}, \forall i \in [K] \). We can obtain that

\[
\mathbb{P}\left[ \bar{\tau}_{i,t} \geq \mu_i' + \eta \delta_i, n_{i,t} \geq U_i^j(\eta) \right] = \sum_{l=0}^{\tau_B-1} \mathbb{P}\left[ \bar{X}_{i,m_i} \geq \mu_i' + \eta \delta_i, n_{i,t} = l \right] \leq \sum_{l=U_i^j(\eta)}^{\tau_B-1} \mathbb{P}\left[ \bar{X}_{i,m_i} \geq \mu_i' + \eta \delta_i \right] \leq \exp(-2\eta^2 \delta_i^2) \quad \text{ (obtained by Hoeffding’s inequality)}
\]

\[
\leq \frac{\exp(2\eta^2 \delta_i^2)}{2\eta^2 \delta_i^2} e^{-f_b(\alpha, \eta)}
\]

Similarly, one can obtain that

\[
\mathbb{P}\left[ \bar{\tau}_{i,t} \leq \mu_i' - \delta_i + d_b(t), n_{i,t} \geq U_i^j(\eta) \right] \leq \frac{\exp(2\eta^2 \delta_i^2)}{\eta^2 \delta_i^2} e^{-f_b(\alpha, \eta)}.
\]

As a result, term \( b \) is upper bounded by \( \frac{\exp(2\eta^2 \delta_i^2)}{\eta^2 \delta_i^2} e^{-f_b(\alpha, \eta)} \). Accordingly, For term \( b \),

\[
\sum_{i, k=1}^{K} \mathbb{P}\left[ \frac{\bar{\tau}_{i,t} + \mathcal{E}_{i,t}}{\tau_a} \geq \frac{\mu_i' + \delta_i}{\mu_i' - \delta_i}, n_{i,t} \geq U_i^j(\eta) \right] \leq \frac{\exp(2\eta^2 \delta_i^2)}{\eta^2 \delta_i^2} e^{-f_b(\alpha, \eta)}.
\]
Therefore, according to (S1), (S2-1)~(S2-2), we have that
\[
E(T_i) \leq 1 + U_i^*(\eta) + \varphi_\delta(J_\alpha^{\text{UCB}}(\alpha)) + \frac{\exp(2\gamma^2)}{\eta\delta^2} I_{B \to \text{in}(\alpha\eta)}.
\] (C.5)

Due to the complexity of optimization over \( \eta \), we simply choose \( \eta = \frac{1}{2} \) and get that
\[
f_\delta(i, \alpha, \eta) = \frac{2\alpha^2\gamma^2\delta^2}{2(\delta_{i+1}^2 + 1)^2} \left( \frac{\mu_i^* - \eta \delta_i}{\delta_{i+1}^2} \right)^2 = \frac{2\alpha^2}{2(\delta_{i+1}^2 + 1)^2} \left( \frac{1}{2} \right)^2 = \alpha^2 \frac{2}{2(\delta_{i+1}^2 + 1)^2};
\]
\[
U_i^*(\frac{1}{2}) = \frac{4\alpha^2 \ln \tau_B}{\delta_i} \left( \frac{1}{2} \right)^2 \leq \frac{4\alpha^2 \ln \tau_B}{\delta_i}.
\]

Therefore, according to Lemma 4, we can obtain that the regret of i-UCB is upper bounded by
\[
\sum_{i=2}^{K} \frac{4\alpha_i^2 \ln \tau_B}{\Delta_i} + \sum_{i=2}^{K} \Delta_i \left( 1 + \varphi_\delta(J_\alpha^{\text{UCB}}(\alpha)) + \frac{7}{\delta_i^2} T_i^{1-J_\alpha^{\text{UCB}}(\alpha)} + X(B) \right) + \frac{2\alpha_i^2}{\mu_i^2}
\] (C.6)

Appendix C.2. The regret bound of i-UCB by setting \( \alpha = \frac{2(1+\frac{1}{2})}{4} \)

For term a, since \( J_\alpha^{\text{UCB}}(\alpha) = \left( \frac{\alpha_i^2}{\delta_i^2} \right)^{\frac{1}{4}} \), we know that
\[
\varphi_\delta(J_\alpha^{\text{UCB}}(\alpha)) = \sum_{i=K+1}^{\infty} 2 \left( \log_2(t-i+1) + 1 \right) \left( \frac{1}{2} \right)^{t-i+1} \leq \sum_{i=K+1}^{\infty} 2 \left( \log_2(t-i+1) + 1 \right) \left( \frac{1}{2} \right)^{t-i+1}
\]
\[
\leq \sum_{i=K+1}^{\infty} \left( \log_2(t-i+1) + 1 \right) \left( \frac{1}{2} \right)^{t-i+1} \leq \int_1^\infty \frac{1}{2} \ln r \, dr + \int_1^\infty \frac{2}{r} \, dr \leq 3.
\]

For term b, we have \( T_i^{1-J_\alpha^{\text{UCB}}(\alpha)} \leq 1 \). As a result,
\[
R_{\text{UCB}} \leq \sum_{i=2}^{K} \frac{16\alpha_i^2 \ln \tau_B}{\Delta_i} \left( \frac{1}{2} \right)^2 + \sum_{i=2}^{K} \Delta_i \left( 4 + X(B) + \frac{7}{\delta_i^2} \right) + \frac{2\alpha_i^2}{\mu_i^2}
\] (C.7)
which is a \( O(\ln B) \) regret bound.

Appendix C.3. Detailed analysis of m-UCB algorithm

Throughout this subsection, (A) let \( T_i \) denote the number of pulling rounds of arm \( i \) from round 1 to round \( \tau_B \) with m-UCB algorithm; (B) denote \( \delta_i(1) \) as \( \delta_i \), \( \forall i \geq 2 \); (C) temporally omit the “\( \alpha \)” from \( E_i^* \), i.e., denote \( E_i^* \) as \( \hat{E}_{ij} \) for any \( i \in [K] \); (D) Define \( U_i^*(\eta) = \frac{\alpha_i^2 \ln \tau_B}{\delta_i(1-\eta)^2} \).

(S1) Decompose \( T_i \) for m-UCB:
\[
T_i \leq 1 + \sum_{j=K+1}^{\infty} 1 \{ \hat{I}_{ij} = i \} \leq U_i^*(\eta) + \sum_{j=K+1}^{\infty} \left\{ \frac{\min\{\hat{I}_{ij} + E_{ij}, 1\}}{\max\{\hat{E}_{ij} - E_{ij}, 0\}} \geq \frac{\min\{\hat{I}_{ij} + E_{ij}, 1\}}{\max\{\hat{E}_{ij} - E_{ij}, 0\}}, \forall j \neq i, n_{ij} \geq [U_i^*(\eta)] \}
\]
\[
\leq 1 + U_i^*(\eta) + \sum_{j=K+1}^{\infty} \left\{ \frac{\min\{\hat{I}_{ij} + E_{ij}, 1\}}{\max\{\hat{E}_{ij} - E_{ij}, 0\}} \geq \frac{\min\{\hat{I}_{ij} + E_{ij}, 1\}}{\max\{\hat{E}_{ij} - E_{ij}, 0\}}, n_{ij} \geq U_i^*(\eta) \}
\]
\[
\leq 1 + U_i^*(\eta) + \sum_{j=K+1}^{\infty} \left\{ \frac{\min\{\hat{I}_{ij} + E_{ij}, 1\}}{\max\{\hat{E}_{ij} - E_{ij}, 0\}} \geq \frac{\min\{\hat{I}_{ij} + E_{ij}, 1\}}{\max\{\hat{E}_{ij} - E_{ij}, 0\}}, n_{ij} \geq U_i^*(\eta) \}
\]
\[
+ 1 \left\{ \frac{\min\{\hat{I}_{ij} + E_{ij}, 1\}}{\max\{\hat{E}_{ij} - E_{ij}, 0\}} \geq \frac{\min\{\hat{I}_{ij} + E_{ij}, 1\}}{\max\{\hat{E}_{ij} - E_{ij}, 0\}}, n_{ij} \geq U_i^*(\eta) \right\}
\]
\[
\leq 1 + U_i^*(\eta) + \sum_{j=K+1}^{\infty} \left\{ \frac{\min\{\hat{I}_{ij} + E_{ij}, 1\}}{\max\{\hat{E}_{ij} - E_{ij}, 0\}} \leq \frac{\mu_j}{\mu_i}, \right\} + 1 \left\{ \frac{\min\{\hat{I}_{ij} + E_{ij}, 1\}}{\max\{\hat{E}_{ij} - E_{ij}, 0\}} \geq \frac{\mu_j + \delta_i}{\mu_i}, \right\}
\]
Since \( \mu'_i < 1 \) and \( \mu'_i > 0 \) \( \forall i \in [K] \), we have \( \mathbb{E}[T_i] \) is upper bounded by

\[
1 + U_i^\pi(\eta) + \sum_{r=1}^{r_{\text{UB}}} \mathbb{P} \left[ \min \{ \tau_{ij}, \tilde{E}_{ij} \}, 1 \right] \leq \frac{\mu'_i}{\mu'_i} + \mathbb{P} \left[ \min \{ \tau_{ij}, \tilde{E}_{ij} \} \right] \geq \frac{\mu'_i + \delta_i}{\mu'_i - \delta_i}, n_j \geq U_i^\pi(\eta) \]

\[
\leq 1 + U_i^\pi(\eta) + \sum_{r=1}^{r_{\text{UB}}} \mathbb{P} \left[ \min \{ \tau_{ij}, \tilde{E}_{ij} \}, 1 \right] \leq \mu'_i \right] + \mathbb{P} \left[ \max \{ \tau_{ij}, \tilde{E}_{ij} \} \right] \geq \mu'_i \]

\[
+ \mathbb{P} \left[ \min \{ \tau_{ij}, \tilde{E}_{ij} \}, 1 \right] \geq \mu'_i + \delta_i, n_j \geq U_i^\pi(\eta) \right] + \mathbb{P} \left[ \max \{ \tau_{ij}, \tilde{E}_{ij} \} \right] \leq \mu'_i - \delta_i, n_j \geq U_i^\pi(\eta) \right].
\]

(S2-1) **Bound the sum of term a**: Similar to the derivation of C.1, we have

\[
\mathbb{P} \left[ \tau_{ij} + \tilde{E}_{ij} \geq \mu'_i \right] + \mathbb{P} \left[ \tau_{ij} - \tilde{E}_{ij} \geq \mu'_i \right] \leq 2 \left( \log_2(t - 1) + 1 \right) \left( \frac{1}{1 - t} \right)^2.
\]

Therefore, the sum of term \( a \) is upper bounded by \( \varphi_B(\alpha^2) \).

(S2-2) **Bound the sum of term b**: \( \forall \eta \in (0, 1) \), given \( \tau_{ij} \), if \( n_j \geq U_i^\pi(\eta) \), we can obtain that \( \delta_i - \tilde{E}_{ij} \geq \eta \delta_i \). Thus, term \( b \) could be bounded as

\[
\mathbb{P} \left[ \tau_{ij} + \tilde{E}_{ij} \geq \mu'_i + \eta \delta_i, n_j \geq U_i^\pi(\eta) \right] + \mathbb{P} \left[ \tau_{ij} - \tilde{E}_{ij} \geq \mu'_i - \eta \delta_i, n_j \geq U_i^\pi(\eta) \right] \leq 2 \tau_B \exp \left\{ -2U_i^\pi(\eta)^2 \eta \delta_i^2 \right\} \leq 2 \tau_B \exp \left\{ - \frac{2 \alpha^2 \eta^2 \ln \tau_B}{(1 - \eta)^2} \right\},
\]

where the last step is obtained by Hoeffding's inequality and union bound (similar to the derivation of (C.3)). Thus,

\[
\mathbb{E} \left[ \sum_{i=1}^{r_{\text{UB}}} \mathbb{P} \left[ \tau_{ij} \geq \mu'_i + \delta_i, n_j \geq U_i^\pi(\eta) \right] \right] \leq 2 \tau_B 1 - \frac{2 \alpha^2 \eta^2 \ln \tau_B}{(1 - \eta)^2}.
\]

According to (S1) ~ (S2-2), we can obtain \( \mathbb{E}[T_i^\text{m-UCB}] \) is upper bounded by

\[
1 + \frac{\alpha^2 (\frac{\mu'_i}{1 - \eta})^2}{(1 - \eta)^2 \Delta_i^2} \ln \tau_B + \varphi_B(\alpha^2) + 2 \tau_B 1 - \frac{2 \alpha^2 \eta^2 \ln \tau_B}{(1 - \eta)^2}.
\]

According to Lemma 4, by setting \( \frac{2 \alpha^2 \eta^2}{(1 - \eta)^2} = 1 \), we have the regret bound of m-UCB is

\[
\sum_{i=2}^{K} \frac{2 \mu'_i}{\Delta_i (1 + \frac{\alpha^2 (\frac{\mu'_i}{1 - \eta})^2}{(1 - \eta)^2 \Delta_i^2})} \ln \tau_B + \left( 3 + \varphi_B(\alpha^2) + X(B) \right) \sum_{i=2}^{K} \Delta_i + \frac{2 \mu'_i}{\mu'_i^2}.
\]

By choosing \( \alpha = 2 \), we can obtain that

\[
\sum_{i=2}^{K} \frac{2 \mu'_i}{\Delta_i (1 + \frac{\alpha^2 (\frac{\mu'_i}{1 - \eta})^2}{(1 - \eta)^2 \Delta_i^2})} \ln \tau_B + \left( 6 + X(B) \right) \sum_{i=2}^{K} \Delta_i + \frac{2 \mu'_i}{\mu'_i^2}.
\]

**Appendix C.4. Detailed Analysis on b-GREEDY Algorithm**

Throughout this subsection, (A) let \( T_i \) denote the number of pulling rounds of arm \( i \) from round 1 to round \( \tau_B \) with b-GREEDY algorithm.

(S1) **Decompose \( \mathbb{E}[T_i] \)**: Obviously, \( \mathbb{E}[T_i] = \sum_{t=1}^{r_{\text{UB}}} \mathbb{P} [I_t = i] \). We will bound \( \mathbb{P} [I_t = i] \) in the next four steps:

(S2-1) **Decompose \( \mathbb{P} [I_t = i] \)**: It is obvious that

\[
\mathbb{P} [I_t = i] \leq \frac{\epsilon_i}{K} + (1 - \epsilon_i) \mathbb{P} \left[ \frac{\tau_{ij}}{\tilde{E}_{ij}} \geq \frac{\tau_{ij}}{\tilde{E}_{ij}} \right] \leq \frac{\epsilon_i}{K} + \mathbb{P} \left[ \frac{\tau_{ij}}{\tilde{E}_{ij}} \geq \frac{\tau_{ij}}{\tilde{E}_{ij}} \right].
\]

and the term \( \mathbb{P} \left[ \frac{\tau_{ij}}{\tilde{E}_{ij}} \geq \frac{\tau_{ij}}{\tilde{E}_{ij}} \right] \) can be further upper bounded by

\[
\mathbb{P} \left[ \frac{\tau_{ij}}{\tilde{E}_{ij}} \geq \frac{\tau_{ij}}{\tilde{E}_{ij}} \right] \leq \mathbb{P} \left[ \frac{\tau_{ij}}{\tilde{E}_{ij}} \geq \frac{\tau_{ij}}{\tilde{E}_{ij}} \right] \leq \mathbb{P} \left[ \frac{\tau_{ij}}{\tilde{E}_{ij}} \geq \frac{\tau_{ij}}{\tilde{E}_{ij}} \right] \leq \mathbb{P} \left[ \frac{\tau_{ij}}{\tilde{E}_{ij}} \geq \frac{\tau_{ij}}{\tilde{E}_{ij}} \right] \leq \mathbb{P} \left[ \frac{\tau_{ij}}{\tilde{E}_{ij}} \geq \frac{\tau_{ij}}{\tilde{E}_{ij}} \right].
\]

\[
\mathbb{P} \left[ \frac{\tau_{ij}}{\tilde{E}_{ij}} \geq \frac{\tau_{ij}}{\tilde{E}_{ij}} \right] \leq \mathbb{P} \left[ \frac{\tau_{ij}}{\tilde{E}_{ij}} \geq \frac{\tau_{ij}}{\tilde{E}_{ij}} \right] \leq \mathbb{P} \left[ \frac{\tau_{ij}}{\tilde{E}_{ij}} \geq \frac{\tau_{ij}}{\tilde{E}_{ij}} \right] \leq \mathbb{P} \left[ \frac{\tau_{ij}}{\tilde{E}_{ij}} \geq \frac{\tau_{ij}}{\tilde{E}_{ij}} \right] \leq \mathbb{P} \left[ \frac{\tau_{ij}}{\tilde{E}_{ij}} \geq \frac{\tau_{ij}}{\tilde{E}_{ij}} \right].
\]

\[
\mathbb{P} \left[ \frac{\tau_{ij}}{\tilde{E}_{ij}} \geq \frac{\tau_{ij}}{\tilde{E}_{ij}} \right] \leq \mathbb{P} \left[ \frac{\tau_{ij}}{\tilde{E}_{ij}} \geq \frac{\tau_{ij}}{\tilde{E}_{ij}} \right] \leq \mathbb{P} \left[ \frac{\tau_{ij}}{\tilde{E}_{ij}} \geq \frac{\tau_{ij}}{\tilde{E}_{ij}} \right] \leq \mathbb{P} \left[ \frac{\tau_{ij}}{\tilde{E}_{ij}} \geq \frac{\tau_{ij}}{\tilde{E}_{ij}} \right] \leq \mathbb{P} \left[ \frac{\tau_{ij}}{\tilde{E}_{ij}} \geq \frac{\tau_{ij}}{\tilde{E}_{ij}} \right].
\]
Given any rule for budgeted Bernoulli MAB such that as $B \to \infty$, we will only consider the regret at the first round as $\sum_{i=1}^{r_{i-1}} \epsilon_i$ and denote $n_i^B$ as the number of random plays of arm $i$ before (excluding) round $t$. According to Hoeffding’s inequality,

$$\mathbb{P}(\hat{r}_i \geq \mu_i + \epsilon_i) = \sum_{l=1}^{\epsilon_i} \mathbb{P}(n_{l,i} = l; \hat{r}_i \geq \mu_i + \epsilon_i) \leq \sum_{l=1}^{\epsilon_i} \mathbb{P}(\hat{r}_i \geq \mu_i + \epsilon_i, n_{l,i} = l) + \sum_{l=1}^{\epsilon_i} \mathbb{P}(n_{l,i} = l)$$

(C.15)

$$\leq \sum_{l=1}^{\epsilon_i} \exp\{-2l^2\epsilon_i^2\} + \mathbb{P}(n_{l,i} \leq [\epsilon_i]) \leq \frac{4}{\epsilon_i} e^{-2\epsilon_i^2} + \mathbb{P}(n_{l,i} \leq [\epsilon_i]).$$

Considering $\mathbb{E}[n_i^B] = 2\epsilon_i$ and

$$\text{Var}[n_i^B] = \sum_{l=1}^{\epsilon_i} \frac{e_i}{K} (1 - \frac{e_i}{K}) \leq \sum_{l=1}^{\epsilon_i} \frac{e_i}{K} = 2\epsilon_i,$$

according to the Bernstein inequality (Fact 3), we have $\mathbb{P}[n_i^B \leq \epsilon_i] \leq e^{-\frac{\epsilon_i^2}{2}},$ and thus

$$\mathbb{P}(\hat{r}_i \geq \mu_i + \epsilon_i) \leq \frac{4}{\epsilon_i} e^{-2\epsilon_i^2} + e^{-\frac{\epsilon_i^2}{2}}.$$  

(C.16)

Similarly, we can obtain the upper bounds for the other three terms as follows.

(A) $\mathbb{P}(\hat{r}_i \leq \mu_i - \epsilon_i) \leq \frac{4}{\epsilon_i} e^{-2\epsilon_i^2} + e^{-\frac{\epsilon_i^2}{2}}$;  
(B) $\mathbb{P}(\hat{r}_i \geq \mu_i + \epsilon_i) \leq \frac{4}{\epsilon_i} e^{-2\epsilon_i^2} + e^{-\frac{\epsilon_i^2}{2}}$;  
(C) $\mathbb{P}(\hat{r}_i \geq \mu_i + \epsilon_i) \leq \frac{4}{\epsilon_i} e^{-2\epsilon_i^2} + e^{-\frac{\epsilon_i^2}{2}}.$

(C.17)

(S2-3) Lower bound $\epsilon_i$: Assume $t' = \alpha K$, then we can easily verify that

$$\epsilon_i \geq \frac{t'}{2K} \geq \frac{\epsilon_i}{2K} \sum_{i=1}^{\epsilon_i} \frac{\alpha K}{s} \geq \frac{\alpha K}{2K} + \frac{\alpha}{2} \ln \frac{t}{\alpha K + 1} \geq \frac{\alpha}{2} \ln \left(\frac{1}{\kappa''(s)}\right).$$

(C.18)

where $\kappa''(s) = \frac{s^2 + 1}{s^3}.$

(S2-4) Bound $\mathbb{E}[T_i]$; Combining inequalities (C.13) to (C.18),

$$\mathbb{P}(\hat{r}_i \leq \epsilon_i) \leq \epsilon_i + 4e^{-\frac{\epsilon_i^2}{2}} + \frac{16}{\epsilon_i} e^{-\frac{\epsilon_i^2}{2}} \leq \frac{\alpha}{4} \ln \frac{t}{\alpha K + 1} + \frac{16\alpha}{\epsilon_i} \varphi_{\epsilon_i}(\epsilon_i) + g_i(\epsilon_i).$$

As a result,

$$\mathbb{E}[T_i] \leq \alpha \ln \frac{t}{\alpha K + 1} + \frac{16\alpha}{\epsilon_i} \varphi_{\epsilon_i}(\epsilon_i) + g_i(\epsilon_i).$$

(C.19)

where

$$g_i(\epsilon_i) = \alpha + 4(\frac{1}{\kappa''(s)}) \epsilon_i + \frac{16\alpha}{\epsilon_i} \varphi_{\epsilon_i}(\epsilon_i),$$

function $\varphi_{\epsilon_i}(x) = \left(\frac{1}{\kappa''(s)}\right)^{\frac{\alpha}{2}} \ln \left(\frac{t}{\alpha K + 1}\right) \sum_{i=1}^{\epsilon_i} \frac{1}{t}$ ∀ $x > 0$.

(C.20)

(S3) We can get the regret bound of b-GREEDY by Lemma 4.

Appendix D. Proof of Theorem 5 and Corollary 6

Denote the pulling number of arm $i$ at the first $B$ rounds as $\hat{T}_i$. (Note that in budgeted MAB, the pulling number is at least $B$.) The regret of the first $B$ rounds is certainly a lower bound of the budgeted Bernoulli MAB. To remove the randomness of the stopping time, we will only consider the regret at the first $B$ rounds.

To prove Theorem 5, we first need to prove Lemma 11.

Lemma 11. Given any rule for budgeted Bernoulli MAB such that as $B \to \infty$, for any $a > 0$, $\sum_{i=1}^{K} \mathbb{E}[T_i] = o(B^a)$, we have that for any $\epsilon > 0, \gamma > 0, \mu_j^* + \delta_j(\gamma) < 1$,

$$\lim_{B \to \infty} \mathbb{P}\left[\hat{T}_i \geq \frac{(1 - \delta) \ln B}{kl(\mu_j^*, \mu_j^* + \delta_j(\gamma)) + kl(\mu_j^*, \mu_j^* - \gamma \delta_j(\gamma))}\right] = 1$$

(D.1)

Proof of Lemma 11: Consider a $K$-armed budgeted Bernoulli bandit: $\forall i \in [K]$, the expected reward and cost of arm $i$ are $\mu_i$ and $\mu_i^*$ respectively. (We call this bandit “original bandit”.) Fix $j \geq 2$, given the proper $\gamma$ s.t. $\mu_j^* + \delta_j(\gamma) < 1$, for any $0 < \phi < 1$, we can always find an $x^*$ and an $\chi^*$ such that

$$kl(\mu_j^*, \mu_j^* + \delta_j(\gamma) + x^*) < (1 + \phi)kl(\mu_j^*, \mu_j^* + \delta_j(\gamma)) \text{ and } x^* > 0, \mu_j^* + \delta_j(\gamma) + x^* < 1;$$

$$kl(\mu_j^*, \mu_j^* - \gamma \delta_j(\gamma) - x^*) < (1 + \phi)kl(\mu_j^*, \mu_j^* - \gamma \delta_j(\gamma)) \text{ and } x^* > 0, \mu_j^* - \gamma \delta_j(\gamma) - x^* > 0.$$
Define $\chi' = \mu'_j + \delta_j (\gamma + x')$ and $\chi'' = \mu'_j - \gamma \delta_j (\gamma) - x'$. Consider a modified $K$-armed budgeted Bernoulli bandit: the expected reward and cost of arm $i \neq j$ are $\mu'_i$ and $\mu''_i$ respectively; the expected reward and cost of arm $j$ are $\chi'$ and $\chi''$ respectively.

We use the notation $\mathbb{P}_M$ and $\mathbb{E}_M$ when we integrate with respect to the original bandit. And we use the notation $\mathbb{P}_M'$ and $\mathbb{E}_M'$ when we integrate with respect to the modified bandit.

According to the assumption, we have

$$\mathbb{E}_M'(B - \tilde{T}_j) = \sum_{r \in \mathcal{R}_j} \mathbb{E}_M'(\tilde{T}_j) = o(B^r). \quad (D.2)$$

On the other hand, we can obtain that with $0 < a < \varrho$,

$$\mathbb{E}_M'(B - \tilde{T}_j) = \mathbb{E}_M'(B - \tilde{T}_j) \tilde{T}_j < \frac{(1 - \varrho) \ln B}{k(l(\mu'_j, \chi') + k(l(\mu'_j, \chi'')))} \mathbb{P}_M'(\tilde{T}_j < \frac{(1 - \varrho) \ln B}{k(l(\mu'_j, \chi') + k(l(\mu'_j, \chi''))})$$

$$+ \mathbb{E}_M'(B - \tilde{T}_j) \tilde{T}_j \geq \mathbb{P}_M'(\tilde{T}_j \geq \frac{(1 - \varrho) \ln B}{k(l(\mu'_j, \chi') + k(l(\mu'_j, \chi''))})$$

$$\geq \mathbb{E}_M'(B - \tilde{T}_j) \tilde{T}_j < \frac{(1 - \varrho) \ln B}{k(l(\mu'_j, \chi') + k(l(\mu'_j, \chi''))}) \mathbb{P}_M'(\tilde{T}_j < \frac{(1 - \varrho) \ln B}{k(l(\mu'_j, \chi') + k(l(\mu'_j, \chi''))})$$

$$\geq (B - \frac{(1 - \varrho) \ln B}{k(l(\mu'_j, \chi') + k(l(\mu'_j, \chi''))}) \mathbb{P}_M'(\tilde{T}_j < \frac{(1 - \varrho) \ln B}{k(l(\mu'_j, \chi') + k(l(\mu'_j, \chi''))}) \quad (D.3)$$

Accordingly, we can obtain that

$$\mathbb{P}_M'(\tilde{T}_j \leq \frac{(1 - \varrho) \ln B}{k(l(\mu'_j, \chi') + k(l(\mu'_j, \chi''))}) = o(B^{-r-1}). \quad (D.4)$$

Let $X'_1, X'_2, X'_2, \ldots, X'_B, X''_B$ denote successive reward and cost observations from the $j$-th arm. Note that $X'_i, X''_i \in \{0, 1\} \forall i \in [B]$. Define

$$L_m \equiv \sum_{i=1}^m \mu'_i X'_i + (1 - \mu'_i)(1 - \chi') X''_i + \sum_{i=1}^m \chi X''_i + (1 - \chi') X''_i$$

An important property is the following: for any event $A$ in the $\sigma$-algebra generated by $X'_1, X'_2, X'_2, \ldots, X'_B, X''_B$, the following change-of-measure identity holds:

$$\mathbb{P}_M'(A) = \mathbb{E}_M[I(A) \exp(-L_{\tilde{T}_j})]. \quad (D.5)$$

According to (D.4), we know that $\mathbb{P}_M'(\xi) = o(B^{-r-1})$ where

$$\xi = \tilde{T}_j \leq \frac{(1 - \varrho) \ln B}{k(l(\mu'_j, \chi') + k(l(\mu'_j, \chi''))) \text{ and } L_{\tilde{T}_j} \leq (1 - a) \ln B}. \quad (D.6)$$

Given $n_1, n_2, \ldots, n_K$ s.t. $\sum_{i=1}^K n_i = B, n_i \geq 0 \forall i \in [K]$, we have

$$\mathbb{P}_M'(\tilde{T}_1 = 1, \tilde{T}_2 = 2, \ldots, \tilde{T}_K = n_K, L_{\tilde{T}_j} \leq (1 - a) \ln B)$$

$$\mathbb{E}_M[I(\tilde{T}_1 = 1, \tilde{T}_2 = 2, \ldots, \tilde{T}_K = n_K, L_{\tilde{T}_j} \leq (1 - a) \ln B) \exp(-L_{\tilde{T}_j})]$$

$$\geq \exp((-1 - a) \ln B) \mathbb{P}_M'(\tilde{T}_1 = 1, \tilde{T}_2 = 2, \ldots, \tilde{T}_K = n_K, L_{\tilde{T}_j} \leq (1 - a) \ln B). \quad (D.7)$$

Since

$$\xi = \bigcup_{\tilde{T}_1 = 1, \tilde{T}_2 = 2, \ldots, \tilde{T}_K = n_K, L_{\tilde{T}_j} \leq (1 - a) \ln B} \tilde{T}_1 = 1, \tilde{T}_2 = 2, \ldots, \tilde{T}_K = n_K, L_{\tilde{T}_j} \leq (1 - a) \ln B]. \quad (D.8)$$

we know that $\xi$ can be decomposed into a group of the disjoint events. Therefore,

$$\mathbb{P}_M(\xi) \leq B^{-r} \mathbb{P}_M'(\xi) \rightarrow 0 \quad \text{as } B \rightarrow \infty. \quad (D.9)$$

By the strong law of large numbers, $\frac{L_m}{m} \rightarrow k(l(\mu'_j, \chi') + k(l(\mu'_j, \chi'')) > 0$, therefore, according to Lemma 10.5 in [11], as $m \rightarrow \infty$

$$\max_{i \in [n]} \frac{L_i}{m} \rightarrow k(l(\mu'_j, \chi') + k(l(\mu'_j, \chi'')) \quad \text{a.s.} \quad \mathbb{P}_M]. \quad (D.10)$$

Since $1 - a > 1 - \varrho$, it then follows that

$$\mathbb{P}_M(L_i > (1 - a) \ln B \text{ for some } i < \frac{(1 - \varrho) \ln B}{k(l(\mu'_j, \chi') + k(l(\mu'_j, \chi''))}) \rightarrow 0 \quad (D.11)$$

as $B \rightarrow \infty$. Therefore,

$$\lim_{B \rightarrow \infty} \mathbb{P}_M'(\tilde{T}_j < \frac{(1 - \varrho) \ln B}{k(l(\mu'_j, \chi') + k(l(\mu'_j, \chi''))})$$

$$= \lim_{B \rightarrow \infty} \mathbb{P}_M'(\tilde{T}_j < \frac{(1 - \varrho) \ln B}{k(l(\mu'_j, \chi') + k(l(\mu'_j, \chi''))), L_{\tilde{T}_j} > (1 - a) \ln B}] + \lim_{B \rightarrow \infty} \mathbb{P}_M(\xi) \quad (D.12)$$

$$= \lim_{B \rightarrow \infty} \mathbb{P}_M(\xi) = 0. \quad (D.13)$$
As a result, we have
\[
\lim_{B \to \infty} \Pr[\tilde{T}_i < \frac{1 - \epsilon}{1 + \epsilon} \frac{\ln B}{k l(\mu_i', \mu_i') + \delta_i(\gamma)} + kl(\mu_i', \mu_i' - \gamma \delta_i(\gamma)))] \leq \lim_{B \to \infty} \Pr[\tilde{T}_i < \frac{(1 - \epsilon) \ln B}{kl(\mu_i', \mu_i') + kl(\mu_i', \mu_i')} = 0. 
\] (D.13)

Therefore, by replacing \( \frac{1 - \epsilon}{1 + \epsilon} \rightarrow \epsilon \) with \( 1 - \epsilon \), we can obtain Lemma 11.

Given any algorithm “alg” for budgeted Bernoulli MAB such that as \( B \to \infty \), \( \forall a > 0 \), \( \sum_{i=2}^{K} E(T_i^{alg}) = o(B^2) \) means that \( \sum_{i=2}^{K} E(\tilde{T}_i) = o(B^2) \). By Lemma 11, we can obtain that for any \( \epsilon > 0 \), \( \gamma \geq 0 \), \( \mu_i' + \delta_i(\gamma) < 1 \).

\[
\lim_{B \to \infty} \Pr(\tilde{T}_i^{alg} \geq \frac{(1 - \epsilon) \ln B}{kl(\mu_i', \mu_i') + kl(\mu_i', \mu_i' - \gamma \delta_i(\gamma))}) = 1. 
\] (D.14)

Next we prove that \( L_i^\gamma \) is positive and it always exists. Let \( L_i^\gamma \) denote \( kl(\mu_i', \mu_i' + \delta_i(\gamma)) + kl(\mu_i', \mu_i' - \gamma \delta_i(\gamma)) \). Denote \( \gamma^* \) as the corresponding \( \gamma \) of \( L_i^\gamma \). Define \( \Gamma_j = \{ \gamma \geq 0 | \mu_j' + \delta_j(\gamma) < 1 \} \). Minimizing \( L_i^\gamma \) w.r.t \( \gamma \in \Gamma_j \) is equivalent to maximizing

\[
\tilde{L}_j = \mu_j' \ln \left( \frac{\Delta_j}{\gamma_j^{\mu_j'} + 1} + (1 - \mu_j') \ln \left( 1 - \mu_j' - \frac{\Delta_j}{\gamma_j^{\mu_j'} + 1} + \mu_j' \ln \left( \frac{\gamma_j^{\Delta_j}}{\gamma_j^{\mu_j'} + 1} \right) + (1 - \mu_j') \ln \left( 1 - \mu_j' + \frac{\gamma_j^{\Delta_j}}{\gamma_j^{\mu_j'} + 1} \right) \right) \right) .
\] (D.15)

The first order derivative of \( \tilde{L}_j \) w.r.t \( \gamma \) is:

\[
\frac{\partial \tilde{L}_j}{\partial \gamma} = \frac{\Delta_j}{(\gamma_j^{\mu_j'} + 1)^2} \left\{ - \frac{\mu_j^{\mu_j'}}{\gamma_j^{\mu_j' + 1}} + \frac{(1 - \mu_j') \gamma_j^{\Delta_j}}{(1 - \mu_j') \gamma_j^{\mu_j' + 1} - \mu_j' \gamma_j^{\mu_j' + 1}} - \frac{\mu_j' \gamma_j^{\Delta_j}}{(1 - \mu_j') \gamma_j^{\mu_j' + 1} - \mu_j' \gamma_j^{\mu_j' + 1}} + \frac{1 - \mu_j' \gamma_j^{\Delta_j}}{(1 - \mu_j') \gamma_j^{\mu_j' + 1} - \mu_j' \gamma_j^{\mu_j' + 1}} \right\} .
\] (D.16)

It is easy to see that (1) \( \lim_{\gamma \to \infty} \frac{\partial \tilde{L}_j}{\partial \gamma} < 0 \); (2) the terms inside \( \{ \} \) decreases w.r.t \( \gamma \in \Gamma_j \). Thus,

1. If \( \mu_j' + \delta_j(0) \geq 1 \) we can always find a unique non-negative solution \( \gamma_0 \) s.t. \( \mu_j' + \delta_j(\gamma_0) = 1 \). It is obvious that \( \lim_{\gamma \to \gamma_0^+} \tilde{L}_j = \infty \).

According to intermediate value theorem\(^5\), there exists a \( \gamma^* \) s.t. \( \frac{\partial \tilde{L}_j}{\partial \gamma} \bigg|_{\gamma = \gamma^*} = 0 \). And \( \gamma^* \) is just the \( \gamma^* \).

2. If \( \mu_j' + \delta_j(0) < 1 \) when \( \frac{\partial \tilde{L}_j}{\partial \gamma} \bigg|_{\gamma = 0} \leq 0 \), \( \gamma^* = 0 \); otherwise, \( \gamma^* \) is the one satisfies \( \frac{\partial \tilde{L}_j}{\partial \gamma} = 0 \).

Therefore, we have proved that \( L_j^\gamma \) always exists. Since at least one of \( \delta_j(\gamma^*) \) and \( \gamma^* \delta_j(\gamma^*) \) is positive, we know \( L_j^\gamma \) is positive. Finally, according to Lemma 2 of [31], we know for Bernoulli bandits, the pseudo regret can be written as:

\[
\text{Regret} = \sum_{i=2}^{K} \Delta_i E(T_i^{alg}).
\] (D.17)

During the first \( B \) rounds, if the budget spend on suboptimal arm \( j \) is spent on the optimal arm, the player can get more expected reward, which is \( \Omega(\frac{\Delta_j}{\tau_j} \ln B) \) as \( B \to \infty \). Summing this for all the arms, the regret is \( \Omega(\sum_{j=2}^{K} \Delta_j \ln B) \) as \( B \to \infty \).

Appendix E. Proof of the fractional KUBE in [27]

In the budgeted MAB with deterministic cost [27], the reward distribution of each arm has support in \([0, 1]\). The expected reward of arm \( i \) is denoted as \( \mu_i' \forall i \in [K] \). The pulling cost of arm \( i \in [K] \) is denoted as \( c_i \), which is deterministic. In [27], \( c_i \geq 1 \forall i \in [K] \). Denote \( c_{\text{min}} = \min_{i \in [K]} c_i \). W.l.o.g. set \( \arg \max_{i \in [K]} c_i' = 1 \). Define \( \Delta_i = c_i' - c_i \geq 0 \). Define \( \Delta_i' = c_i - c_i' \geq 2 \) (i.e., the \( \Delta_i \) in [27]). Define \( \Delta_i' = c_i - c_i' \geq 2 \) (i.e., the \( \Delta_i \) in [27]). Define the stopping time as \( T_B \).

The algorithm “fractional KUBE” proposed in [27] is: at each round, pull the arm with the maximum index \( \frac{r_{ii} + E_{ii}}{c_i} \), in which \( E_{ii} = \sqrt{\frac{2 \ln(d-1)}{m_i}} \). Denote the pulling time of arm \( i \) as \( T_i^{\text{KUBE}} \) when the algorithm fKUBE stops.

---

Under the setting that the cost of each arm is deterministic, the pulling time is at most \( \frac{B}{c_{\min}} \). Given \( \eta \in (0, 1) \), \( U_i^f(\eta) \) denotes 
\[
2 \ln \frac{B}{c_{\min}(1 - \eta)^2} \rho_c^2(0). \]
Therefore, we have

\[
T_i^{KBUBE} \leq 1 + \sum_{i = K + 1}^{B/c_{\text{min}}} \mathbb{1}(U_i = i) \leq \left[ U_i^f(\eta) \right] + \sum_{i = K + 1}^{B/c_{\text{min}}} \mathbb{1}(U_i = i; n_{ij} \geq \left[ U_i^f(\eta) \right])
\]

\[
\leq 1 + U_i^f(\eta) + \sum_{i = K + 1}^{B/c_{\text{min}}} \mathbb{1}\left( \frac{\tilde{T}_{ij} + \tilde{E}_{ij}}{c_i} \geq \frac{\tilde{T}_{ij} + \tilde{E}_{ij}}{c_1}, n_{ij} \geq \left[ U_i^f(\eta) \right] \right)
\]

\[
\leq 1 + U_i^f(\eta) + \sum_{i = K + 1}^{B/c_{\text{min}}} \mathbb{1}\left( \frac{\tilde{T}_{ij} + \tilde{E}_{ij}}{c_i} \geq \frac{\tilde{T}_{ij} + \tilde{E}_{ij}}{c_1}, n_{ij} \geq \left[ U_i^f(\eta) \right], \frac{\tilde{T}_{ij} + \tilde{E}_{ij}}{c_1} \leq \frac{\mu_i^f + \delta_i(0)}{c_i} \right)
\]

\[
+ \sum_{i = K + 1}^{B/c_{\text{min}}} \mathbb{1}\left( \frac{\tilde{T}_{ij} + \tilde{E}_{ij}}{c_i} \geq \frac{\tilde{T}_{ij} + \tilde{E}_{ij}}{c_1}, n_{ij} \geq \left[ U_i^f(\eta) \right], \frac{\tilde{T}_{ij} + \tilde{E}_{ij}}{c_1} \leq \frac{\mu_i^f + \delta_i(0)}{c_i} \right)
\]

\[
\leq 1 + U_i^f(\eta) + \sum_{i = K + 1}^{B/c_{\text{min}}} \mathbb{1}\left( \tilde{T}_{ij} + \tilde{E}_{ij} \geq \mu_i^f + \delta_i(0), n_{ij} \geq \left[ U_i^f(\eta) \right] \right) + \sum_{i = K + 1}^{B/c_{\text{min}}} \mathbb{1}\left( \tilde{T}_{ij} + \tilde{E}_{ij} \leq \mu_i^f \right)
\]

One can verify that \( P[\tilde{T}_{ij} + \tilde{E}_{ij} \leq \mu_i^f] \leq \frac{1}{(2\eta)^2} \) by Hoeffding inequality and union bound. Therefore,

\[
\mathbb{E} \sum_{i = K + 1}^{B/c_{\text{min}}} \mathbb{1}\left( \tilde{T}_{ij} + \tilde{E}_{ij} \leq \mu_i^f \right) \leq \sum_{i = K + 1}^{B/c_{\text{min}}} 1 \leq \frac{1}{2}
\]

(E.2)

On the other hand, if \( n_{ij} \geq U_i^f(\eta), \delta_i(0) - \tilde{E}_{ij} \geq \eta \delta_i(0) \). By Hoeffding inequality and union bound, we can obtain that

\[
P[\tilde{T}_{ij} + \tilde{E}_{ij} \geq \mu_i^f + \delta_i(0), n_{ij} \geq U_i^f(\eta)] \leq \left( \frac{B}{c_{\min}} \right)^{1 - \frac{2\rho_c^2(0)}{\eta}} \]

(E.3)

Therefore,

\[
\mathbb{E} \sum_{i = K + 1}^{B/c_{\text{min}}} \mathbb{1}\left( \tilde{T}_{ij} + \tilde{E}_{ij} \geq \mu_i^f + \delta_i(0), n_{ij} \geq U_i^f(\eta) \right) \leq \left( \frac{B}{c_{\min}} \right)^{2 - \frac{4\rho_c^2(0)}{\eta}}.
\]

(E.4)

By choosing \( \eta = \frac{1}{\sqrt{c_1}} \), we have

\[
\mathbb{E} \sum_{i = K + 1}^{B/c_{\text{min}}} \mathbb{1}\left( \tilde{T}_{ij} + \tilde{E}_{ij} \geq \mu_i^f + \delta_i(0), n_{ij} \geq U_i^f(\eta) \right) \leq 1.
\]

(E.5)

Therefore, we can conclude that

\[
\mathbb{E}[T_i^{KBUBE}] \leq \left( 3 + 2 \sqrt{2} \right) \ln \frac{B}{c_{\min}(c_1\Delta_i)^2} + \frac{5}{2}
\]

(E.6)

In [27], the player can keep pulling until none of the arms are feasible. Thus,

\[
B - c_{\min} \leq \mathbb{E}\left[ \sum_{j = 1}^{K} c_1 \mathbb{1}(L_j = i) \right] = \mathbb{E}\left[ \sum_{j = 1}^{K} (c_j - c_1 + c_1) \mathbb{1}(L_j = i) \right]
\]

\[
= \mathbb{E}[T_B]c_1 + \sum_{j = 2}^{K} \mathbb{E}[T_B]c_1 \mathbb{1}(L_j = i) = \mathbb{E}[T_B]c_1 + \sum_{j = 2}^{K} \Delta_j^f \mathbb{E}[T_i^{KBUBE}]\]

\[
\leq \mathbb{E}[T_B]c_1 + \sum_{\Delta_j^f > 0} \Delta_j^f \left( 3 + 2 \sqrt{2} \right) \ln \frac{B}{c_{\min}(c_1\Delta_i)^2} + \frac{5}{2}.
\]

Therefore, we have

\[
\frac{B}{c_1} - \mathbb{E}[T_B] \leq \frac{c_{\min}}{c_1} \sum_{\Delta_j^f > 0} \Delta_j^f \left( 3 + 2 \sqrt{2} \right) \ln \frac{B}{c_{\min}(c_1\Delta_i)^2} + \frac{5}{2}.
\]

(E.7)
According to inequality (31) in [27],
\[ R^{\text{KUBE}} \leq \frac{B}{c_1} - \mathbb{E}(T_i)|u_i' + \sum_{c \neq A(t)} \Delta^i_c| \mathbb{E}(T_i) \]
\[ \leq \frac{\mu_i'^c \Delta_{\text{min}}}{c_1} + \sum_{c \neq A(t)} \frac{\mu_i'^c A(t) (3 + 2 \sqrt{2}) \ln \frac{B}{\mu_i'^c A(t)}}{c_1 (\Delta^i_c)^2} + \frac{5}{2} \sum_{c \neq A(t)} \Delta^i_c(3 + 2 \sqrt{2}) \ln \frac{B}{\mu_i'^c A(t)} + \frac{5}{2} \frac{\mu_i'^c \Delta_{\text{min}}}{c_1} \]
\[ \leq (3 + 2 \sqrt{2}) \ln \frac{B}{c_{\text{min}}} \left( \sum_{c \neq A(t)} \frac{\mu_i'^c \Delta_{\text{min}}}{c_1} + \sum_{c \neq A(t)} \Delta^i_c + \frac{\mu_i'^c \Delta_{\text{min}}}{c_1} \right). \]
which is certainly smaller than Theorem 7 (in Page 16) of [27].

Appendix F. Improved Results for Budgeted Thompson Sampling [31]

Notations 1 For any suboptimal arm \( i \geq 2 \) and for any \( \gamma \geq 0 \), \( x_i', x_i', y_i', y_i' \) are the constant to be determined satisfying the following constraints:
\[ \frac{\mu_i'}{\mu_i'} \leq \frac{x_i'}{x_i'} \leq \frac{y_i'}{y_i'} \leq \frac{\mu_i'}{\mu_i'} + \delta(y) \]
\[ x_i', x_i', y_i', y_i' \in (0, 1); \quad \mu_i' < x_i' < y_i' < \mu_i' + \delta(y); \quad \mu_i' > x_i' > y_i' > \mu_i' - \gamma \delta(y). \] (F.1)

(2) Define \( L_n(\tau_B) \) as follows:
\[ L_n(\tau_B) = \frac{\ln(\tau_B)}{\sup_{y, \alpha} \min[kl(x_i', y), kl(x_i', y) \alpha]}. \] (F.2)

(3) We introduce some other notations:
\[ k_i(t) = S_i'(t) + F_i'(t); \quad \bar{\mu}_i'(t) = \frac{S_i'(t)}{S_i'(t) + F_i'(t) + 1}; \quad \tilde{\mu}_i'(t) = \frac{S_i'(t)}{S_i'(t) + F_i'(t) + 1}; \]
\[ E_i^\alpha(t): \mu_i' \leq x_i'; \quad E_i^{\alpha'}(t): \mu_i' \leq y_i'; \quad E_i^\beta(t) = E_i^\alpha(t) \cap E_i^{\alpha'}(t); \quad E_i^\gamma(t) : \frac{\delta^i(t)}{\gamma^i(t)} \leq \frac{y_i'}{y_i'}, \]
\[ \text{Analysis of BTS for Bernoulli Budgeted MAB} \]

According to Lemma 4, we only need to bound \( \mathbb{E}[T_i] \) for any \( i \geq 2 \), where \( T_i \) is still defined as the number of pulling time of arm \( i \) from round 1 to round \( \tau_B \).

Similar to the analysis in [31], we can get that
\[ \mathbb{E}[T_i] = \mathbb{E} \left[ \sum_{r=1}^{T_n} \mathbf{1}[I_r = i] \right] = \sum_{r=1}^{T_n} \mathbb{P}[I_r = i, E_i^{\alpha}(t)] \quad \text{and} \quad \sum_{r=1}^{T_n} \mathbb{P}[I_r = i, E_i^{\alpha'}(t)] + \sum_{r=1}^{T_n} \mathbb{P}[I_r = i, E_i^\beta(t)]. \] (F.4)
The first two terms of (F.4) can be bounded by the following lemma:

**Lemma 12.** For any \( i \geq 2 \), we have the following two results:

(1) \[ \sum_{r=1}^{T_n} \mathbb{P}[I_r = i, E_i^{\alpha}(t)] \leq 1 + \frac{1}{kl(x_i', \mu_i')} + \frac{1}{kl(y_i', \mu_i')}; \quad \text{(II)} \quad \sum_{r=1}^{T_n} \mathbb{P}[I_r = i, E_i^{\alpha'}(t)] \leq L_n(\tau_B) + 2. \]

The proof of Lemma 12 will be provided later.

Define \( w_i \) as follows:
\[ w_i = x_i' - y_i', \]
from which we can verify that
\[ x_i' - y_i' = \frac{\bar{\mu}_i' - w_i}{\bar{\mu}_i' + w_i}. \] (F.7)

Given \( \epsilon \in (0, 1) \), we can always get a group of proper \( x_i', x_i', y_i', y_i' \) s.t.
\[ kl(x_i', y_i') = \frac{kl(\mu_i', y_i')}{1 + \epsilon} = \frac{kl(\mu_i', \mu_i' + \delta(y))}{(1 + \epsilon)^2}; \quad kl(x_i', y_i') = \frac{kl(y_i', \mu_i')}{1 + \epsilon} = \frac{kl(\mu_i', \mu_i' - \gamma \delta(y))}{(1 + \epsilon)^2}. \] (F.8)

Temporally denote the \( y_i' \) and \( y_i' \) in (F.8) as \( y_i'(\epsilon) \) and \( y_i'(\epsilon) \). Denote the \( w_i \) in (F.6) as \( w_i(\epsilon) \). We can get that
\[ \frac{dy_i'(\epsilon)}{de} = \frac{y_i'(\epsilon) - \mu_i'}{y_i'(\epsilon) - \mu_i'}, \quad \frac{dy_i'(\epsilon)}{de} = \frac{y_i'(\epsilon) - \mu_i'}{y_i'(\epsilon) - \mu_i'}. \] (F.9)
Appendix G. Detailed Analysis on Budget-UCB Algorithm

Applying total difference to both sides of (F.6), we can obtain that

$$\frac{d \omega_i(e)}{d e} = \frac{\mu'_i - w_i(e)}{\gamma'_i(e) + \mu'_i + \gamma_0 \delta(y)} - \frac{\mu'_i + w_i(e)}{\gamma'_i(e) + \gamma_0 \delta(y)}.$$  \hspace{1cm} (F.10)

Since \( \lim_{e \to 0} \gamma'_i(e) = \mu'_i + \delta(y) \) and \( \lim_{e \to 0} \gamma'_i(e) = \mu'_i - \gamma_0 \delta(y) \), by (F.9) and (F.10), we can verify that

$$\frac{d \omega_i(e)}{d e} \bigg|_{e=0} = \frac{\mu'_i (1 - \gamma_0 \delta(y)) (1 - \mu'_i + \gamma_0 \delta(y)) (1 - \mu'_i + \gamma_0 \delta(y)) k_i \mu'_i + \gamma_0 \delta(y)}{\mu'_i + \gamma_0 \delta(y)}.$$  \hspace{1cm} (F.11)

which is obviously a positive finite constant. As a result, when \( \epsilon \in (0, 1) \), we can say that \( O(\omega_i^{-1}(\epsilon)) \leq O(\epsilon^{-1}) \).

Therefore, by substituting the \( \epsilon_i(y) \) in the Step 3 of [31] to the \( w_i(e) \), we can obtain that when the \( \epsilon \) is chosen small enough such that \( w_i(e) < 0.5(1 - \mu'_i) \), we have\( \sum_{t=1}^{\tau_1} \mathbb{P}[I_t = i, E_i^{A}(t)] \leq O(\epsilon_i^{-c}) \).

According to the above derivations, we can get the improved regret in Theorem 7.

Finally we will prove Lemma 12:

(Proof of (I) in Lemma 12): Let \( s_i \) be the \( k \)-th time that arm \( i \) is pulled. Define \( s_0 = 0 \).

$$\sum_{i=1}^{k} \mathbb{P}[I_t = i; E_i^{A}(t)] \leq \mathbb{E} \left[ \sum_{i=1}^{k} \sum_{s_i = 0}^{s_i + 1} \mathbb{I}[I_t = i|E_i^{A}(t)] \right]$$

$$= \mathbb{E} \left[ \sum_{i=1}^{k} \sum_{s_i = 0}^{s_i + 1} \mathbb{I}[I_t = i] \right] = \mathbb{E} \left[ \sum_{i=1}^{k} \mathbb{I}[E_i^{A}(s_i + 1)] \right] \leq 1 + \mathbb{E} \left[ \sum_{i=1}^{k} \mathbb{I}[E_i^{A}(s_i + 1)] \right] \leq 1 + \sum_{i=1}^{k} \mathbb{P}[\mu'_i(t) \geq \gamma'_i] + \sum_{i=1}^{k} \mathbb{P}[\mu'_i(t) \leq \gamma'_i]$$

$$\leq 1 + \sum_{i=1}^{k} \exp[-k k_i \gamma'_i] + \sum_{i=1}^{k} \exp[-k k_i \mu'_i] \leq 1 + \frac{1}{k} \sum_{i=1}^{k} \exp[-k k_i \gamma'_i] + \frac{1}{k} \sum_{i=1}^{k} \exp[-k k_i \mu'_i]. \quad \Box$$  \hspace{1cm} (F.12)

(Proof of (II) in Lemma 12) First, given the history until round \( t - 1 \) (denoted as \( \mathcal{F}_{t-1} \)), \( k_i(t) \) is a deterministic number rather than variable. Then we have

$$\mathbb{P}[I_t = i; E_i^{A}(t), E_i^{A}(t)[\mathcal{F}_{t-1}] \leq \mathbb{P}[I_t = i; \theta'_i(t) > \gamma'_i, E_i^{A}(t)[\mathcal{F}_{t-1}] + \mathbb{P}[I_t = i; \theta'_i(t) < \gamma'_i, E_i^{A}(t)[\mathcal{F}_{t-1}]$$

$$\leq \mathbb{P}[\Theta'_i(t) > \gamma'_i \mu'_i(t) \leq \gamma'_i, \mathcal{F}_{t-1}] + \mathbb{P}[\Theta'_i(t) < \gamma'_i \mu'_i(t) \geq \gamma'_i, \mathcal{F}_{t-1}]$$

$$\leq \mathbb{P}[\text{Beta} \mu'_i(t)(k_i(t) + 1) + 1, (1 - \mu'_i(t))(k_i(t) + 1) > \gamma'_i]$$

$$+ \mathbb{P}[\text{Beta} \mu'_i(t)(k_i(t) + 1) + 1, (1 - \mu'_i(t))(k_i(t) + 1) < \gamma'_i]$$

$$\leq F_{k_i(t)+1, \gamma'_i}^{-1}(F_{k_i(t)+1, \gamma'_i}) + (1 - F_{k_i(t)+1, \gamma'_i}^{-1}(F_{k_i(t)+1, \gamma'_i}))$$

$$\leq \exp[-k_k(t)(x'_i + \gamma'_i)] + \exp[-k_k(t)(x'_i - \gamma'_i)].$$

As a result, when \( k_i(t) > L_\tau(\tau_B) \), we have

$$\mathbb{P}[I_t = i; E_i^{A}(t), E_i^{A}(t), E_i^{A}(t)] \geq \frac{2}{\tau_B} \quad \hspace{1cm} (F.13)$$

Temporally let \( s \) denote the maximum round that \( k_i(t) \leq L_\tau(\tau_B) \). As a result, we have

$$\sum_{i=1}^{K} \mathbb{P}[I_t = i; E_i^{A}(t), E_i^{A}(t), E_i^{A}(t)]$$

$$\leq \mathbb{E} \left[ \sum_{i=1}^{K} \mathbb{P}[I_t = i; E_i^{A}(t), E_i^{A}(t)] + \sum_{i=1}^{K} \mathbb{P}[I_t = i; E_i^{A}(t), E_i^{A}(t), E_i^{A}(t)] \right]$$

$$\leq \mathbb{E} \left[ \sum_{i=1}^{K} \mathbb{P}[I_t = i] + \sum_{i=1}^{K} \mathbb{P}[I_t = i; E_i^{A}(t), E_i^{A}(t)] \right]$$

$$\leq L_\tau(\tau_B) + 2. \quad \Box$$  \hspace{1cm} (F.14)

Appendix G. Detailed Analysis on Budget-UCB Algorithm

Throughout this subsection, (A) let \( T_i \) denote the number of pulling rounds of arm \( i \) from round 1 to round \( \tau_B \) with Budget-UCB algorithm; (B) denote \( \delta_i(t) \) as \( \delta_i \); (C) temporally omit the “\( \alpha \)” from \( E_i^{A}[t] \); i.e., denote \( E_i^{A}[t] \) as \( E_i[t] \) for any \( i \in [K] \).
(S1) Similar to the derivations of (S1) in Subsection Appendix C.1, we have
\[
\mathbb{E}[T_i] \leq \left[U'\right] + \sum_{i=1}^{n} \left[ \sum_{r=1}^{n_r} \mathbb{P} \left[ \frac{\bar{c}_{ij}}{\bar{c}_{ij}} + H_{i,j} \leq \frac{\mu_i'}{\mu_i'} \right] + \sum_{r=1}^{n_r} \mathbb{P} \left[ \frac{\bar{c}_{ij}}{\bar{c}_{ij}} + H_{i,j} \geq \frac{\mu_i' + \delta_i}{\mu_i'} - \delta_i, n_{ij} \geq \left[U'\right] \right] \right],
\]
where
\[
U_i' = \frac{(3 + 2 \sqrt{2}) \ln \tau_i}{\delta_i^2}; \quad H_{i,j} = \frac{\epsilon_{ij}}{\bar{c}_{ij}} + \frac{\epsilon_{ij}}{\bar{c}_{ij}} \min[\bar{c}_{ij} - \epsilon_{ij}, 0].
\]
Next, we will bound sum of term \(a\) and term \(b\) respectively.

(S2) Bound the sum of term \(a\) in (G.1): It is obvious that
\[
\mathbb{P} \left[ \frac{\bar{c}_{ij}}{\bar{c}_{ij}} + H_{i,j} \leq \frac{\mu_i'}{\mu_i'} \right] \leq \mathbb{P} \left[ \bar{c}_{ij} - \mu_i' \geq \epsilon_{ij} \right] + \mathbb{P} \left[ \frac{\bar{c}_{ij}}{\bar{c}_{ij}} + H_{i,j} \leq \frac{\mu_i'}{\mu_i'} - \bar{c}_{ij} - \mu_i' < \epsilon_{ij} \right].
\]
If the event expressed by the last term of (G.3) holds, we have\(^6\)
\[
\bar{r}_{ij} - \mu_i' = \frac{(\bar{c}_{ij} - \mu_i') \mu_i'}{\bar{c}_{ij}} \leq \frac{\epsilon_{ij}}{\bar{c}_{ij}} \min[\bar{c}_{ij} - \epsilon_{ij}, 1] \leq \frac{\epsilon_{ij}}{\bar{c}_{ij}} \max[\bar{c}_{ij} - \epsilon_{ij}, 0].
\]
Therefore, at least one of the following events is true:
(i) \(\bar{r}_{ij} - \mu_i' \leq -\epsilon_{ij}\) or (ii) \(\frac{\epsilon_{ij}}{\bar{c}_{ij}} \min[\bar{c}_{ij} - \epsilon_{ij}, 1] \leq \frac{\epsilon_{ij}}{\bar{c}_{ij}} \max[\bar{c}_{ij} - \epsilon_{ij}, 0]\).

Given \(\bar{c}_{ij} - \mu_i' < \epsilon_{ij}\), (ii) in (G.5) implies that \(\bar{r}_{ij} + \epsilon_{ij} \leq \mu_i'\). Therefore, (G.3) is upper bounded by \(\mathbb{P}[\bar{r}_{ij} + \epsilon_{ij} \leq \mu_i'] + \mathbb{P}[\bar{c}_{ij} - \mu_i' \geq \epsilon_{ij}]\). By Hoeffding’s inequality and union bound, we have
\[
\mathbb{P}[\bar{r}_{ij} + \epsilon_{ij} \leq \mu_i'] + \mathbb{P}[\bar{c}_{ij} - \mu_i' \geq \epsilon_{ij}] \leq 2(t - 1)^{-3}.
\]
As a result, the sum of term \(a\) in (G.1) can be bounded by
\[
\sum_{i=1}^{n} \left[ \sum_{r=1}^{n_r} \mathbb{P} \left[ \frac{\bar{c}_{ij}}{\bar{c}_{ij}} + H_{i,j} \leq \frac{\mu_i'}{\mu_i'} \right] \right] \leq 2 \sum_{i=1}^{n} \frac{1}{t_i^3} \leq 1.
\]

(S3) Bound the sum of term \(b\) in (G.1): For any suboptimal arm \(i\), we can find a shadow expected reward \(\bar{\mu}_i^*\) and a shadow expected cost \(\bar{\mu}_i^*\) of the optimal arm related to both \(\mu_i'\) and \(\mu_i'\): \(\forall i \neq i: \bar{\mu}_i^* = \mu_i + \delta_i; \bar{\mu}_i^* = \mu_i - \delta_i\). We can verify
\[
\bar{r}_{ij} = \frac{\epsilon_{ij}}{\bar{c}_{ij}} - \frac{\epsilon_{ij}}{\bar{c}_{ij}} = \frac{\bar{c}_{ij} - \bar{c}_{ij}}{\bar{c}_{ij}} + \frac{\bar{c}_{ij} - \mu_i'}{\bar{c}_{ij}}.
\]
Thus, \(\bar{c}_{ij} - \mu_i' \geq \bar{c}_{ij} - \mu_i'\) holds implies that
\[
\bar{r}_{ij} - \bar{\mu}_i^* \leq \epsilon_{ij}; \quad \bar{r}_{ij} + \epsilon_{ij} \geq -\delta_i.
\]
Given \(T_i \geq U_i' = \frac{(3 + 2 \sqrt{2}) \ln \tau_i}{\delta_i^2}\), one can verify that \(\delta_i - \epsilon_{ij} \geq \frac{\delta_i}{\sqrt{2}}\). Therefore,
\[
\sum_{i=1}^{n} \left[ \sum_{r=1}^{n_r} \mathbb{P} \left[ \bar{c}_{ij} - \mu_i' \leq \epsilon_{ij}, n_{ij} \geq \left[U_i'\right] \right] \right] \leq \sum_{i=1}^{n} \left[ \sum_{r=1}^{n_r} \mathbb{P} \left[ \bar{c}_{ij} - \mu_i' \leq -\delta_i \sqrt{2 + 1}, n_{ij} = 1 \right] \right] \leq (\tau_i) \exp[-2U_i' \left( \frac{\delta_i}{\sqrt{2 + 1}} \right)^2] = 1. (\text{Obtained by Chernoff-Hoeffding bound.})
\]
Similarly, we can also obtain that
\[
\sum_{r=1}^{n_r} \mathbb{P} \left[ \bar{r}_{ij} + \epsilon_{ij} \geq -\delta_i', n_{ij} \geq U_i' \right] \leq 1.
\]
Therefore,
\[
\mathbb{E}[T_i] \leq \frac{(3 + 2 \sqrt{2}) \ln \tau_i}{\delta_i^2} + 3.
\]
The regret can be obtained by applying Lemma 4.

\(^6\)For the boundary case that \(\bar{c}_{ij} = 0\), since \(\epsilon_{ij} > 0\) when \(t \geq K + 1\), we know that \((\bar{r}_{ij}/\bar{c}_{ij}) + H_{i,j} \to \infty\), which makes the last \(\mathbb{P}[\cdot]\) of G.3 zero. Therefore, we can safely consider the case \(\bar{c}_{ij} > 0\) only.

\(^7\)Note that the boundary case \(\bar{c}_{ij} = 0\) belongs to the events expressed by (G.10).