Budgeted Multi-armed Bandits with Multiple Plays (Full Version)*

Yingce Xia[1], **Tao Qin**[2], **Weidong Ma**[2], **Nenghai Yu**[1] and **Tie-Yan Liu**[2]
[1]University of Science and Technology of China [2]Microsoft Research Asia
yingce.xia@gmail.com; {taoqin,weima,tie-yan.liu}@microsoft.com; ynh@ustc.edu.cn

Abstract

We study the multi-play budgeted multi-armed bandit (MP-BMAB) problem, in which pulling an arm receives both a random reward and a random cost, and a player pulls $L(\geq 1)$ arms at each round. The player targets at maximizing her total expected reward under a budget constraint B for the pulling costs. We present a multiple ratio confidence bound policy: At each round, we first calculate a truncated upper (lower) confidence bound for the expected reward (cost) of each arm, and then pull the L arms with the maximum ratio of the sum of the upper confidence bounds of rewards to the sum of the lower confidence bounds of costs. We design a 0-1 integer linear fractional programming oracle that can pick such the L arms within polynomial time. We prove that the regret of our policy is sublinear in general and is log-linear for certain parameter settings. We further consider two special cases of MP-BMABs: (1) We derive a lower bound for any consistent policy for MP-BMABs with Bernoulli reward and cost distributions. (2) We show that the proposed policy can also solve conventional budgeted MAB problem (a special case of MP-BMABs with $L = 1$) and provides better theoretical results than existing UCB-based pulling policies.

1 Introduction

Multi-armed bandits (MAB) are 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 her cumulated reward. Multiple real world applications have been modeled as MAB problems, such as auction mechanism design [Mohri and Munoz, 2014], search advertising [Tran-Thanh *et al.*, 2014], UGC mechanism design [Ghosh and Hummel, 2013], and personalized recommendation [Li *et al.*, 2010]. Many policies have been designed for MAB problems and studied from both theoretical and empirical perspectives, including UCB1, ϵ_n-GREEDY [Auer *et al.*, 2002], LinRel [Auer, 2003], UCB-V [Audibert *et al.*, 2009], DMED [Honda and Takemura, 2010], and KL-UCB [Garivier and Cappé, 2011]. A good survey on MAB can be found in [Bubeck and Cesa-Bianchi, 2012].

Recently, budgeted MABs have attracted much research attention. In budgeted MABs, playing an arm needs to pay a cost while receiving a reward, and the player targets at maximizing her cumulative reward under a budget constraint for the total costs. Different settings of costs have been studied in budgeted MABs. Deterministic costs were studied in [Tran-Thanh *et al.*, 2012]. [Vanchinathan *et al.*, 2015] attacked an MAB related problem by taking both deterministic costs and the diversity of the selected items into consideration. UCB based algorithms were adapted to the random discrete cost setting [Ding *et al.*, 2013] and random continuous cost setting [Xia *et al.*, 2015a]. Thompson sampling algorithm for budgeted MAB was studied in [Xia *et al.*, 2015b]. Besides minimizing the regret, the best arm identification problem for budgeted MAB was studied in [Xia *et al.*, 2016].

Multiple-play MABs, in which the player pulls multiple arms at each round, have been studied in conventional settings without considering budget [Anantharam *et al.*, 1987; Agrawal *et al.*, 2010; Komiyama *et al.*, 2015; Liu and Zhao, 2010; Chen *et al.*, 2013]. In some applications, a decision maker needs to take multiple actions at each round and consider a budget constraint. For example, consider an advertiser who creates an ad campaign to promote her products in a search engine. To participate in search ad auctions, she needs to choose multiple keywords for her campaign and set a monthly/quarterly budget. Since each keyword (together with a bid price) can be regarded as an arm [Ding *et al.*, 2013], this keyword selection and bid optimization problem can be modeled as a budgeted MAB with multiple plays. In this work, we study this new setting, the *Multiple-Play Budgeted Multi-armed Bandit* (denote as MP-BMAB) problem. For simplicity, we refer the simple case of the budgeted MAB, playing a single arm at each round, as *Single-Play Budgeted Multi-armed Bandit* (denoted as SP-BMAB).

Consider a bandit with K arms in total and the player needs to pull $L \geq 1$ different arms at each round. There are $\binom{K}{L}$ different ways of pulling L different arms, and the number could be of order $O(2^K)$ in the worst case. Therefore, we need to carefully design policies that can efficiently deal with large number of possible pullings. Our work can be summarized from the following three aspects:

*This work was conducted at Microsoft Research Asia.

Policy Design: Intuitively, a good policy for MP-BMABs should try to pull the L arms with the maximum ratio of the sum of the expected rewards to the sum of the expected costs. Since the reward and cost distributions of all the arms are unknown, the policy needs to allocate necessary explorations to all the arms. We design an efficient policy for the MP-BMAB problem, called *Multiple Ratio Confidence Bound* policy (denoted as MRCB), which works as follows. For each arm, we introduce a truncated upper confidence bound for the estimated expected reward and a truncated lower confidence bound for the estimated expected cost. A hyper parameter is introduced to the confidence bound, which brings flexibility to the policy. At each round, we pull the L arms with the maximum ratio of the sum of the upper bounds of rewards to the sum of the lower bounds of costs. How to find such L arms with the maximum ratio is an 0-1 *integer linear fractional program* (denoted as 0-1 ILFP) [Seerengasamy and Jeyaraman, 2013]. We design an efficient algorithm that can find the optimal solution of the 0-1 ILFP in our setting within polynomial time.

Theoretical Analysis: We conduct theoretical analysis on MRCB, and show that it enjoys a sublinear regret bound with respect to budget B. By properly setting the hyper parameter, we show that the policy theoretically achieves a log-linear regret. Comparing with conventional MABs, there are two challenges to analyze MRCB: (1) One needs to pull L different arms at each round (for simplicity, we say any L different arms constitute a *super arm*) and there are exponential number of possible super arms, which might bring the combinatorial number into the regret bound and make the bound very loose. (2) The randomness of both the rewards and costs brings difficulties when decomposing the probabilities that suboptimal super arms are pulled[1]. To address the first challenge, we carefully divide the exponential number of suboptimal super arms into K subsets and design intermediate events related to the pulling time of each super arm in each subset. Doing so we can eliminate the affects brought by the exponential number of super arms. To address the second one, we introduce the δ-gap in Eqn.(11a), based on which we can separate the ratio related terms which depend on both rewards and costs into terms that depend on rewards only and costs only.

Special Cases: We further study two special cases of MP-BMABs. First, for Bernoulli MP-BMABs (whose rewards and costs are either 0 or 1), we give a lower bound to any consistent policy and show that our proposed policy can match the lower bound in terms of the order of B. Second, for conventional budgeted MABs (i.e., SP-BMABs), we show that our policy can be directly applied and achieves a better regret bound than existing UCB based policies [Ding *et al.*, 2013]. We also provide a lower bound for SP-BMABs, which is missing in the literature.

2 The Problem

An MP-BMAB problem can be described as follows. Given a slot machine with K arms ($K \geq 2$), at each round, the player needs to pull $L(\geq 1)$ different arms of the bandit. Denote the set of arms pulled at round t as I_t. For each pulled arm $i \in [K]$ at round t (let $[K]$ denote the set $\{1, 2, \cdots, K\}$), she needs to pay a random cost $c_i(t)$ and receives a random reward $r_i(t)$. Both $c_i(t)$ and $r_i(t)$ are drawn from distributions supported in $[0, 1]$. We study the *semi-bandit* setting [Kveton *et al.*, 2015], in which the player can only observe $r_i(t)$ and $c_i(t)$ for pulled arms, i.e., for all $i \in I_t$. The player can keep pulling until her budget, B, runs out. B is a positive number and does not need to be known to the player in advance.

Following the common practice in standard MABs, 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 rounds are independently drawn from the same distribution with expectation μ_i^r (and μ_i^c). For ease of reference, denote the vector $(\mu_1^r, \mu_2^r, \cdots, \mu_K^r)$ as μ^r, and so for μ^c. Note that we do not assume that the rewards of an arm are independent of its costs. Without loss of generality, we assume $0 < \mu_i^r, \mu_i^c < 1$ for all $i \in [K]$. The player wants to minimize the regret, which is usually defined as the differences between R^*, the maximum expect cumulative reward that a pulling policy can obtain when the reward/cost distributions of all the arms are known, and the expected reward that a policy can obtain, both under the budget constraint. Mathematically,

$$\text{Regret} = R^* - \mathbb{E}\sum\nolimits_{t=1}^{\infty}\sum\nolimits_{i\in I_t} r_i(t)\mathbb{I}\{B_t \geq 0\}, \qquad (1)$$

where B_t is the remaining budget at round t, i.e., $B_t = B - \sum_{s=1}^{t}\sum_{i\in I_s} c_i(s)$, and $\mathbb{I}\{\cdot\}$ is the indicator function. $\mathbb{I}\{E\} = 1$ if the event E is true; otherwise, 0.

3 Pulling Policy

It is hard to find the optimal policy for MP-BMABs. Even for a simplified setting, in which the reward and cost of each arm are deterministic and $L = 1$, the problem is an unbounded knapsack problem, which is NP-hard [Lueker, 1975]. For the semi-bandit setting, this problem becomes even harder. To solve the MP-BMAB problem, in this section, we first consider a simple case with known reward and cost distributions for all the arms, and show that a simple greedy policy $\mathcal{M}_g$ can obtain almost the same expected reward as R^*. Then we design a pulling policy for the setting with unknown reward and cost distributions by leveraging $\mathcal{M}_g$.

3.1 $\mathcal{M}_g$ for Known Distributions

Remind that any L different arms from the K candidates constitute a *super arm*. Let $\mathcal{C}_L^K$ denote the set of all the super arms, which is mathematically defined as follows.

$$\{\{j : x_j = 1\} | \textstyle\sum_{j=1}^{K} x_j = L; x_j \in \{0, 1\}\ \forall j \in [K]\}.$$

Let I_* denote the super arm defined as follows:

$$I_* = \text{argmax}_{I\in\mathcal{C}_L^K}(\textstyle\sum_{k\in I}\mu_k^r)/(\sum_{k\in I}\mu_k^c). \qquad (2)$$

Without loss of generality, assume I_* is unique. Define ϱ_L^* as $(\sum_{k\in I_*}\mu_k^r)/(\sum_{k\in I_*}\mu_k^c)$.

The greedy policy $\mathcal{M}_g$ is shown in Algorithm 1. Lemma 1 shows that $\mathcal{M}_g$ is close to the optimal policy for the case with known reward/cost distributions, and therefore we call I_* the *nearly-optimal* super arm.

[1]The super arms which do not have the maximum ratio of the sum of the expected rewards to the sum of expected costs are suboptimal.

Algorithm 1: $\mathcal{M}_g$ for Known Distributions

1 *Input*: The reward and cost distributions of the K arms; the budget B; $L \in [K]$;
2 For any arm $i \in [K]$, calculate the expected reward μ_i^r and expected cost μ_i^c; find the I_* of the bandit in (2);
3 Keep pulling the L arms in I_*, until the budget runs out.

Lemma 1 *When the reward and cost distributions of all the arms are known, we have $R^* \leq (B+L)\varrho_L^*$ and the expected reward of $\mathcal{M}_g$ is at least $(B-L)\varrho_L^*$.*

Due to space limitations, we leave the proof of Lemma 1 to Appendix A. Lemma 1 tells that the gap between R^* and the expected reward of $\mathcal{M}_g$ is at most $2L\varrho_L^*$, which is very small when B is sufficiently large.

Step 2 of Algorithm 1 needs to find the I_* defined in (2), which is actually a 0-1 *Integer Linear Fractional Programming* problem defined as follows.

$$\max \left(\sum_{i\in I} a_i\right)/\left(\sum_{i\in I} b_i\right) \quad \text{s.t. } I \in \mathcal{C}_L^K, \tag{3}$$

where a and b are K-element vectors with the i-th element $a_i > 0, b_i \geq 0$ for any $i \in [K]$. We design a 0-1 *ILFP Oracle* $\mathcal{O}(a, b, L)$ that can efficiently solve the optimization problem in (3). The oracle is shown in Algorithm 2.

Algorithm 2: 0-1 ILFP Oracle $\mathcal{O}(a, b, L)$

1 *Input*: Vectors a and b with $a_i > 0, b_i \geq 0\, \forall i \in [K]$; $L \in [K]$;
2 *Boundary Cases*: Denote $Z_0 = \{i | b_i = 0, \forall i \in [K]\}$. *If* $|Z_0| \geq L$, then randomly return L elements in Z_0; *Else if* L is 1, return $\text{argmax}_i(a_i/b_i)$ for any $i \in [K]$ directly; *Else*, go to the next step;
3 Solve the LP problem marked with $(\triangle)$ by Interior Point Method. Denote the solution as y^* and z^*.
$$\max\ a^T y \ \text{ s.t. } \ \sum_{i=1}^K y_i - Lz = 0;\ b^T y = 1; \qquad (\triangle)$$
$$z \geq 0;\ 0 \leq y_i \leq z\ \forall i \in [K];$$
4 Let $\mathcal{I} = \{i | y_i^* = z^*; i \in [K]\}, \mathscr{F} = \{i | 0 < y_i^* < z^*; i \in [K]\}$; If $|\mathcal{I}| = L$, *return* $\mathcal{I}$; otherwise, pick any $L - |\mathcal{I}|$ elements from $\mathscr{F}$ forming $\mathscr{F}'$ and *return* $\mathcal{I} \cup \mathscr{F}'$.

Lemma 2 *The $\mathcal{O}(a, b, L)$ in Algorithm 2 can output the optimal solution of* (3) *within polynomial time*[2].

The proof of Lemma 2 is constructive: (1) Relax the 0-1 integer constraints to continuous ones, (2) solve the relaxed linear fractional programming, and then (3) convert the fractional solutions to integer ones. Complete proof is in Appendix B.

3.2 Multiple Ratio Confidence Bound Policy

Now we turn to the MP-BMAB problem with unknown reward/cost distributions. We can only observe the rewards/costs of the pulled arms at each round. Our idea is simple and straightforward: We estimate the expected reward/cost of each arm using historical observations and then apply Algorithm 2 with estimated expected rewards/costs as input to select the pulled arms at each round.

For any $i \in [K]$, let $T_i(t)$, $\hat{\mu}_i^r(t)$, $\hat{\mu}_i^c(t)$ and $\mathcal{E}_{i,t}^\kappa$ denote the number of pulling rounds, the empirical average reward and cost, and a confidence term of arm i at round t respectively:

$$\begin{aligned} T_i(t) &= \sum_{s=1}^t \mathbb{I}\{i \in I_s\},\ \hat{\mu}_i^r(t) = \frac{1}{T_i(t)} \sum_{s=1}^t r_i(s)\mathbb{I}\{i \in I_s\}, \\ \hat{\mu}_i^c(t) &= \frac{1}{T_i(t)} \sum_{s=1}^t c_i(s)\mathbb{I}\{i \in I_s\},\ \mathcal{E}_{i,t}^\kappa = \sqrt{\frac{\kappa \ln(t-1)}{T_i(t-1)}}, \end{aligned} \tag{4}$$

where κ is a positive hyper parameter, which brings flexibility[3] to our policy.

Note that for each arm, we do not directly replace the expected reward and cost by the empirical average reward and cost. Instead, we take the uncertainty of the estimation into consideration. Define $\tilde{\mu}_i^r(t)$ and $\tilde{\mu}_i^c(t)$ as the truncated upper confidence bound for the empirical average reward (see (5)) and truncated lower confidence bound for the empirical average cost (see (6)) respectively.

$$\tilde{\mu}_i^r(t) = \min\{\hat{\mu}_i^r(t-1) + \mathcal{E}_{i,t}^\kappa, 1\}; \tag{5}$$
$$\tilde{\mu}_i^c(t) = \max\{\hat{\mu}_i^c(t-1) - \mathcal{E}_{i,t}^\kappa, 0\}. \tag{6}$$

Our proposed policy, *Multiple Ratio Confidence Bound* policy (briefly denoted as MRCB), is shown in Algorithm 3, in which $\tilde{\mu}^r(t)$ is a K-dimensional vector[4] with the i-th element $\tilde{\mu}_i^r(t)$, and so for $\tilde{\mu}^c(t)$.

Algorithm 3: Multiple Ratio Confidence Bound (MRCB)

1 *Input*: hyper parameter $\kappa > 0$, the budget B; $L \in [K]$;
2 **for** $t \to 1 : \lceil K/L \rceil$ **do**
3 Pull arms $\{([(t-1)L + j - 1] \bmod K) + 1 | j \in [L]\}$;
4 **for** $t \to \lceil K/L \rceil + 1 : \infty$ **do**
5 Update the $T_i(t)$, $\hat{\mu}_i^r(t)$, $\hat{\mu}_i^c(t)$, $\tilde{\mu}_i^r(t)$, $\tilde{\mu}_i^c(t)$ for any i;
6 Pull the arms output by $\mathcal{O}(\tilde{\mu}^r(t), \tilde{\mu}^c(t), L)$; update B_t; *if* $B_t \geq 0$, obtain the reward; *else*, return;

4 Theoretical Analysis

In this section we theoretically analyze and upper bound the regret of the MRCB policy.

We first define some notations. (1) Let $\mathcal{C}_s$ denote $\mathcal{C}_L^K \backslash \{I_*\}$. (2) For any $i \in [K]$, let $\mathcal{S}_i$ denote $\{I | I \in \mathcal{C}_s, i \in I\}$. (3) For any $i \in [K]$, define

$$\begin{aligned} \Delta_{\min}^i &= \min_{I\in\mathcal{S}_i}(\varrho_L^* \textstyle\sum_{k\in I} \mu_k^c - \sum_{k\in I} \mu_k^r); \\ \Delta_{\max}^i &= \max_{I\in\mathcal{S}_i}(\varrho_L^* \textstyle\sum_{k\in I} \mu_k^c - \sum_{k\in I} \mu_k^r). \end{aligned} \tag{7}$$

Define $\mathcal{B} = \{i | i \in [K], \Delta_{\min}^i > 0\}$.
(4) $\mathcal{T}_L(B) = \lfloor 2B/(L\mu_{\min}^c) \rfloor$, where $\mu_{\min}^c = \min_{i\in[K]} \mu_i^c$.
(5) $\mathcal{X}_L(B) = O\big([B/(L\mu_{\min}^c)] \exp\{-(B\mu_{\min}^c)/2\}\big)$.

[2]We follow the common practice in combinatorial optimization literature that the "polynomial time" means "polynomial time in the number of bits of precision in which the inputs are specified".

[3]This trick has also been used in [Li *et al.*, 2010].

[4]Keep in mind that both $\tilde{\mu}^r(t)$ and $\tilde{\mu}^c(t)$ depend on the κ.

The above notations can be interpreted as follows. (1) $\mathcal{C}_s$ can be regarded as the set of all suboptimal super arms, since it is very likely that these arms are not as good as the near-optimal arm I_* in terms of the ratio of expected rewards to expected costs. (2) $\mathcal{S}_i$ is the collection of suboptimal super arms containing arm i. (3) $\Delta^i_{\min}$ and $\Delta^i_{\max}$ are two gaps measuring the suboptimality of the super arms in $\mathcal{S}_i$. $\mathcal{B}$ is a collection of "bad" arms, which can lead to regret after pulling. (4) $\mathcal{T}_L(B)$ can be seen as the *pseudo stopping time* of the bandit, since when B is large, the probability that the pulling rounds of an MP-BMAB can exceed $\mathcal{T}_L(B)$, bounded by $\mathcal{X}_L(B)$, is very small. Mathematically,

$$\textstyle\sum_{t=\mathcal{T}_L(B)+1}^{\infty}\mathbb{P}\{B_t \geq 0\} \leq \mathcal{X}_L(B). \tag{8}$$

Note $\mathcal{X}_L(B)$ decreases exponentially w.r.t. B. The proof of the above inequality is left in Appendix C. In our MP-BMAB problem, the stopping time is not given in advance like those in [Auer *et al.*, 2002; Badanidiyuru *et al.*, 2013]; instead, the stopping time is controlled by the budget B. To leverage the proof techniques from conventional bandits, we introduce the pseudo stopping time $\mathcal{T}_L(B)$. We will see how to use it later.

Define $\zeta_\kappa(\mathcal{T}_L(B)) = \sum_{t=1}^{\mathcal{T}_L(B)}(\log_2(t)+1)t^{-\kappa}$.

One can verify that when $\kappa > 1$, $\zeta_\kappa(\mathcal{T}_L(B))$ can be bounded by a term depending on κ only; when $\kappa = 1$, $\zeta_\kappa(\mathcal{T}_L(B))$ is of order $O(\ln^2(B))$; when $\kappa < 1$, $\zeta_\kappa(\mathcal{T}_L(B))$ is of order $O(B^{1-\kappa}\ln(B)/(1-\kappa))$. (See Appendix D for details.)

We can upper bound the regret of our policy as follows.

Theorem 3 *The regret of MRCB is upper bounded by*

$$\varphi_\iota \ln \mathcal{T}_L(B) + \varphi_s \zeta_\kappa(\mathcal{T}_L(B)) + \varphi_0, \tag{9}$$

where $\varphi_\iota = (\varrho_L^*+1)^2L^2(\sqrt{\kappa}+1)^2\sum_{i\in\mathcal{B}}(2/\Delta^i_{\min} - 1/\Delta^i_{\max})$, $\varphi_s = 2L\sum_{i\in\mathcal{B}}\Delta^i_{\max}$, *and* $\varphi_0 = 0.5(L-1)\varphi_\iota \ln K + \varphi_s + L\varrho_L^*\mathcal{X}_L(B) + 2L\varrho_L^* + 2\sum_{i\in\mathcal{B}}\Delta^i_{\max}$.

When $\kappa \in (0,1)$, the regret shown in (9) can be written as $\varphi_s\mathcal{T}_L^{1-\kappa}(B)\ln(\mathcal{T}_L(B))/(1-\kappa) + o(\mathcal{T}_L(B))$, which is sublinear in terms of $\mathcal{T}_L(B)$, and thus B. When $\kappa > 1$, the regret improves to $\varphi_\iota \ln \mathcal{T}_L(B) + O(1)$, which is of order $O(\kappa \ln B)$.

Proof outline: The proof of Theorem 3 is quite technical. Here we only give a proof sketch. The omitted derivation details are left in Appendix E.

∘ *Step 1: Bridge the regret and the expected pulling number of each suboptimal super arm.* With some derivations, we can get that the regret can be bounded as

$$\text{Regret} \leq \textstyle\sum_{I\in\mathcal{C}_s}\Delta^I\mathbb{E}\{\mathcal{N}_I\} + L\varrho_L^*\mathcal{X}_L(B) + 2L\varrho_L^*, \tag{10}$$

where for any $I \in \mathcal{C}_s$, Δ^I is defined as $(\sum_{k\in I}\mu_k^c)[\varrho_L^* - (\sum_{k\in I}\mu_k^r)/(\sum_{k\in I}\mu_k^c)]$, $\mathcal{N}_I$ is the pulling number of super arm I from round 1 to round $\mathcal{T}_L(B)$. The insight behind (10) is very intuitive: if the player pulls a suboptimal super arm I once, the expected cost is $\sum_{k\in I}\mu_k^c$; if she spends such cost on the near optimal super arm, she can gain Δ^I more reward. (10) frees us from the randomness of the stopping time, and allows us to only consider the expected pulling number of suboptimal arms before round $\mathcal{T}_L(B)$, which is deterministic (even though the budget might run out before it).

∘ *Step 2: Bridge the regret and each arm.* It is not convenient to work on the super arms directly. Therefore, we need to further decompose (10).

Let K_i denote that number of super arms in $\mathcal{S}_i$ for any $i \in [K]$, and $S(i,j)$ denote one super arm in $\mathcal{S}_i$ indexed by $j \in [K_i]$. Assume the super arms in $\mathcal{S}_i$ are sorted by the order $\Delta^{S(i,1)} \geq \Delta^{S(i,2)} \geq \cdots \geq \Delta^{S(i,K_i)}$. For simplicity of use, denote $\Delta^{S(i,j)}$ as $\Delta^{i,j}$.

For any suboptimal super arm $S(i,j)$, define the δ-gap $\delta^{i,j}(\gamma)$ in Eqn.(11-a), which can be seen as a weighted version of $\Delta^{i,j}$. We can verify that the gap satisfies Eqn.(11-b).

$$\text{(a) } \delta^{i,j}(\gamma) = \frac{\Delta^{i,j}}{\gamma\varrho_L^*+1}; \text{(b) } \varrho_L^* = \frac{(\sum_{k\in S(i,j)}\mu_k^r) + \delta^{i,j}(\gamma)}{(\sum_{k\in S(i,j)}\mu_k^c) - \gamma\delta^{i,j}(\gamma)}. \tag{11}$$

In the analysis of the upper bound of the regret, we only need to consider the case of [5] $\gamma = 1$. For ease of reference, let $\delta^{i,j}$ denote $\delta^{i,j}(1)$.

Define $f_{i,j} = L^2(\sqrt{\kappa}+1)^2\ln[\sqrt{K^{L-1}}\mathcal{T}_L(B)]/(\delta^{i,j})^2$. According to [Chen *et al.*, 2013], the $\sum_{I\in\mathcal{C}_s}\Delta^I\mathbb{E}\{\mathcal{N}_I\}$ of (10) can be bounded by $\sum_{i\in\mathcal{B}}\mathcal{R}_i$, in which $\mathcal{R}_i$ is

$$\begin{aligned}\mathcal{R}_i \leq & 2\Delta^i_{\max} + L^2(1+\sqrt{\kappa})^2(\varrho_L^*+1)^2(2/\Delta^i_{\min} - 1/\Delta^i_{\max})\\ & \ln[\sqrt{K^{L-1}}\mathcal{T}_L(B)] + \mathbb{E}\sum_{t=t_0}^{\mathcal{T}_L(B)}\sum_{j=1}^{K_i}\Delta^{i,j}\mathbb{I}\{I_t = S(i,j),\\ & \forall k \in I_t\ T_k(t-1) > \lfloor f_{i,j}\rfloor\},\end{aligned} \tag{12}$$

where $t_0 = \lceil K/L\rceil + 1$.

∘ *Step 3: Bound the* $\mathbb{E}\{\cdot\}$ *in* (12). For ease of reference, let $U_{i,j}(t)$ denote the event $\{I_t = S(i,j), \forall k \in I_t\ T_k(t-1) > \lfloor f_{i,j}\rfloor\}$ in (12). Define the event $\mathcal{Q}_o(t)$ as:

$$\mathcal{Q}_o(t) = \textstyle\bigcup_{k\in I_*}\{\tilde{\mu}_k^r(t) \leq \mu_k^r\} \cup \{\tilde{\mu}_k^c(t) \geq \mu_k^c\}. \tag{13}$$

Accordingly, the $\mathbb{E}\{\cdot\}$ in (12) can be decomposed as:

$$\mathbb{E}\textstyle\sum_{t=t_0}^{\mathcal{T}_L(B)}\sum_{j=1}^{K_i}\Delta^{i,j}\mathbb{I}\{U_{i,j}(t), \mathcal{Q}_o(t)\} \tag{14}$$

$$+\mathbb{E}\textstyle\sum_{t=t_0}^{\mathcal{T}_L(B)}\sum_{j=1}^{K_i}\Delta^{i,j}\mathbb{I}\{U_{i,j}(t), \overline{\mathcal{Q}_o(t)}\}, \tag{15}$$

where $\overline{\mathcal{Q}_o(t)}$ means that the event $\mathcal{Q}_o(t)$ does not hold.

Step 3-1: Bound (14). Since $U_{i,j}(t)$ are disjoint for different $j \in [K_i]$, we have that $\sum_{j=1}^{K_i}\mathbb{I}\{U_{i,j}(t), \mathcal{Q}_o(t)\} \leq \mathbb{I}\{\mathcal{Q}_o(t)\}$. Since we do not need to consider the randomness of the stopping time, we can apply Hoeffding's maximal inequality and union bound, and obtain that

$$\mathbb{P}\{\mathcal{Q}_o(t)\} \leq 2L\{\log_2(t-1)+1\}(t-1)^{-\kappa}. \tag{16}$$

Thus, (14) is bounded by $2L\Delta^i_{\max}\zeta_\kappa(\mathcal{T}_L(B))$.

Step 3-2: Bound (15). If super arm $S(i,j)$ is pulled where $i \in \mathcal{B}$ and $j \in [K_i]$, conditioned on $\overline{\mathcal{Q}_o(t)}$, we know that $\mathbb{P}\{U_{i,j}(t), \overline{\mathcal{Q}_o(t)}\}$ is upper bounded by

$$\begin{aligned}&\mathbb{P}\Big\{\textstyle\bigcup_{k\in S(i,j)}\{\tilde{\mu}_k^r(t) \geq \mu_k^r + \frac{\delta^{i,j}}{L}, T_k(t-1) > \lfloor f_{i,j}\rfloor\} \cup\\ &\textstyle\bigcup_{k\in S(i,j)}\{\tilde{\mu}_k^c(t) \leq \mu_k^c - \frac{\delta^{i,j}}{L}, T_k(t-1) > \lfloor f_{i,j}\rfloor\}\Big\}.\end{aligned} \tag{17}$$

[5] The case of $\gamma \neq 1$ will be considered when analyzing the lower bound of MP-BMAB in the next section.

With some derivations, for any $k \in S(i,j)$, we have

$$\mathbb{P}\{\tilde{\mu}_k^r(t) \geq \mu_k^r + \frac{\delta^{i,j}}{L}, T_k(t-1) > \lfloor f_{i,j} \rfloor\} \leq 1/[K^{L-1}\mathcal{T}_L(B)];$$
$$\mathbb{P}\{\tilde{\mu}_k^c(t) \leq \mu_k^c - \frac{\delta^{i,j}}{L}, T_k(t-1) > \lfloor f_{i,j} \rfloor\} \leq 1/[K^{L-1}\mathcal{T}_L(B)].$$

Therefore, $\mathbb{P}\{U_{i,j}(t), \overline{\mathcal{Q}_o(t)}\} \leq (2L)/[K^{L-1}\mathcal{T}_L(B)]$. Accordingly, (15) can be bounded by $2L\Delta^i_{\max}$.

According to the above three steps, by combining (10), (12), the bound of (14) in Step 3-1, and the result of (15) in Step 3-2, we can eventually get Theorem 3. $\square$

5 Special Cases

In this section, we consider two special cases of MP-BMABs: the Bernoulli MP-BMABs, in which the reward and cost distributions of all the arms are Bernoulli, and the SP-BMABs, in which the player can only pull $L = 1$ arm at each round.

5.1 Bernoulli MP-BMABs

In this subsection, we present a lower bound for the regret of any consistent policy (defined later) for Bernoulli MP-BMABs and compare it with the regret of MRCB.

For any policy w, let $I_k^w(T)$ denote the pulling number of arm $k \in [K]$ in the first T rounds, and $I_I^w(T)$ for super arm $I \in \mathcal{C}_L^K$, where $T \in \mathbb{Z}_+$. If $\sum_{I \in \mathcal{C}_s} \mathbb{E}\{I_I^w(T)\} = o(T^a)$ holds for any $a \in (0,1)$ and any bandit, we say policy w is *consistent*. According to the analysis in Section 4, we can get that the regret of any consistent policy is sublinear to the pseudo stopping time $\mathcal{T}_L(B)$, and so to the budget B.

Since the costs are no larger than 1, the stopping time of a policy is at least B/L (assume B/L is an integer for simplicity). The regret of the first B/L rounds is certainly a lower bound of the total regret, thus we will only consider the regret in these rounds. Let $kl(x,y)$ denote the KL divergence of two Bernoulli distributions with parameters x and y:

$$kl(x,y) = x\ln\frac{x}{y} + (1-x)\ln\frac{1-x}{1-y} \quad \forall x, y \in (0,1). \tag{18}$$

For ease of reference, define $\delta^i_{\min}(\gamma) = \min_{j \in [K_i]} \delta^{i,j}(\gamma)$ for any $i \notin I_*$. Define the following optimization problem:

$$\begin{aligned} &\min_\gamma \ \ kl(\mu_i^r, \mu_i^r + \delta^i_{\min}(\gamma)) + kl(\mu_i^c, \mu_i^c - \gamma\delta^i_{\min}(\gamma)) \\ &\text{s.t.} \quad \mu_i^r + \delta^i_{\min}(\gamma) < 1, \ \mu_i^c - \gamma\delta^i_{\min}(\gamma) > 0, \ \gamma \geq 0. \end{aligned} \tag{19}$$

As shown in Appendix F.1 and F.2, we can prove that: (1) the feasible set of (19) is non-empty; (2) the optimal solution of (19) is an interior point of its constraint set. Thus, the optimal value exists and is strictly positive. Denote the optimal value of (19) as $\mathcal{L}_i^*$.

Theorem 4 *For Bernoulli MP-BMABs, if the rewards are independent to the costs for each arm, for any consistent policy w (i.e., $\sum_{I \in \mathcal{C}_s} \mathbb{E}\{I_I^w(T)\} = o(T^a)$ holds for any $a \in (0,1)$ and $T \in \mathbb{Z}_+$), we have that for any $i \notin I_*$ and $\epsilon > 0$,*

$$\lim_{B\to\infty} \mathbb{P}\Big\{ I_i^w(B/L) \geq \frac{(1-\epsilon)\ln(B/L)}{\mathcal{L}_i^*} \Big\} = 1,$$

and $\liminf_{B\to\infty} \mathbb{E}[I_i^w(B/L)]/[\ln(B/L)] \geq 1/\mathcal{L}_i^*$.

From Theorem 4, we can get that for Bernoulli MP-BMABs, the pulling time of arm $i \notin I_*$ for any consistent policy is at least $\Omega(\ln(B/L)/\mathcal{L}_i^*)$. After some derivations, we can get that the regret is $\Omega(\sum_{i \notin I_*}(\Delta^i_{\min}/\mathcal{L}_i^*)\ln(B/L))$. The theorem can be proved by using the change-of-measure techniques and large number laws, as shown in Appendix F.3.

First, we can see that, for Bernoulli bandits, the upper bound of the regret of MRCB is $O(\kappa \ln B)$ when $\kappa > 1$, which matches the lower bound in terms of the order of B.

Second, we make some discussion about the coefficients of $\ln B$. We specify MRCB by setting $\kappa = 2$. For ease of reference, denote the upper bound and lower bound as $O(o\ln B)$ and $\Omega(\omega \ln B)$ respectively. Similar to the UCB-based policies for conventional MABs (without budget constraints), our MRCB cannot match the lower bound perfectly, i.e., $o > \omega$. The following example shows that o in the upper bound of MRCB and ω in the lower bound share similar trends.

Example 5 *We study the relationship between the regret and the ratio gap $\Delta^i_{\min}$ $\forall i \in \mathcal{B}$ for an MP-BMAB. Suppose $p \in (0, 0.5)$. Consider a Bernoulli bandit with $\mu_j^r, \mu_j^c \in [p, 1-p]$ $\forall j \in [K]$ and $\Delta^i_{\min} < p/2$ $\forall i \in \mathcal{B}$. In this case, we have (a) $o = \sum_{i\in\mathcal{B}} L^2/(p^2\Delta^i_{\min})$; (b) $\omega = \sum_{i \notin I_*} p^2/\Delta^i_{\min}$. That is, the coefficients of $\ln B$ in both the upper and lower bounds of the regret are linear to $\sum_{i\notin I_*} 1/\Delta^i_{\min}$.*

5.2 Single-Play Budgeted MAB

Since SP-BMABs are a special case of MP-BMABs with $L = 1$, our MRCB policy (Algorithm 3) can be directly applied.

While applying Algorithm 3 to the SP-BMAB problem, I_* degenerates to the arm with the maximum ratio of the expected reward to the expected cost, i.e., $i_* = \arg\max_{i\in[K]} \mu_i^r/\mu_i^c$ and $\varrho_1^* = \mu_{i_*}^r/\mu_{i_*}^c$. For any $i \neq i_*$, (a) $\Delta^i_{\max}$ equals $\Delta^i_{\min}$ and we denote them as $\Delta^i = \mu_i^c\varrho_1^* - \mu_i^r$; (b) $\delta^i_{\min}(\gamma)$ degenerates to $\delta^i(\gamma) = \Delta^i/(\gamma\varrho_1^* + 1)$; (c) the optimization problem of (19) degenerates as follows:

$$\begin{aligned} &\min_\gamma \ \ kl(\mu_i^r, \mu_i^r + \delta^i(\gamma)) + kl(\mu_i^c, \mu_i^c - \gamma\delta^i(\gamma)) \\ &\text{s.t.} \quad \mu_i^r + \delta^i(\gamma) < 1, \ \gamma \geq 0. \end{aligned} \tag{20}$$

One can verify the existence of the optimal solution and optimal value of (20). Denote the optimal value as $\mathcal{L}_i^{**}$. Theorem 3 and 4 degenerate to the following two corollaries:

Corollary 6 *For SP-BMABs, the regret of MRCB is upper bounded by*

$$\sum_{i\neq i_*} \frac{[(\sqrt{\kappa}+1)(\varrho_1^*+1)]^2}{\Delta^i}\ln\mathcal{T}_1(B) + 2\zeta_\kappa(\mathcal{T}_1(B))\sum_{i\neq i_*}\Delta^i$$

$+4\sum_{i\neq i_*}\Delta^i + [2 + \mathcal{X}_1(B)]\varrho_1^*$, *where the $\mathcal{T}_1(B)$ and $\mathcal{X}_1(B)$ are obtained by setting the L's in $\mathcal{T}_L(B)$ and $\mathcal{X}_L(B)$ (defined at the beginning of Section 4) as* 1.

Corollary 7 *For Bernoulli SP-BMABs, if the rewards are independent to the costs for each arm, for any consistent policy w (i.e., $\sum_{i\neq i_*} \mathbb{E}\{I_i^w(T)\} = o(T^a)$ holds for any $a \in (0,1)$ and $T \in \mathbb{Z}_+$), we have that for any $i \neq i_*$ and $\epsilon > 0$,*

$$\lim_{B\to\infty} \mathbb{P}\{I_i^w(B) \geq [(1-\epsilon)/\mathcal{L}_i^{**}]\ln B\} = 1;$$

consequently, $\liminf_{B\to\infty} \mathbb{E}[I_i^w(B)]/[\ln B] \geq 1/\mathcal{L}_i^{**}$.

Corollary 7 tells that the regret for Bernoulli SP-BMAB is at least $\Omega(\sum_{i\neq i_*}(\Delta^i/\mathcal{L}_i^{**})\ln B)$. So far as we know, it is the first non-trivial lower bound for SP-BMABs. Similar to the Example 5, for SP-BMABs, the coefficients of $\ln B$ in both the upper bound of MRCB's regret and the lower bound of the regret of any consistent policy are linear to $\sum_{i\neq i_*} 1/\Delta^i$.

SP-BMABs with random costs have been studied in [Ding *et al.*, 2013]. Compared with the above literature, MRCB has two advantages: (1) Since there is a hyper parameter of our policy, by carefully setting the parameter, the empirical performance of our policy can outperform previous algorithms (see Section 6). (2) The theoretical guarantees of our policy are better than previous UCB-based policies. For example, Corollary 6 outperforms the regret bound in [Ding *et al.*, 2013]. (See Appendix G for the details.)

6 Empirical Evaluations

We conducted a set of numerical simulations to test the empirical performance of our policy. We compared with the following baselines. (1) The ε-first policy first pulls the arms one by one when the spent budget is less than εB; after that we have two schemes to recommend L arms: *Scheme T* always pulling the top L arms with the largest average reward to average cost ratio, and *Scheme R* always pulling the L arms with the maximum ratio of the sum of average rewards to the sum of average costs. We followed the practice in [Tran-Thanh *et al.*, 2010; Xia *et al.*, 2015b] to set $\varepsilon = 0.1$. (2) Fractional KUBE [Tran-Thanh *et al.*, 2012] with both schemes T and R. (3) BTS policy [Xia *et al.*, 2015b] with schemes T and R. (4) the UCB-BV1 [Ding *et al.*, 2013] with scheme T only, since the confidence term is added to the ratio of the average rewards to average costs, which makes it hard to be associated with scheme R. For ε-first, we set the budget as $\{5K, 10K, 15K, \cdots, 50K\}$; for the other policies, we set the budget as $50K$ and record the regret at each budget.

We simulated the bandit with two distributions: one with multinomial distribution, and the other with beta distribution. For each distribution, we simulated a 10-armed bandit and a 50-armed bandit. Detailed parameters of the distributions are left in Appendix H.1 due to limited space. We individually run each policy under each setting for 100 times and report the average regret and standard derivation over the 100 runs.

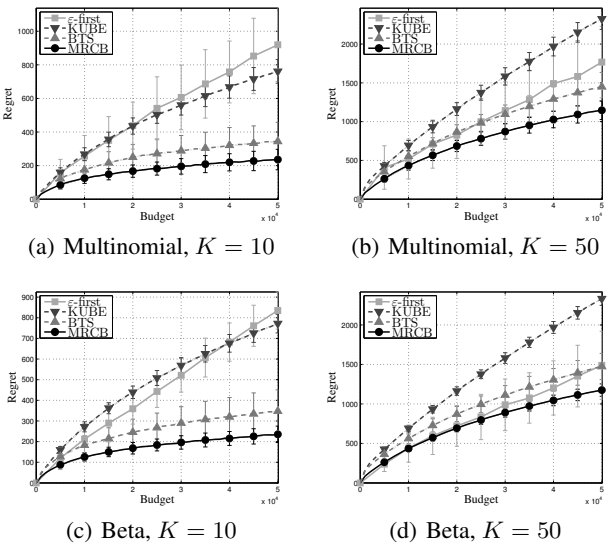
MRCB has a hyper parameter κ. We searched the κ in the set $\{2^{-10}, 2^{-7}, 2^{-4}, 2^{1}\}$ and found that $\kappa = 2^{-4}$ worked well for most cases. Therefore, we fix 2^{-4} in the following experiments. Though asymptotically MRCB enjoys log-linear regret when $\kappa > 1$, it is not good to set large values for κ since B is limited in our experiments.

The results of the first three baselines with different schemes are shown in Table 1. It is obvious that Scheme R is better than Scheme T. Thus, in the following experiments, we will only show the results for Scheme R. We will not show the regrets for UCB-BV1 neither since they are too large.

The average regret and the standard deviation of each policy w.r.t different K and different reward/cost distributions are shown in Figure 1. We can see that our MRCB has clear advantages over the 3 baselines: It achieves smaller regrets and lower standard derivations. When the number of arms increases from 10 to 50, the regrets of all the policies increase. This is in accord with our intuition, since more candidate arms can make the nearly-optimal super arm harder to be found.

Table. 1 Comparison of Baselines

	ε-first	KUBE	BTS	UCB-BV1
Scheme T	1159.0	831.3	568.5	2273.8
Scheme R	919.4	760.8	344.3	- - -



(a) Multinomial, $K = 10$ (b) Multinomial, $K = 50$

(c) Beta, $K = 10$ (d) Beta, $K = 50$

Figure 1: The Regrets

We also tested the performance of MRCB under the SP-BMAB setting (i.e., $L = 1$). The results are in Table 2, which are carried out on the bandits with multinomial reward/cost distributions and $B = 50K$. The average regrets and the standard derivations are reported. Again, MRCB performs the best, which shows the MRCB can handle the SP-BMAB. Additional experiments can be found at Appendix H.2.

Table 2. Regrets for SP-BMAB

	10-armed bandit	50-armed bandit
ε-first	2183.5 ± 51.9	2403.8 ± 54.9
KUBE	552.9 ± 34.4	2722.3 ± 66.9
BTS	226.9 ± 38.3	1182.0 ± 93.6
MRCB	$\mathbf{103.3 \pm 13.5}$	$\mathbf{521.9 \pm 31.1}$

7 Conclusion and Future Work

In this work, we studied the MP-BMAB problem and proposed a policy for it. The policy theoretically enjoys a sublinear regret (log-linear under some conditions) and empirically outperforms several baselines in different settings.

There are several aspects to study in the future for MP-BMABs. (1) multi-play budgeted linear/contextual bandit, in which each arm is associated with a multi-dimensional feature vector, is an attractive topic; (2) the distribution-free upper/lower bound of MP-BMABs is still unknown and remains to be explored.

Acknowledgments

This work is partially supported by National Natural Science Foundation of China (NSFC, NO.61371192).

References

[Agrawal *et al.*, 2010] R Agrawal, M Hegde, and D Teneketzis. Multi-armed bandit problems with multiple plays and switching cost. 2010.

[Anantharam *et al.*, 1987] Venkatachalam Anantharam, Pravin Varaiya, and Jean Walrand. Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-part i: Iid rewards. *Automatic Control, IEEE Transactions on*, 32(11):968–976, 1987.

[Audibert *et al.*, 2009] Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. Exploration-exploitation tradeoff using variance estimates in multi-armed bandits. *Theor. Comput. Sci.*, 410(19):1876–1902, 2009.

[Auer *et al.*, 2002] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.

[Auer, 2003] Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *The Journal of Machine Learning Research*, 3:397–422, 2003.

[Badanidiyuru *et al.*, 2013] Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. Bandits with knapsacks. In *Foundations of Computer Science (FOCS), 2013 IEEE 54th Annual Symposium on*, pages 207–216. IEEE, 2013.

[Boyd and Vandenberghe, 2004] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[Bubeck and Cesa-Bianchi, 2012] Sébastien Bubeck and Nicolo Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *arXiv preprint arXiv:1204.5721*, 2012.

[Bubeck, 2010] Sébastien Bubeck. *Bandits games and clustering foundations*. PhD thesis, Université des Sciences et Technologie de Lille-Lille I, 2010.

[Chen *et al.*, 2013] Wei Chen, Yajun Wang, and Yang Yuan. Combinatorial multi-armed bandit: General framework and applications. In *Proceedings of the 30th International Conference on Machine Learning*, pages 151–159, 2013.

[Ding *et al.*, 2013] Wenkui Ding, Tao Qin, Xu-Dong Zhang, and Tie-Yan Liu. Multi-armed bandit with budget constraint and variable costs. In *Twenty-Seventh AAAI Conference on Artificial Intelligence*, 2013.

[Flajolet and Jaillet, 2015] Arthur Flajolet and Patrick Jaillet. Low regret bounds for bandits with knapsacks. *arXiv preprint arXiv:1510.01800*, 2015.

[Garivier and Cappé, 2011] Aurélien Garivier and Olivier Cappé. The kl-ucb algorithm for bounded stochastic bandits and beyond. *arXiv preprint arXiv:1102.2490*, 2011.

[Ghosh and Hummel, 2013] Arpita Ghosh and Patrick Hummel. Learning and incentives in user-generated content: Multi-armed bandits with endogenous arms. In *Proceedings of the 4th conference on Innovations in Theoretical Computer Science*, pages 233–246. ACM, 2013.

[Honda and Takemura, 2010] Junya Honda and Akimichi Takemura. An asymptotically optimal bandit algorithm for bounded support models. In *COLT*, pages 67–79. Citeseer, 2010.

[Komiyama *et al.*, 2015] Junpei Komiyama, Junya Honda, and Hiroshi Nakagawa. Optimal regret analysis of thompson sampling in stochastic multi-armed bandit problem with multiple plays. In *ICML*, 2015.

[Kveton *et al.*, 2015] Branislav Kveton, Zheng Wen, Azin Ashkan, and Csaba Szepesvári. Tight regret bounds for stochastic combinatorial semi-bandits. *AISTATS*, 2015.

[Lai and Robbins, 1985] Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

[Li *et al.*, 2010] Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *WWW*, pages 661–670. ACM, 2010.

[Liu and Zhao, 2010] Keqin Liu and Qing Zhao. Decentralized multi-armed bandit with multiple distributed players. In *Information Theory and Applications Workshop (ITA), 2010*, pages 1–10. IEEE, 2010.

[Lueker, 1975] George S Lueker. *Two NP-complete problems in nonnegative integer programming*. Princeton University. Department of Electrical Engineering, 1975.

[Mohri and Munoz, 2014] Mehryar Mohri and Andres Munoz. Optimal regret minimization in posted-price auctions with strategic buyers. In *Advances in Neural Information Processing Systems*, pages 1871–1879, 2014.

[Seerengasamy and Jeyaraman, 2013] V Seerengasamy and K Jeyaraman. An alternative method to find the solution of zero one integer linear fractional programming problem with the help of θ-matrix. *International Journal of Scientific and Research Publications*, 2013.

[Tran-Thanh *et al.*, 2010] Long Tran-Thanh, Archie Chapman, Enrique Munoz de Cote, Alex Rogers, and Nicholas R. Jennings. Epsilon-first policies for budget limited multi-armed bandits. In *AAAI*, 2010.

[Tran-Thanh *et al.*, 2012] Long Tran-Thanh, Archie C Chapman, Alex Rogers, and Nicholas R Jennings. Knapsack based optimal policies for budget-limited multi-armed bandits. In *AAAI*, 2012.

[Tran-Thanh *et al.*, 2014] Long Tran-Thanh, Lampros C Stavrogiannis, Victor Naroditskiy, Valentin Robu, Nicholas R Jennings, and Peter Key. Efficient regret bounds for online bid optimisation in budget-limited sponsored search auctions. pages 809–818, 2014.

[Vanchinathan *et al.*, 2015] Hastagiri P Vanchinathan, Andreas Marfurt, Charles-Antoine Robelin, Donald Kossmann, and Andreas Krause. Discovering valuable items from massive data. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1195–1204. ACM, 2015.

[Xia *et al.*, 2015a] Yingce Xia, Wenkui Ding, Xu-Dong Zhang, Nenghai Yu, and Tao Qin. Budgeted bandit problems with continuous random costs. In *The 7th Asian Conference on Machine Learning*, 2015.

[Xia *et al.*, 2015b] Yingce Xia, Haifang Li, Tao Qin, Nenghai Yu, and Tie-Yan Liu. Thompson sampling for budgeted multi-armed bandits. In *24th International Joint Conference on Artificial Intelligence*, pages 3960–3966, 2015.

[Xia *et al.*, 2016] Yingce Xia, Tao Qin, Nenghai Yu, and Tie-Yan Liu. Best action selection in a stochastic environment. In *15th International Conference on Autonomous Agents and Multiagent Systems*, 2016.

Appendix

In the proofs, we will often come across the sum of the specific elements in a vector indexed by a set. Assume x is a K-dimensional vector, with the i-th element x_i ($i \in [K]$), and I is a set s.t. $I \subset [K]$. Define $\Sigma(x, I) = \sum_{i \in I} x_i$. Let μ^r denote the vector $(\mu_1^r, \mu_2^r, \cdots \mu_K^r)$. Let μ^c denote the vector $(\mu_1^c, \mu_2^c, \cdots \mu_K^c)$. For any $t \in \mathbb{Z}_+$, let $r(t)$ denote the vector $(r_1(t), r_2(t), \cdots, r_K(t))$ and let $c(t)$ denote the vector $(c_1(t), c_2(t), \cdots, c_K(t))$. Consequently, for any $I \subset [K]$, $\Sigma(\mu^r, I) = \sum_{i \in I} \mu_i^r$ and $\Sigma(\mu^c, I) = \sum_{i \in I} \mu_i^c$. $\Sigma(r(t), I) = \sum_{i \in I} r_i(t)$, $\Sigma(c(t), I) = \sum_{i \in I} c_i(t)$ for any $t \geq 1$ and $I \in \mathcal{C}_L^K$. The super arm pulled at round t is denoted as I_t. Denote the history before round t as $\mathcal{H}_{t-1}$, which is defined as follows:

$$\mathcal{H}_{t-1} = \{I_\tau, r_k(\tau) \text{ for each } k \in I_\tau \ , c_k(\tau) \text{ for each } k \in I_\tau, \ \tau = 1, 2, \cdots t-1\}. \tag{21}$$

A Bound Expected Rewards of the Optimal Policy and $\mathcal{M}_g$

We give the proof of Lemma 1 in this section.
(1) *Bound* R^*: Define $B_0 = B$. Denote the pulling sequences generated by an algorithm w as $\mathcal{I}_w = \{I_t\}_{t=1}^\infty$. When the reward and cost distributions are all known, the decision of I_t can depend on (1) $\mathcal{H}_{t-1}$, (2) the reward and cost distributions, but cannot depend on the reward and cost at and after round t. Denote the expected reward of algorithm w before the budget runs out as $\tilde{R}_w$. $\tilde{R}_w$ can be bounded as follows:

$$\begin{aligned}
\tilde{R}_w &= \mathbb{E}\sum_{t=1}^{\infty}\sum_{I\in\mathcal{C}_L^K} \Sigma(r(t),I)\mathbb{I}\{I_t = I, B_t \geq 0\} \leq \sum_{t=1}^{\infty}\sum_{I\in\mathcal{C}_L^K} \mathbb{E}\big[\Sigma(r(t),I)\mathbb{I}\{I_t = I, B_{t-1} \geq 0\}\big] \quad \triangleright B_t \geq 0 \text{ implies that } B_{t-1} \geq 0. \\
&= \sum_{t=1}^{\infty}\sum_{I\in\mathcal{C}_L^K} \mathbb{E}\big[\mathbb{E}[\Sigma(r(t),I)\mathbb{I}\{I_t = I, B_{t-1} \geq 0\}|\mathcal{H}_{t-1}]\big] \quad \triangleright \text{ The first } \mathbb{E} \text{ is taken w.r.t } \mathcal{H}_{t-1}. \\
&= \sum_{t=1}^{\infty}\sum_{I\in\mathcal{C}_L^K} \mathbb{E}\big[\mathbb{E}\{\Sigma(r(t),I)|\mathcal{H}_{t-1}\}\mathbb{E}[\mathbb{I}\{I_t = I, B_{t-1} \geq 0\}|\mathcal{H}_{t-1}]\big] \quad \triangleright \mathbb{I}\{I_t = I, B_{t-1} \geq 0\} \text{ is determined by } \mathcal{H}_{t-1}. \\
&= \sum_{t=1}^{\infty}\sum_{I\in\mathcal{C}_L^K} \mathbb{E}\big[\Sigma(\mu^r,I)\mathbb{E}[\mathbb{I}\{I_t = I, B_{t-1} \geq 0\}|\mathcal{H}_{t-1}]\big] \quad \triangleright \text{ Reward of a fixed super arm } I \text{ at round } t \text{ are independent of } \mathcal{H}_{t-1}. \\
&= \sum_{t=1}^{\infty}\sum_{I\in\mathcal{C}_L^K} \frac{\Sigma(\mu^r,I)}{\Sigma(\mu^c,I)}\mathbb{E}\big[\Sigma(\mu^c,I)\mathbb{E}[\mathbb{I}\{I_t = I, B_{t-1} \geq 0\}|\mathcal{H}_{t-1}]\big] \\
&= \sum_{t=1}^{\infty}\sum_{I\in\mathcal{C}_L^K} \frac{\Sigma(\mu^r,I)}{\Sigma(\mu^c,I)}\mathbb{E}\big[\mathbb{E}\{\Sigma(c(t),I)|\mathcal{H}_{t-1}\}\mathbb{E}[\mathbb{I}\{I_t = I, B_{t-1} \geq 0\}|\mathcal{H}_{t-1}]\big] \quad \triangleright \text{ The first } \mathbb{E} \text{ is taken w.r.t } \mathcal{H}_{t-1}. \\
&= \sum_{t=1}^{\infty}\sum_{I\in\mathcal{C}_L^K} \frac{\Sigma(\mu^r,I)}{\Sigma(\mu^c,I)}\mathbb{E}[\Sigma(c(t),I)\mathbb{I}\{I_t = I, B_{t-1} \geq 0\}] \\
&\leq \sum_{t=1}^{\infty}\sum_{I\in\mathcal{C}_L^K} \varrho_L^*\mathbb{E}\{\Sigma(c(t),I)\mathbb{I}\{I_t = I, B_{t-1} \geq 0\}\} \quad \triangleright \text{ Recall that } \varrho_L^* = \frac{\Sigma(\mu^r,I_*)}{\Sigma(\mu^c,I_*)} \geq \frac{\Sigma(\mu^r,I)}{\Sigma(\mu^c,I)} \text{ for any } I \in \mathcal{C}_L^K. \\
&= \varrho_L^*\mathbb{E}\sum_{t=1}^{\infty}\sum_{I\in\mathcal{C}_L^K} \Sigma(c(t),I)\mathbb{I}\{I_t = I, B_{t-1} \geq 0\} = \varrho_L^*\mathbb{E}\sum_{t=1}^{\infty} \Sigma(c(t),I_t)\mathbb{I}\{B_{t-1} \geq 0\} \leq^{\triangle} (B+L)\varrho_L^*. \qquad \square
\end{aligned}$$

There are two things to be claimed in the above derivations:

1. We need to prove that

$$\mathbb{E}\sum_{t=1}^{\infty}\sum_{I\in\mathcal{C}_L^K} \Sigma(r(t),I)\mathbb{I}\{I_t = I, B_t \geq 0\} = \sum_{t=1}^{\infty}\sum_{I\in\mathcal{C}_L^K} \mathbb{E}\big[\Sigma(r(t),I)\mathbb{I}\{I_t = I, B_t \geq 0\}\big]; \tag{22}$$

$$\mathbb{E}\sum_{t=1}^{\infty}\sum_{I\in\mathcal{C}_L^K} \Sigma(c(t),I)\mathbb{I}\{I_t = I, B_t \geq 0\} = \sum_{t=1}^{\infty}\sum_{I\in\mathcal{C}_L^K} \mathbb{E}\big[\Sigma(c(t),I)\mathbb{I}\{I_t = I, B_t \geq 0\}\big]. \tag{23}$$

Also, the two B_t's in (22) or (23) need to be replaced with two B_{t-1}'s.

We need to use the following conclusion to prove the above two equations:

$\triangleright$ *If $X_0, X_1, \cdots$ is a sequence of random variables such that $\sum_{j=0}^{\infty}\mathbb{E}\{|X_j|\}$ converges, then the linearity of expectations holds: $\mathbb{E}[\sum_{j=0}^{\infty}X_j]=\sum_{j=0}^{\infty}\mathbb{E}[X_j]$.*

The above conclusion is from Page 23 (or, Exercise 2.29), Book *"Probability and Computing: Randomized Algorithms and Probabilistic Analysis", authored by Michael Mitzenmacher, Eli Upfal, Cambridge University Press, 2005, eBook at: https://books.google.co.jp/books?id=0bAYl6d7hvkC&lpg=PA23&pg=PA23#v=onepage&q&f=false*

Note that each term in the r.h.s of (22) is nonnegative. The r.h.s of (22) can be bounded as

$$\begin{aligned}&\sum_{t=1}^{\infty}\sum_{I\in\mathcal{C}_L^K}\mathbb{E}\big[\Sigma(r(t),I)\mathbb{I}\{I_t=I,B_t\geq 0\}\big]=\sum_{t=1}^{\infty}\mathbb{E}\big[\Sigma(r(t),I_t)\mathbb{I}\{B_t\geq 0\}\big]\leq L\sum_{t=1}^{\infty}\mathbb{E}\big[\mathbb{I}\{B_t\geq 0\}\big]\\ =&L\sum_{t=1}^{\mathcal{T}_L(B)}\mathbb{P}\{B_t\geq 0\}+L\sum_{t=\mathcal{T}_L(B)+1}^{\infty}\mathbb{P}\{B_t\geq 0\}\leq L(\mathcal{T}_L(B)+\tilde{\mathcal{X}}_L(B)),\end{aligned}\tag{24}$$

where the last inequality is obtained by Lemma 9, which is shown later. Both $\mathcal{T}_L(B)$ and $\tilde{\mathcal{X}}_L(B)$ are deterministic and finite numbers, which tells that $\sum_{t=1}^{\infty}\sum_{I\in\mathcal{C}_L^K}\mathbb{E}\big[|\Sigma(r(t),I)\mathbb{I}\{I_t=I,B_t\geq 0\}|\big]$ converges. Therefore, (22) holds. (23) can be similarly proved. If the B_t's in (22) or (23) are replaced with B_{t-1}'s, they can also be similarly proved. In the next context, we can safely exchange the position of $\mathbb{E}$ when meeting similar situations like those in (22) and (23).

2. Another thing is the inequality marked with $\triangle$ holds because: assume $B_{\tau-1}\geq 0$ but $B_\tau<0$. In this case, we have

$$\sum_{t=1}^{\infty}\Sigma(c(t),I_t)\mathbb{I}\{B_{t-1}\geq 0\}=\Big(\sum_{t=1}^{\tau-1}\Sigma(c(t),I_t)\Big)+\Sigma(c(\tau),I_\tau)\leq B-B_{\tau-1}+L\leq B+L.\tag{25}$$

Adding $\mathbb{E}$ to both sides of (25), we can get that the inequality marked with $\triangle$ holds.

(2) *Bound the expected reward of $\mathcal{M}_g$.* Denote the expected reward of always pulling I_* (i.e., $\mathcal{M}_g$) as R_g. We have

$$\begin{aligned}R_g&=\mathbb{E}\sum_{t=1}^{\infty}\Sigma(r(t),I_*)\mathbb{I}\{B_t\geq 0\}\geq\mathbb{E}\sum_{t=1}^{\infty}\Sigma(r(t),I_*)\mathbb{I}\{B_{t-1}\geq L\}\quad\triangleright\text{ Note } B_{t-1}\geq L \text{ implies that } B_t\geq 0.\\ &=\sum_{t=1}^{\infty}\mathbb{E}\{\Sigma(r(t),I_*)\}\mathbb{P}\{B_{t-1}\geq L\}\quad\triangleright\text{ The reward of super arm } I_* \text{ is independent to the history until round } t-1.\\ &=\sum_{t=1}^{\infty}\Sigma(\mu^r,I_*)\mathbb{P}\{B_{t-1}\geq L\}=\sum_{t=1}^{\infty}\frac{\Sigma(\mu^r,I_*)}{\Sigma(\mu^c,I_*)}\Sigma(\mu^c,I_*)\mathbb{P}\{B_{t-1}\geq L\}=\sum_{t=1}^{\infty}\varrho_L^*\mathbb{E}\{\Sigma(c(t),I_*)\}\mathbb{P}\{B_{t-1}\geq L\}\\ &=\sum_{t=1}^{\infty}\varrho_L^*\mathbb{E}\{\Sigma(c(t),I_*)\mathbb{I}\{B_{t-1}\geq L\}\}\geq(B-L)\varrho_L^*.\quad\triangleright\text{ Can be similarly obtained like (25).}\quad\square\end{aligned}$$

B Proof of Lemma 2

The original problem is (P1). We give a relaxed version in (P2). We only need to consider the case that $a_i>0, b_i\geq 0$ for any $i\in[K]$, and there are at most $L-1$ b_i's s.t. $b_i=0$.

$$\begin{aligned}(P1)\quad&\max\ \frac{\sum_{i=1}^K a_ix_i}{\sum_{i=1}^K b_ix_i}\\ \text{s.t.}\quad&\sum_{i=1}^K x_i=L;\\ &x_i\in\{0,1\}\quad\text{for any } i\in[K].\end{aligned}\qquad\qquad\begin{aligned}(P2)\quad&\max\ \frac{\sum_{i=1}^K a_ix_i}{\sum_{i=1}^K b_ix_i}\\ \text{s.t.}\quad&\sum_{i=1}^K x_i=L;\\ &0\leq x_i\leq 1\quad\text{for any } i\in[K].\end{aligned}\tag{26, 27}$$

We first need to prove the following lemma:

Lemma 8 *Given any optimal solution of (P2) in* (27)*, it can be converted to the optimal solution in* (26) *within K steps.*

Proof. Assume $x^{(0)}$ is one optimal solution of (P2). If each element in $x^{(0)}$ is either 0 or 1, then it is the optimal solution of (P1), and we can directly return $x^{(0)}$.

Otherwise, denote $\sum_{i=1}^K a_ix_i^{(0)}$ as A_1, and denote $\sum_{i=1}^K b_ix_i^{(0)}$ as A_2 (Note that $A_1,A_2>0$, since $a_i>0\ \forall i\in[K]$ and there are at most $L-1$ b_i's equaling zero. The number of non-zero x_i's is at least L.). If there exist some fractional elements in $x^{(0)}$, there are at least two fractional elements in $x^{(0)}$, since L is an integer. Denote two indexes of the fractional element as m and n, which means that $0<x_m^{(0)},x_n^{(0)}<1$. Denote ν as a real number s.t. $0<\nu\leq\min\{x_m^{(0)},1-x_n^{(0)}\}$. We have that

$$\frac{\sum_{i\neq m,n}a_ix_i^{(0)}+a_m(x_m^{(0)}-\nu)+a_n(x_n^{(0)}+\nu)}{\sum_{i\neq m,n}b_ix_i^{(0)}+b_m(x_m^{(0)}-\nu)+b_n(x_n^{(0)}+\nu)}\leq\frac{\sum_{i=1}^K a_ix_i^{(0)}}{\sum_{i=1}^K b_ix_i^{(0)}}.\tag{28}$$

Please note that the two denominators in (28) are both strictly larger than zero. Therefore, (28) implies that

$$A_2(a_m - a_n) \geq A_1(b_m - b_n). \tag{29}$$

On the other hand, we can obtain for any $0 < \nu \leq \min\{1 - x_m^{(0)}, x_n^{(0)}\}$.

$$\frac{\sum_{i \neq m,n} a_i x_i^{(0)} + a_m(x_m^{(0)} + \nu) + a_n(x_n^{(0)} - \nu)}{\sum_{i \neq m,n} b_i x_i^{(0)} + b_m(x_m^{(0)} + \nu) + b_n(x_n^{(0)} - \nu)} \leq \frac{\sum_{i=1}^K a_i x_i^{(0)}}{\sum_{i=1}^K b_i x_i^{(0)}},$$

which implies that

$$A_2(a_m - a_n) \leq A_1(b_m - b_n). \tag{30}$$

Combining (29) and (30), we have that

$$A_2(a_m - a_n) = A_1(b_m - b_n). \tag{31}$$

We propose two rounding methods to get $x^{(1)}$: (Recall we have assumed that $0 < x_m^{(0)}, x_n^{(0)} < 1$)

- (R1) Define $d_{m,n} = \min\{x_m^{(0)}, 1 - x_n^{(0)}\}$. $x_i^{(1)} = x_i^{(0)}$ for any $i \neq m, n$; $x_m^{(1)} = x_m^{(0)} - d_{m,n}$; $x_n^{(1)} = x_n^{(0)} + d_{m,n}$.
- (R2) Define $d'_{m,n} = \min\{1 - x_m^{(0)}, x_n^{(0)}\}$; $x_i^{(1)} = x_i^{(0)}$ for any $i \neq m, n$; $x_m^{(1)} = x_m^{(0)} + d'_{m,n}$; $x_n^{(1)} = x_n^{(0)} - d'_{m,n}$.

We will prove that the $x^{(1)}$ obtained by (R1) from $x^{(0)}$ is the optimal solution of (P2):

1. By Eqn. (31), we know that if $a_m = a_n$, we must have $b_m = b_n$; if $b_m = b_n$, we must have $a_m = a_n$. That is, if $a_m = a_n$ or $b_m = b_n$, $x^{(1)}$ is another optimal solution of (P2). The number of non-integer element in $x^{(1)}$ is strictly smaller than that of $x^{(0)}$.
2. If $a_m \neq a_n$ and $b_m \neq b_n$, we can conclude that

$$\frac{a_m - a_n}{b_m - b_n} = \frac{A_1}{A_2}. \tag{32}$$

Thus,

$$\frac{\sum_{i=1}^K a_i x_i^{(1)}}{\sum_{i=1}^K b_i x_i^{(1)}} = \frac{\sum_{i=1}^K a_i x_i^{(0)} - d_{m,n}(a_m - a_n)}{\sum_{i=1}^K b_i x_i^{(0)} - d_{m,n}(b_m - b_n)} = \frac{A_1 - (A_1/A_2)d_{m,n}(b_m - b_n)}{A_2 - d_{m,n}(b_m - b_n)} = \frac{A_1}{A_2}. \tag{33}$$

Therefore, $x^{(1)}$ is still the optimal solution of (P2), and the number of non-integer element in $x^{(1)}$ is strictly smaller than that of $x^{(0)}$.

We can similarly prove that the $x^{(1)}$ obtained by (R2) from $x^{(0)}$ is also the optimal solution of (P2):

Therefore, we can keep adjusting with either (R1) or (R2) (Note (R1) and (R2) can be used alternatively), and get the sequences

$$x^{(0)} \xrightarrow[\text{(R2)}]{\text{(R1)}} x^{(1)} \xrightarrow[\text{(R2)}]{\text{(R1)}} x^{(2)} \xrightarrow[\text{(R2)}]{\text{(R1)}} \cdots x^{(K)}. \tag{34}$$

The number of non-integer elements in $x^{(i)}$ is at least one fewer than that of $x^{(i-1)}$ for any $i \in [\tau]$, where x^τ is the first vector whose elements are all integer. This process will stop within at most K steps. $\square$

The proof of Lemma 8 is constructive, which shows that: given $x^{(0)}$, assume that $x^{(0)}$ has L_1 1's, L_0 0's. To get an optimal solution of (P1), we can keep the 1's in $x^{(0)}$; set any $L - L_0$ elements in $x^{(0)}$ s.t. $0 < x_i^{(0)} < 1$ as 1; and set the left elements as 0. This is an optimal solution of (P1).

(P2) is a linear fractional programming, which can be efficiently solved. Here we show the method proposed in the Section 4.3.2. of [Boyd and Vandenberghe, 2004]. (P2) can be transformed into

$$\begin{aligned} (P3) \qquad & \max\ a^T y \\ & \text{s.t.} \quad 0 \leq y_i \leq z \quad \forall i \in [K]; \\ & \quad \mathbf{1}^T y - Lz = 0; \\ & \quad b^T y = 1, \\ & \quad z \geq 0. \end{aligned} \tag{35}$$

By Interior Point Method, we can solve the optimization problem (P3) in (35) within polynomial time. Denote the optimal solution of (P3) as y^* and z^*. It is obvious that $z^* > 0$. Thus, the solution of (P2) is y^*/z^*.

C Proof of (8)

The $\mathcal{X}_L(B)$ described at the beginning of Section 4 is just an order. The complete form of $\mathcal{X}_L(B)$ is denoted as $\tilde{\mathcal{X}}_L(B)$, which is shown below:

$$\tilde{\mathcal{X}}_L(B) = (\lfloor \frac{2B}{L\mu^c_{\min}} \rfloor + 1) \exp\Big\{ - \frac{B\mu^c_{\min}}{2} \Big\} + \frac{1}{L(\mu^c_{\min})^2} \exp\{L(\mu^c_{\min})^2 - 2B\mu^c_{\min}\}. \tag{36}$$

We will prove the following lemma:

Lemma 9 $\sum_{t=\mathcal{T}_L(B)+1}^{\infty} \mathbb{P}\{B_t \geq 0\} \leq \tilde{\mathcal{X}}_L(B)$.

The proof of Lemma 9 relies on the following fact:

Fact 1 (Lemma 1 of [Flajolet and Jaillet, 2015]) *Consider n random variables $X_1, X_2 \cdots, X_n$ with support in $[0,1]$, if for any $t \in [n]$, $\mathbb{E}\{X_t | X_1, \cdots, X_{t-1}\} \geq \mu$, then $\mathbb{P}\{X_1 + X_2 + \cdots + X_n \leq n\mu - a\} \leq \exp\{-\frac{2a^2}{n}\}$ for any $a \geq 0$.*

We will use two sub steps to prove Lemma 9.

($\mathcal{S}1$) *We prove that for any $t \geq \mathcal{T}_L(B)+1$, $\mathbb{P}\{B_t \geq 0\} \leq \exp\big\{ - 2(B - tL\mu^c_{\min})^2/(tL)\big\}$.*

In ($\mathcal{S}1$), temporally let $X^c_{\sigma(i),t}$ denote the cost of arm $\sigma(i)$ at round t, in which $\sigma(i)$ is the i-th smallest element in I_t. Obviously, given $i \in [L]$ and $t \geq 1$, we have that $\mathbb{E}[X^c_{\sigma(i),t} | \{X^c_{\sigma(j),t}\}_{j<i} \cup \{X^c_{\sigma(j),s}\}_{s<t, j\in[L]}] \geq \mu^c_{\min}$. Also, we can verify that for any $t \geq \mathcal{T}_L(B)+1$, $B - tL\mu^c_{\min} \leq B - (\lfloor \frac{2B}{L\mu^c_{\min}} \rfloor + 1) L\mu^c_{\min} \leq B - 2B < 0$. As a result, we have

$$\mathbb{P}\{B_t \geq 0\} = \mathbb{P}\Big\{ \sum_{s=1}^{t} \sum_{i=1}^{L} X^c_{\sigma(i),t} \leq B \Big\} = \mathbb{P}\Big\{ \sum_{s=1}^{t} \sum_{i=1}^{L} X^c_{\sigma(i),t} \leq tL\mu^c_{\min} + B - tL\mu^c_{\min} \Big\} \leq \exp\Big\{ - \frac{2(B - tL\mu^c_{\min})^2}{tL} \Big\}, \tag{37}$$

where the last inequality is obtained by Fact 1.

($\mathcal{S}2$) *Bound $\sum_{t=\mathcal{T}_L(B)+1}^{\infty} \mathbb{P}\{B_t \geq 0\}$.* According to ($\mathcal{S}1$),

$$\begin{aligned}
&\sum_{t=\mathcal{T}_L(B)+1}^{\infty} \mathbb{P}\{B_t \geq 0\} \leq \sum_{l=0}^{\infty} \exp\Big\{ - \frac{2(B + lL\mu^c_{\min})^2}{\frac{2B}{\mu^c_{\min}} + lL} \Big\} \quad \triangleright \text{ function } h(t) = -\frac{2(B - tL\mu^c_{\min})^2}{tL} \text{ decreases w.r.t } t \in [\frac{B}{L\mu^c_{\min}}, \infty) \\
\leq & (\lfloor \frac{2B}{L\mu^c_{\min}} \rfloor + 1) \exp\Big\{ - \frac{B\mu^c_{\min}}{2} \Big\} + \sum_{l=\mathcal{T}_L(B)+1}^{\infty} \exp\{-lL(\mu^c_{\min})^2\} \leq \tilde{\mathcal{X}}_L(B).
\end{aligned}$$

D Proof of the Upper Bound for $\zeta_\kappa(\mathcal{T}_L(B))$

The proof contains four steps:

1. When $\kappa > 2$: We know that $\ln x < x$ holds for any $x \in [1, \infty)$. Therefore,

$$\begin{aligned}
\zeta_\kappa(\mathcal{T}_L(B)) &\leq 1 + \sum_{t=2}^{\infty} (\log_2(t) + 1) t^{-\kappa} \leq 1 + \frac{1}{\ln 2} \sum_{t=2}^{\infty} t^{-(\kappa-1)} + \sum_{t=2}^{\infty} t^{-\kappa} \\
&\leq 1 + \frac{1}{\ln 2} \int_1^{\infty} t^{-(\kappa-1)} \mathrm{d}t + \int_1^{\infty} t^{-\kappa} \mathrm{d}t \leq \frac{1}{\ln 2} \frac{1}{\kappa - 2} + \frac{\kappa}{\kappa - 1}.
\end{aligned} \tag{38}$$

2. When $1 < \kappa \leq 2$: For any $\epsilon > 0$, we can verify that $\ln(x) < x^\epsilon$ when $x > \max\{1, (\frac{1}{\epsilon})^{\frac{1}{\epsilon}}\}$. This is because that function $x^\epsilon - \ln(x)$ monotonically increases in $[(\frac{1}{\epsilon})^{\frac{1}{\epsilon}}, \infty)$, and we can verify that $\ln(1) < 1^\epsilon$. Given any $\epsilon \in (0, 1]$, we know that $1 \leq (\frac{1}{\epsilon})^{\frac{1}{\epsilon}}$. Therefore,

$$\begin{aligned}
\zeta_\kappa(\mathcal{T}_L(B)) &\leq 1 + \sum_{t=2}^{\infty} (\log_2(t) + 1) t^{-\kappa} \leq 1 + (\frac{1}{\epsilon})^{\frac{1}{\epsilon}+1} \log_2(\frac{1}{\epsilon}) + \frac{1}{\ln 2} \sum_{t=2}^{\infty} t^{-(\kappa-\epsilon)} + \sum_{t=2}^{\infty} t^{-\kappa} \\
&\leq 1 + (\frac{1}{\epsilon})^{\frac{1}{\epsilon}+1} \log_2(\frac{1}{\epsilon}) + \frac{1}{\ln 2} \int_1^{\infty} t^{-(\kappa-\epsilon)} \mathrm{d}t + \int_1^{\infty} t^{-\kappa} \mathrm{d}t \\
&\leq (\frac{1}{\epsilon})^{\frac{1}{\epsilon}+1} \log_2(\frac{1}{\epsilon}) + \frac{1}{\ln 2} \frac{1}{\kappa - \epsilon - 1} + \frac{1}{\kappa - 1} + 1.
\end{aligned} \tag{39}$$

   By choosing $\epsilon = \frac{\kappa - 1}{2}$, (39) can be further written as

$$\zeta_\kappa(\mathcal{T}_L(B)) \leq (\frac{2}{\kappa - 1})^{\frac{\kappa+1}{\kappa-1}} \log_2(\frac{2}{\kappa - 1}) + (\frac{2}{\ln 2} + 1) \frac{1}{\kappa - 1} + 1.$$

   Case 1 & 2 show that when $\kappa > 1$, $\zeta_\kappa(\mathcal{T}_L(B))$ is related to κ only.

3. When $\kappa = 1$:

$$\zeta_\kappa(\mathcal{T}_L(B)) = \sum_{t=1}^{\mathcal{T}_L(B)} (\log_2(t)+1)t^{-1} \leq \sum_{t=1}^{\mathcal{T}_L(B)} t^{-1} + \log_2(\mathcal{T}_L(B)) \sum_{t=1}^{\mathcal{T}_L(B)} t^{-1} \leq O(\ln^2(\mathcal{T}_L(B))) = O(\ln^2 B).$$

4. When $0 < \kappa < 1$:

$$\zeta_\kappa(\mathcal{T}_L(B)) = \sum_{t=1}^{\mathcal{T}_L(B)} (\log_2(t)+1)t^{-\kappa} \leq \frac{\ln \mathcal{T}_L(B)}{\ln 2} \frac{\mathcal{T}_L^{1-\kappa}(B)}{1-\kappa} + \frac{\mathcal{T}_L^{1-\kappa}(B)}{1-\kappa} + O(1) = O\big(\frac{B^{1-\kappa}\ln B}{1-\kappa}\big).$$

E Detailed Regret Analysis of the MRCB Policy

E.1 Derivations of the Step 1 in the Proof for Theorem 3

For any algorithm under the (semi-)bandit setting, the decision at round t (i.e., I_t) should only depends on $\mathcal{H}_{t-1}$. Given any algorithm w, denote the expected reward before the budget runs out as $\tilde{R}_w$. We have that

$$\begin{aligned}
\tilde{R}_w &= \mathbb{E}\sum_{t=1}^{\infty} \Sigma(r(t), I_t)\mathbb{I}\{B_t \geq 0, I_t = I_*\} + \mathbb{E}\sum_{t=1}^{\infty} \Sigma(r(t), I_t)\mathbb{I}\{B_t \geq 0, I_t \in \mathcal{C}_s\} \\
&\geq \mathbb{E}\sum_{t=1}^{\infty} \Sigma(r(t), I_t)\mathbb{I}\{B_{t-1} \geq L, I_t = I_*\} + \mathbb{E}\sum_{t=1}^{\infty} \Sigma(r(t), I_t)\mathbb{I}\{B_{t-1} \geq L, I_t \in \mathcal{C}_s\} \\
&= \mathbb{E}\sum_{t=1}^{\infty} \varrho_L^* \Sigma(c(t), I_t)\mathbb{I}\{B_{t-1} \geq L, I_t = I_*\} + \mathbb{E}\sum_{t=1}^{\infty} \varrho_L^* \Sigma(c(t), I_t)\mathbb{I}\{B_{t-1} \geq L, I_t \in \mathcal{C}_s\} \qquad (40) \\
&\quad - \mathbb{E}\sum_{t=1}^{\infty} (\varrho_L^* \Sigma(c(t), I_t) - \Sigma(\mu^r, I_t))\mathbb{I}\{B_{t-1} \geq L, I_t \in \mathcal{C}_s\} \qquad (41) \\
&= \mathbb{E}\sum_{t=1}^{\infty} \varrho_L^* \Sigma(c(t), I_t)\mathbb{I}\{B_{t-1} \geq L\} - \mathbb{E}\sum_{t=1}^{\infty} (\varrho_L^* \Sigma(c(t), I_t) - \Sigma(\mu^c, I_t))\mathbb{I}\{B_{t-1} \geq L, I_t \in \mathcal{C}_s\} \\
&\geq (B-L)\varrho_L^* - \mathbb{E}\sum_{t=1}^{\infty} (\varrho_L^* \Sigma(c(t), I_t) - \Sigma(\mu^r, I_t))\mathbb{I}\{B_{t-1} \geq L, I_t \in \mathcal{C}_s\}.
\end{aligned}$$

In the above derivations, the first term in (40) can be obtained, because:

$$\begin{aligned}
&\mathbb{E}\sum_{t=1}^{\infty} \Sigma(r(t), I_t)\mathbb{I}\{B_{t-1} \geq L, I_t = I_*\} = \mathbb{E}\sum_{t=1}^{\infty} \Sigma(r(t), I_*)\mathbb{I}\{B_{t-1} \geq L, I_t = I_*\} = \sum_{t=1}^{\infty} \mathbb{E}\{\Sigma(r(t), I_*)\}\mathbb{P}\{B_{t-1} \geq L, I_t = I_*\} \\
=&\sum_{t=1}^{\infty} \Sigma(\mu^r, I_*)\mathbb{P}\{B_{t-1} \geq L, I_t = I_*\} = \sum_{t=1}^{\infty} \varrho_L^* \Sigma(\mu^c, I_*)\mathbb{P}\{B_{t-1} \geq L, I_t = I_*\} = \sum_{t=1}^{\infty} \varrho_L^* \mathbb{E}\{\Sigma(c(t), I_*)\}\mathbb{P}\{B_{t-1} \geq L, I_t = I_*\} \\
=&\mathbb{E}\sum_{t=1}^{\infty} \varrho_L^* \Sigma(c(t), I_*)\mathbb{I}\{B_{t-1} \geq L, I_t = I_*\} = \mathbb{E}\sum_{t=1}^{\infty} \varrho_L^* \Sigma(c(t), I_t)\mathbb{I}\{B_{t-1} \geq L, I_t = I_*\}, \qquad (42)
\end{aligned}$$

and (41) can be obtained because

$$\begin{aligned}
&\mathbb{E}\sum_{t=1}^{\infty} \Sigma(r(t), I_t)\mathbb{I}\{B_{t-1} \geq L, I_t \in \mathcal{C}_s\} = \mathbb{E}\sum_{t=1}^{\infty}\sum_{I\in\mathcal{C}_s} \Sigma(r(t), I)\mathbb{I}\{B_{t-1} \geq L, I_t = I\} = \sum_{t=1}^{\infty}\sum_{I\in\mathcal{C}_s} \mathbb{E}\{\Sigma(r(t), I)\}\mathbb{P}\{B_{t-1} \geq L, I_t = I\} \\
=&\sum_{t=1}^{\infty}\sum_{I\in\mathcal{C}_s} \Sigma(\mu^r, I)\mathbb{P}\{B_{t-1} \geq L, I_t = I\} = \mathbb{E}\sum_{t=1}^{\infty} \Sigma(\mu^r, I_t)\mathbb{I}\{B_{t-1} \geq L, I_t \in \mathcal{C}_s\}.
\end{aligned}$$

As a result,

$$\begin{aligned}
\text{Regret} &\leq R^* - \Big((B-L)\varrho_L^* - \mathbb{E}\sum_{t=1}^{\infty} (\varrho_L^* \Sigma(c(t), I_t) - \Sigma(\mu^r, I_t))\mathbb{I}\{B_{t-1} \geq L, I_t \in \mathcal{C}_s\}\Big) \\
&\leq \mathbb{E}\sum_{t=1}^{\infty} (\varrho_L^* \Sigma(c(t), I_t) - \Sigma(\mu^r, I_t))\mathbb{I}\{B_{t-1} \geq L, I_t \in \mathcal{C}_s\} + 2L\varrho_L^*. \quad \triangleright \text{ By Lemma 1, we know } R^* \leq (B+L)\varrho_L^*
\end{aligned} \qquad (43)$$

(43) can be further decomposed as

$$\begin{aligned}
\text{Regret} \leq & \mathbb{E}\sum_{t=1}^{\mathcal{T}_L(B)}(\varrho_L^*\Sigma(c(t),I_t)-\Sigma(\mu^r,I_t))\mathbb{I}\{B_{t-1}\geq L, I_t\in\mathcal{C}_s\}+\mathbb{E}\sum_{t=\mathcal{T}_L(B)+1}^{\infty}(\varrho_L^*\Sigma(c(t),I_t)-\Sigma(\mu^r,I_t))\mathbb{I}\{B_{t-1}\geq L, I_t\in\mathcal{C}_s\}+2L\varrho_L^*\\
\leq & \mathbb{E}\sum_{t=1}^{\mathcal{T}_L(B)}(\varrho_L^*\Sigma(c(t),I_t)-\Sigma(\mu^r,I_t))\mathbb{I}\{B_{t-1}\geq L, I_t\in\mathcal{C}_s\}+\mathbb{E}\sum_{t=\mathcal{T}_L(B)+1}^{\infty}L\varrho_L^*\mathbb{I}\{B_t\geq 0\}+2L\varrho_L^*\\
\leq & \mathbb{E}\sum_{t=1}^{\mathcal{T}_L(B)}(\varrho_L^*\Sigma(c(t),I_t)-\Sigma(\mu^r,I_t))\mathbb{I}\{B_{t-1}\geq L, I_t\in\mathcal{C}_s\}+L\varrho_L^*\mathcal{X}_L(B)+2L\varrho_L^* \quad \triangleright \text{According to Lemma 9}\\
\leq & \sum_{t=1}^{\mathcal{T}_L(B)}\sum_{I\in\mathcal{C}_s}(\varrho_L^*\mathbb{E}\{\Sigma(c(t),I)\}-\Sigma(\mu^r,I))\mathbb{E}\mathbb{I}\{B_{t-1}\geq L, I_t=I\}+L\varrho_L^*\mathcal{X}_L(B)+2L\varrho_L^* \quad \triangleright \text{Similar to the derivations in (42).}\\
= & \mathbb{E}\sum_{t=1}^{\mathcal{T}_L(B)}\sum_{I\in\mathcal{C}_s}(\varrho_L^*\Sigma(\mu^c,I)-\Sigma(\mu^r,I))\mathbb{I}\{I_t=I\}+L\varrho_L^*\mathcal{X}_L(B)+2L\varrho_L^* \quad \triangleright \text{Similar to the derivations in (42).}\\
= & \sum_{I\in\mathcal{C}_s}(\varrho_L^*\Sigma(\mu^c,I)-\Sigma(\mu^r,I))\mathbb{E}\{\mathcal{N}^I\}+L\varrho_L^*\mathcal{X}_L(B)+2L\varrho_L^*=\sum_{I\in\mathcal{C}_s}\Delta^I\mathbb{E}\{\mathcal{N}^I\}+L\varrho_L^*\mathcal{X}_L(B)+2L\varrho_L^*.
\end{aligned}$$

E.2 Derivations of the Step 2 in the Proof for Theorem 3

Define a counter $N_{i,t}$ for any $i\in[K]$ as follows (the same as the one in [Chen *et al.*, 2013]): If a suboptimal super arm I is pulled at round t, let $i=\arg\min_{j\in I}N_{j,t-1}$, and $N_{i,t}=N_{i,t-1}+1$. $N_{i,0}=0$ for any $i\in[K]$. If $\arg\min_{j\in I}N_{j,t-1}$ is not unique, randomly pick one from the smallest counters and increase it. According to Eqn.(18), Eqn.(28) in Appendix A of [Chen *et al.*, 2013] (, which is in the supplementary document), we can get that for any $i\in[K]$,

$$\begin{aligned}
\mathcal{R}_i \leq & 2\Delta_{\max}^i+(\varrho_L^*+1)^2\frac{2L^2(\sqrt{\kappa}+1)^2\ln[\sqrt{K^{L-1}}\mathcal{T}_L(B)]}{\Delta_{\min}^i}-(\varrho_L^*+1)^2\frac{L^2(\sqrt{\kappa}+1)^2\ln[\sqrt{K^{L-1}}\mathcal{T}_L(B)]}{\Delta_{\max}^i}+\\
& +\mathbb{E}\sum_{t=t_0}^{\mathcal{T}_L(B)}\sum_{j=1}^{K_i}\Delta^{i,j}\mathbb{I}\{I_t=S(i,j),N_{i,t}>N_{i,t-1},N_{i,t-1}>\lfloor f_{i,j}\rfloor\}.
\end{aligned}$$

The $2\Delta_{\max}^i$ in the above inequalities is obtained because the suboptimal arms in $\mathcal{S}_i$ are pulled at most twice in the first t_0-1 rounds. For the update rules of the counters, $\mathbb{I}\{I_t=S(i,j),N_{i,t}>N_{i,t-1},N_{i,t-1}>\lfloor f_{i,j}\rfloor\}$ implies $\mathbb{I}\{I_t=S(i,j),\forall k\in S(i,j),N_{k,t-1}>\lfloor f_{i,j}\rfloor\}$, and further, $\mathbb{I}\{I_t=S(i,j),\forall k\in S(i,j),T_k(t-1)>\lfloor f_{i,j}\rfloor\}$

E.3 Derivations of the Step 3 in the Proof for Theorem 3

Proof of (16) **in Step 3-1** For ease of reference, $X_{k,l}$ denotes the l-th observations of the reward of the k-th arm for any $k\in[K]$. $X_{k,l}$'s are independent for different l. $\bar{X}_{k,s}=\frac{1}{s}\sum_{l=1}^{s}X_{k,l}$. We can get that for any $k\in[K]$,

$$\begin{aligned}
& \mathbb{P}\{\tilde{\mu}_k^r(t)\leq\mu_k^r\}\leq\mathbb{P}\Big\{\hat{\mu}_k^r(t)\leq\mu_k^r-\sqrt{\kappa\frac{\ln(t-1)}{T_k(t-1)}}\Big\}\\
\leq & \mathbb{P}\Big\{\exists s\in\{1,2,\cdots,t-1\}\text{ s.t. }\bar{X}_{k,s}\leq\mu_k^r-\sqrt{\kappa\frac{\ln(t-1)}{s}}\Big\} \quad \triangleright T_k(t-1)\text{ cannot be zero, since each arm is pulled in the initial phase.}\\
= & \mathbb{P}\Big\{\exists s\in\{1,2,\cdots,t-1\}\text{ s.t. }\sum_{l=1}^{s}(X_{k,l}-\mu_k^r)\leq-\sqrt{\kappa s\ln(t-1)}\Big\}\\
\leq & \sum_{h=0}^{\lfloor\frac{\ln(t-1)}{\ln 2}\rfloor}\mathbb{P}\Big\{\exists s\in(\frac{1}{2})^{h+1}(t-1)<s\leq(\frac{1}{2})^h(t-1)\text{ s.t. }\sum_{l=1}^{s}(X_{k,l}-\mu_k^r)\leq-\sqrt{\kappa(\frac{1}{2})^{h+1}(t-1)\ln(t-1)}\Big\}\\
\leq^{\triangle} & \sum_{h=0}^{\lfloor\frac{\ln(t-1)}{\ln(2)}\rfloor}(\frac{1}{t-1})^{\kappa}\leq\Big(\log_2(t-1)+1\Big)(\frac{1}{t-1})^{\kappa},
\end{aligned}\tag{44}$$

in which the "$\leq$" marked with $^{\triangle}$ is obtained by Hoeffing's maximal inequality [Bubeck, 2010] (in Page 30). Similarly, we also have

$$\mathbb{P}\{\tilde{\mu}_k^c\geq\mu_k^c\}\leq\mathbb{P}\Big\{\hat{\mu}_k^c(t)\geq\mu_k^c+\sqrt{\kappa\frac{\ln(t-1)}{T_k(t-1)}}\Big\}\leq\Big(\log_2(t-1)+1\Big)(\frac{1}{t-1})^{\kappa};$$

By union bound, we can obtain that $\mathbb{P}\{\mathcal{Q}_o(t)\}\leq 2L\Big(\log_2(t-1)+1\Big)(t-1)^{-\kappa}$. $\square$

Detailed Proofs of Step 3-2 We give detailed proofs to bound (15). Assume $I_t = S(i,j)$ for some $i \in \mathcal{B}$ and $j \in [K_i]$.
$\triangleright$ First we prove why $\mathbb{P}\{U_{i,j}(t), \overline{\mathcal{Q}_o(t)}\}$ can be bounded by (17): if $\overline{\mathcal{Q}_o(t)}$ is true, we know that

$$\frac{\Sigma(\tilde{\mu}^r(t), I_*)}{\Sigma(\tilde{\mu}^c(t), I_*)} \geq \frac{\Sigma(\mu^r(t), I_*)}{\Sigma(\mu^c(t), I_*)}.$$

Since super arm I_t is pulled, we have that

$$\frac{\Sigma(\tilde{\mu}^r(t), I_t)}{\Sigma(\tilde{\mu}^c(t), I_t)} \geq \frac{\Sigma(\tilde{\mu}^r(t), I_*)}{\Sigma(\tilde{\mu}^c(t), I_*)} \geq \frac{\Sigma(\mu^r(t), I_*)}{\Sigma(\mu^c(t), I_*)}. \tag{45}$$

If for any $k \in I_t = S(i,j)$, we have $\tilde{\mu}_k^r(t) < \mu_k^r + \frac{\delta^{i,j}}{L}$ and $\tilde{\mu}_k^c(t) > \mu_k^c - \frac{\delta^{i,j}}{L}$, we know that

$$\frac{\Sigma(\tilde{\mu}^r(t), I_t)}{\Sigma(\tilde{\mu}^c(t), I_t)} < \frac{\Sigma(\mu^r(t), I_t) + \delta^{i,j}}{\Sigma(\mu^c(t), I_t) - \delta^{i,j}} = \frac{\Sigma(\mu^r(t), I_*)}{\Sigma(\mu^c(t), I_*)} \leq \frac{\Sigma(\tilde{\mu}^r(t), I_*)}{\Sigma(\tilde{\mu}^c(t), I_*)}. \tag{46}$$

(45) and (46) are contradictions. That is, at round t, $\{U_{i,j}(t), \overline{\mathcal{Q}_o(t)}\}$ implies that the following event is true:

$$\Big\{ \bigcup_{k \in S(i,j)} \{\tilde{\mu}_k^r(t) \geq \mu_k^r + \frac{\delta^{i,j}}{L}, T_k(t-1) > \lfloor f_{i,j} \rfloor\} \cup \bigcup_{k \in S(i,j)} \{\tilde{\mu}_k^c(t) \leq \mu_k^c - \frac{\delta^{i,j}}{L}, T_k(t-1) > \lfloor f_{i,j} \rfloor\} \Big\}. \tag{47}$$

Thus, $\mathbb{P}\{U_{i,j}(t), \overline{\mathcal{Q}_o(t)}\}$ can be bounded by (17).
$\triangleright$ We will prove the two inequalities below (17). Again, temporally let $\{X_{k,l}\}_{l=1}^{\infty}$ denote the independent reward observations from arm $k \in [K]$, and let $\bar{X}_{k,s}$ denote the average of the s independent observations, i.e, $\bar{X}_{k,s} = \frac{1}{s}\sum_{l=1}^{s} X_{k,l}$. For any $k \in S(i,j)$, we have that

$$\begin{aligned}
&\mathbb{P}\{\tilde{\mu}_k^r(t) \geq \mu_k^r + \frac{\delta^{i,j}}{L}, T_k(t-1) > \lfloor f_{i,j} \rfloor\} \\
\leq &\mathbb{P}\{\min\{\hat{\mu}_k^r(t) + \mathcal{E}_{k,t}^{\kappa}, 1\} \geq \mu_k^r + \frac{\delta^{i,j}}{L}, T_k(t-1) > \lfloor f_{i,j} \rfloor\} \\
\leq &\mathbb{P}\{\hat{\mu}_k^r(t) + \mathcal{E}_{k,t}^{\kappa} \geq \mu_k^r + \frac{\delta^{i,j}}{L}, T_k(t-1) > \lfloor f_{i,j} \rfloor\} \leq^{\triangle} \mathbb{P}\{\hat{\mu}_k^r(t) \geq \mu_k^r + \frac{1}{\sqrt{\kappa}+1}\frac{\delta^{i,j}}{L}, T_k(t-1) > \lfloor f_{i,j} \rfloor\} \qquad (48) \\
\leq &\sum_{l=\lfloor f_{i,j} \rfloor + 1}^{\mathcal{T}_L(B)} \mathbb{P}\{\bar{X}_{k,l} \geq \mu_k^r + \frac{1}{\sqrt{\kappa}+1}\frac{\delta^{i,j}}{L}\} \leq \mathcal{T}_L(B) \exp\Big\{ -2\ln[\sqrt{K^{L-1}}\mathcal{T}_L(B)] \Big\} \qquad \triangleright \text{ According to Hoeffding's inequality} \\
\leq &\frac{1}{K^{L-1}\mathcal{T}_L(B)}. \qquad (49)
\end{aligned}$$

Note that the inequality marked with $^{\triangle}$ in (48) hold because

$$\frac{\delta^{i,j}}{L} - \mathcal{E}_{i,t}^{\kappa} \geq \frac{\delta^{i,j}}{L} - \sqrt{\kappa \frac{(\delta^{i,j})^2 \ln \mathcal{T}_L(B)}{L^2(\sqrt{\kappa}+1)^2 \ln[\sqrt{K^{L-1}}\mathcal{T}_L(B)]}} \geq \frac{1}{\sqrt{\kappa}+1}\frac{\delta^{i,j}}{L}.$$

Similarly, we have that for any $k \in S(i,j)$,

$$\mathbb{P}\{\tilde{\mu}_k^c(t) \leq \mu_k^c - \frac{\delta^{i,j}}{L}, T_k(t-1) > \lfloor f_{i,j} \rfloor\} \leq \frac{1}{K^{L-1}\mathcal{T}_L(B)}. \tag{50}$$

Thus, according to (49), (50) and union bound, we have that (17) can be bounded as

$$\mathbb{P}\{U_{i,j}(t), \overline{\mathcal{Q}_o(t)}\} \leq \frac{2L}{K^{L-1}\mathcal{T}_L(B)}. \tag{51}$$

Accordingly,

$$\mathbb{E}\sum_{t=1}^{\mathcal{T}_L(B)}\sum_{j=1}^{K_i} \Delta^{i,j}\mathbb{I}\{U_{i,j}(t), \overline{\mathcal{Q}_o(t)}\} \leq \Delta_{\max}^i \sum_{t=1}^{\mathcal{T}_L(B)}\sum_{j=1}^{K_i} \mathbb{P}\{U_{i,j}(t), \overline{\mathcal{Q}_o(t)}\} \leq 2L\Delta_{\max}^i. \tag{52}$$

F Proofs about the Lower Bound

The proof for Theorem 4 consists of three main steps:

F.1 Proof of the Non-Empty of the Constraint Set in (19)

Given a suboptimal super arm $S(i,j)$ where $i \notin I_*$, $j \in [K_i]$, to make $\mu_i^r + \delta^{i,j}(\gamma) < 1$ and $\mu_i^c - \gamma\delta^{i,j}(\gamma) > 0$, we must have

$$\begin{cases} \mu_i^r + \dfrac{\Delta^{i,j}}{\gamma\varrho_L^* + 1} < 1; \\ \mu_i^c - \dfrac{\gamma\Delta^{i,j}}{\gamma\varrho_L^* + 1} > 0. \end{cases} \tag{53}$$

Define the solution set of (53) as $\mathcal{D}^{i,j}$. If $\mu_i^r + \Delta^{i,j} \leq 1$ or $\mu_i^c - \frac{\Delta^{i,j}}{\varrho_L^*} \geq 0$, $\mathcal{D}^{i,j}$ is not empty. Otherwise, to make $\mathcal{D}^{i,j}$ non-empty, γ has to satisfy that $\gamma''(i,j) < \gamma < \gamma'(i,j)$, where

$$\gamma''(i,j) = (\frac{\Delta^{i,j}}{1-\mu_i^r} - 1)\frac{\Sigma(\mu^c, I_*)}{\Sigma(\mu^r, I_*)}; \quad \gamma'(i,j) = \frac{\mu_i^c}{\Delta^{i,j} - \mu_i^c\varrho_L^*}. \tag{54}$$

Then we only need to ensure that $\gamma''(i,j) < \gamma'(i,j)$, which is equivalent to

$$\frac{\Sigma(\mu^r, I_*)}{\Sigma(\mu^c, I_*)} < \frac{\Sigma(\mu^r, S(i,j)) + 1 - \mu_i^r}{\Sigma(\mu^c, S(i,j)) - \mu_i^c}. \tag{55}$$

Given any fixed $k \in I_*$, let $I_{\setminus k}$ denote $\{I_* \setminus \{k\}\} \cup \{i\}$. It is obvious that $I_{\setminus k} = S(i,j')$ for some $j' \in [K_i]$. Next we will use two steps to discuss whether $\mathcal{D}^{i,j'}$ is empty or not.

1. When $\mu_i^r + \Delta^{i,j'} \leq 1$ or $\mu_i^c - \frac{\Delta^{i,j'}}{\varrho_L^*} \geq 0$: $\mathcal{D}^{i,j'}$ is certainly non-empty. Correspondingly, we can find the proper γ_k s.t.

$$\frac{\Sigma(\mu^r, I_*)}{\Sigma(\mu^c, I_*)} = \frac{\Sigma(\mu^r, I_{\setminus k}) + \delta^{i,j'}(\gamma_k)}{\Sigma(\mu^c, I_{\setminus k}) - \gamma_k\delta^{i,j'}(\gamma_k)}; \quad \delta^{i,j'}(\gamma_k) < 1 - \mu_i^r; \quad \gamma_k\delta^{i,j'}(\gamma_k) < \mu_i^c. \tag{56}$$

2. When $\mu_i^r + \Delta^{i,j'} > 1$ and $\mu_i^c - \frac{\Delta^{i,j'}}{\varrho_L^*} < 0$: We can verify that

$$\Sigma(\mu^r, I_*) = \sum_{l \in I_* \setminus \{k\}} \mu_l^r + \mu_k^r < \sum_{l \in I_* \setminus \{k\}} \mu_l^r + [\mu_i^r + (1 - \mu_i^r)] = \Sigma(\mu^r, I_{\setminus k}) + 1 - \mu_i^r;$$
$$\Sigma(\mu^c, I_*) = \sum_{l \in I_* \setminus \{k\}} \mu_l^c + \mu_k^c > \sum_{l \in I_* \setminus \{k\}} \mu_l^c + [\mu_i^c - \mu_i^c] = \Sigma(\mu^c, I_{\setminus k}) - \mu_i^c.$$

Therefore, (55) holds for the super arm $I_{\setminus k}$. Correspondingly, we can also find the proper γ_k satisfying (56).

Since $\delta_{\min}^i(\gamma) \leq \delta^{i,j}(\gamma)$ for any $j \in [K_i]$, we have that (56) implies that $\delta_{\min}^i(\gamma_k) < 1 - \mu_i^r$ and $\gamma_k\delta_{\min}^i(\gamma_k) < \mu_i^c$, from which we know that the constraint set of (19) not empty.

F.2 Proof of the Existence of the Optimal Solution and the Optimal Value of (19)

Assume the $i \notin I_*$ is fixed in this section. Let $\mathcal{L}_i(\gamma)$ denote $kl(\mu_i^r, \mu_i^r + \delta_{\min}^i(\gamma)) + kl(\mu_i^c, \mu_i^c - \gamma\delta_{\min}^i(\gamma))$. Let $\mathcal{D}_i$ denote the constraint set of (19). Throughout this section, let super arm I denote $S(i, K_i)$. Remind that we have already assumed that $\Delta^{i,1} \geq \Delta^{i,2} \geq \cdots \Delta^{i,K_i}$. Minimizing $\mathcal{L}_i(\gamma)$ w.r.t $\gamma \in \mathcal{D}_i$ is equivalent to maximizing

$$\begin{aligned} \tilde{\mathcal{L}}_i(\gamma) = {} & \mu_i^r \ln\left(\mu_i^r + \frac{\Delta_{\min}^i}{\gamma\varrho_L^* + 1}\right) + (1 - \mu_i^r)\ln\left(1 - \mu_i^r - \frac{\Delta_{\min}^i}{\gamma\varrho_L^* + 1}\right) \\ & + \mu_i^c \ln\left(\mu_i^c - \frac{\gamma\Delta_{\min}^i}{\gamma\varrho_L^* + 1}\right) + (1 - \mu_i^c)\ln\left(1 - \mu_i^c + \frac{\gamma\Delta_{\min}^i}{\gamma\varrho_L^* + 1}\right) \end{aligned} \tag{57}$$

w.r.t $\gamma \in \mathcal{D}_i$. The first order derivative of $\tilde{\mathcal{L}}_i(\gamma)$ w.r.t γ is:

$$\frac{\partial\tilde{\mathcal{L}}_i(\gamma)}{\partial\gamma} = \frac{\Delta_{\min}^i}{(\gamma\varrho_L^* + 1)^2}\left\{ - \frac{\mu_i^r\varrho_L^*}{\mu_i^r + \frac{\Delta_{\min}^i}{\gamma\varrho_L^* + 1}} + \frac{(1 - \mu_i^r)\varrho_L^*}{1 - \mu_i^r - \frac{\Delta_{\min}^i}{\gamma\varrho_L^* + 1}} - \frac{\mu_i^c}{\mu_i^c - \frac{\gamma\Delta_{\min}^i}{\gamma\varrho_L^* + 1}} + \frac{1 - \mu_i^c}{1 - \mu_i^c + \frac{\gamma\Delta_{\min}^i}{\gamma\varrho_L^* + 1}} \right\}. \tag{58}$$

In the following context, let $W(\gamma)$ denote terms within $\{\cdot\}$ of (58). One can easily verify that $W(\gamma)$ decreases w.r.t $\gamma \in \mathbb{R}$, thus, decreases w.r.t $\gamma \in \mathcal{D}_i$. According to the derivations in Appendix F.1, we know $\mathcal{D}_i$ is a continuous region.

The proof of the existence of the optimal solution is divided into the following four cases:

1. When $\mu_i^r + \Delta_{\min}^i \leq 1$ and $\mu_i^c - \frac{\Delta_{\min}^i}{\varrho_L^*} \leq 0$: We can verify that

$$\lim_{\gamma\to 0_+}\Big\{-\frac{\mu_i^r\varrho_L^*}{\mu_i^r+\frac{\Delta_{\min}^i}{\gamma\varrho_L^*+1}}+\frac{(1-\mu_i^r)\varrho_L^*}{1-\mu_i^r-\frac{\Delta_{\min}^i}{\gamma\varrho_L^*+1}}\Big\}=-\frac{\mu_i^r\varrho_L^*}{\mu_i^r+\Delta_{\min}^i}+\frac{(1-\mu_i^r)\varrho_L^*}{1-\mu_i^r-\Delta_{\min}^i}=\frac{\Delta_{\min}^i\varrho_L^*}{(\mu_i^r+\Delta_{\min}^i)(1-\mu_i^r-\Delta_{\min}^i)}>0. \quad (59)$$

Please note that in the boundary case that $\mu_i^r + \Delta_{\min}^i = 1$, (59) will tend to $+\infty$, which is still positive.
Let γ' denote $\mu_i^c - (\gamma'\Delta_{\min}^i)/(\gamma'\varrho_L^* + 1) = 0$. Note that γ' can be $+\infty$. We have that $\lim_{\gamma\to 0_+} W(\gamma) > 0$ and $\lim_{\gamma\to\gamma'_-} W(\gamma) < 0$. According to *intermediate value theorem* [6], there exists a unique γ^* s.t. $\frac{\partial\tilde{\mathcal{L}}_i(\gamma)}{\partial\gamma}\big|_{\gamma=\gamma^*} = 0$. The γ^* is the optimal solution of (19).

2. When $\mu_i^r + \Delta_{\min}^i \leq 1$ and $\mu_i^c - \frac{\Delta_{\min}^i}{\varrho_L^*} > 0$: First we have that $\lim_{\gamma\to 0_+} W(\gamma) > 0$. Since

$$\lim_{\gamma\to\infty}\Big\{-\frac{\mu_i^c}{\mu_i^c-\frac{\gamma\Delta_{\min}^i}{\gamma\varrho_L^*+1}}+\frac{1-\mu_i^c}{1-\mu_i^c+\frac{\gamma\Delta_{\min}^i}{\gamma\varrho_L^*+1}}\Big\}=-\frac{\mu_i^c}{\mu_i^c-\frac{\Delta_{\min}^i}{\varrho_L^*}}+\frac{1-\mu_i^c}{1-\mu_i^c+\frac{\Delta_{\min}^i}{\varrho_L^*}}=-\frac{\Delta_{\min}^i}{\varrho_L^*(\mu_i^c-\frac{\Delta_{\min}^i}{\varrho_L^*})(1-\mu_i^c+\frac{\Delta_{\min}^i}{\varrho_L^*})}<0, \quad (60)$$

we know $\lim_{\gamma\to\infty} W(\gamma) < 0$. According to intermediate value theorem, there exists a unique γ^* s.t. $\frac{\partial\tilde{\mathcal{L}}_i(\gamma)}{\partial\gamma}\big|_{\gamma=\gamma^*} = 0$. The γ^* is the optimal solution of (19).

3. When $\mu_i^c - \frac{\Delta_{\min}^i}{\varrho_L^*} \geq 0$ and $\mu_i^r + \Delta_{\min}^i > 1$: Let γ'' denote $\mu_i^r + \Delta_{\min}^i/(\gamma''\varrho_L^* + 1) = 1$. It is obvious that $\lim_{\gamma\to\gamma''_+} W(\gamma) > 0$ and $\lim_{\gamma\to\infty} W(\gamma) < 0$. Again, using intermediate value theorem, there exists a unique γ^* s.t. $\frac{\partial\tilde{\mathcal{L}}_i(\gamma)}{\partial\gamma}\big|_{\gamma=\gamma^*} = 0$. The γ^* is the optimal solution of (19).

4. When $\mu_i^c - \frac{\Delta_{\min}^i}{\varrho_L^*} < 0$ and $\mu_i^r + \Delta_{\min}^i > 1$: According to the above derivations, we have that $\lim_{\gamma\to\gamma'_-} W(\gamma) < 0$ and $\lim_{\gamma\to\gamma''_+} W(\gamma) > 0$. According to intermediate value theorem, there exists a γ^* s.t. $\frac{\partial\tilde{\mathcal{L}}_i(\gamma)}{\partial\gamma}\big|_{\gamma=\gamma^*} = 0$. The γ^* is the unique optimal solution of (19).

F.3 Proof of the Lower Bound for Bernoulli MP-BMABs

In this section, let Γ_i denote $\Gamma_i^w(B/L)$, and let Γ_I denote $\Gamma_I^w(B/L)$ for ease of reference. We will prove the lower bound of Γ_i for the any given $i \notin I_*$. Denote the optimal solution of (19) as γ^*. We have that $\mu_i^r + \delta_{\min}^i(\gamma^*) < 1$, $\mu_i^c - \gamma^*\delta_{\min}^i(\gamma^*) > 0$ and $\gamma^* \geq 0$. For any $\rho > 0$, we can always find an x^r and an x^c such that

$$kl(\mu_i^r, \mu_i^r + \delta_{\min}^i(\gamma^*) + x^r) < (1+\rho)kl(\mu_i^r, \mu_i^r + \delta_{\min}^i(\gamma^*)), \text{ and } x^r > 0,\ \mu_i^r + \delta_{\min}^i(\gamma^*) + x^r < 1;$$
$$kl(\mu_i^c, \mu_i^c - \gamma^*\delta_{\min}^i(\gamma^*) - x^c) < (1+\rho)kl(\mu_i^c, \mu_i^c - \gamma^*\delta_{\min}^i(\gamma^*)), \text{ and } x^c > 0,\ \mu_i^c - \gamma^*\delta_{\min}^i(\gamma^*) - x^c > 0.$$

Define $\chi^r = \mu_i^r + \delta_{\min}^i(\gamma^*) + x^r$ and $\chi^c = \mu_i^c - \gamma^*\delta_{\min}^i(\gamma^*) - x^c$. Define $\tilde{D}_i = kl(\mu_i^r, \chi^r) + kl(\mu_i^c, \chi^c)$. Define a modified bandit as follows: For any arm $j \in [K]\backslash\{i\}$, the reward and cost distributions are Bern(μ_j^r) and Bern(μ_j^c); The reward and cost distributions for arm i are Bern(χ^r) and Bern(χ^c), where Bern(p) represents the Bernoulli distribution with success probability $p \in (0,1)$. 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. Please note that the best super arm of the modified bandit (denoted as $\tilde{I}_*$) is different from that of the original one[7], because

$$\frac{\Sigma(\mu^r, S(i,K_i)\backslash\{i\}) + \chi^r}{\Sigma(\mu^c, S(i,K_i)\backslash\{i\}) + \chi^c} > \frac{\Sigma(\mu^r, S(i,K_i)) + \delta_{\min}^i(\gamma^*)}{\Sigma(\mu^c, S(i,K_i)) - \gamma^*\delta_{\min}^i(\gamma^*)} = \frac{\Sigma(\mu^r, I_*)}{\Sigma(\mu^c, I_*)}. \quad (61)$$

Without the modified arm i, the maximum achievable ratio is ϱ_L^*. Arm i must belong to $\tilde{I}_*$. Thus, if arm i is not pulled, the suboptimal super arms are certainly pulled. Therefore, according to the assumption of the policy in Theorem 4, we have

[6] See http://en.wikipedia.org/wiki/Intermediate_value_theorem for a quick introduction.

[7] Please note that if there are more than one "best super arm" with the modified arm i, put these super arms in a set $\tilde{O}_i^*$. Denote the original bandit with parameters μ_i^r and μ_i^c $\forall i \in [K]$ as $\mathscr{B}_0$. Denote the "modified bandit" obtained by modifying arm i from $\mathscr{B}_0$ as $\mathscr{B}_i'(\mathscr{B}_0)$ (i.e., obtained by the method listed under the title of Subsection F.3). Given a super arm $sa \in \tilde{O}_i^*$, we can find a group of arms Y s.t. $Y \subset sa$ but Y does not belong to x for any $x \in \tilde{O}_i^*\backslash\{sa\}$. Construct a K-armed bandit derived from $\mathscr{B}_0$ as follows (denoted as $\mathscr{B}_z^{sa}(\mathscr{B}_0)$): for any $i \in [K]\backslash Y$, the rewards and costs of arm i follow Bern(μ_i^r) and Bern(μ_i^c) *resp.*; for any $j \in Y$, the rewards and costs of arm j follow Bern$(\mu_j^r + z)$ and Bern$(\mu_j^c - z)$ *resp.* where $z > 0$. When z is very close to zero, (1) the I_*'s obtained from both $\mathscr{B}_0$ and $\mathscr{B}_z^{sa}(\mathscr{B}_0)$ are the same and unique; (2) there is only one $\tilde{I}_*$ obtained from $\mathscr{B}_i'(\mathscr{B}_z^{sa}(\mathscr{B}_0))$; (3) the $\tilde{I}_*$ obtained from $\mathscr{B}_i'(\mathscr{B}_z^{sa}(\mathscr{B}_0))$ belongs to the set formed by the $\tilde{I}_*$ from $\mathscr{B}_i'(\mathscr{B}_0)$. By letting $z \to 0$, we can remove the effect brought by z.

$\mathbb{E}_{M'}\big\{(B/L)-\Gamma_i\big\} \le \sum_{I\in\mathcal{C}_s}\mathbb{E}_{M'}\{\Gamma_I\} = o((B/L)^a)$. On the other hand, we can obtain that with $0<a<\rho$,

$$\begin{aligned}
\mathbb{E}_{M'}\{B/L-\Gamma_i\} &= \mathbb{E}_{M'}\Big\{(B/L-\Gamma_i)|\Gamma_i < \frac{(1-\rho)\ln(B/L)}{\tilde{D}_i}\Big\}\mathbb{P}_{M'}\Big\{\Gamma_i < \frac{(1-\rho)\ln(B/L)}{\tilde{D}_i}\Big\} \\
&\quad + \mathbb{E}_{M'}\Big\{(B/L-\Gamma_i)|\Gamma_i \ge \frac{(1-\rho)\ln(B/L)}{\tilde{D}_i}\Big\}\mathbb{P}_{M'}\Big\{\Gamma_i \ge \frac{(1-\rho)\ln(B/L)}{\tilde{D}_i}\Big\} \\
&\ge \mathbb{E}_{M'}\Big\{(B/L-\Gamma_i)|\Gamma_i < \frac{(1-\rho)\ln(B/L)}{\tilde{D}_i}\Big\}\mathbb{P}_{M'}\Big\{\Gamma_i < \frac{(1-\rho)\ln(B/L)}{\tilde{D}_i}\Big\} \\
&\ge \Big(B/L-\frac{(1-\rho)\ln(B/L)}{\tilde{D}_i}\Big)\mathbb{P}_{M'}\Big\{\Gamma_i < \frac{(1-\rho)\ln(B/L)}{\tilde{D}_i}\Big\}.
\end{aligned}$$

Accordingly, we can obtain that

$$\mathbb{P}_{M'}\Big\{\Gamma_i < \frac{(1-\rho)\ln(B/L)}{\tilde{D}_i}\Big\} = o((B/L)^{a-1}). \tag{62}$$

Let $\{X^r_{i,t}\}_{t\in[B/L]}$ and $\{X^c_{i,t}\}_{t\in[B/L]}$ denote the successive reward and cost observations from the arm i. Define

$$\mathscr{L}(m) = \sum_{t=1}^{m}\ln\frac{\mu^r_i X^r_{i,t}+(1-\mu^r_i)(1-X^r_{i,t})}{\chi^r X^r_{i,t}+(1-\chi^r)(1-X^r_{i,t})} + \sum_{t=1}^{m}\ln\frac{\mu^c_i X^c_{i,t}+(1-\mu^c_i)(1-X^c_{i,t})}{\chi^c X^c_{i,t}+(1-\chi^c)(1-X^c_{i,t})}.$$

An important property is: For any event A in the σ-algebra generated by $\{X^r_{j,t}\}_{j\in[K],t\in[B/L]}$ and $\{X^c_{j,t}\}_{j\in[K],t\in[B/L]}$, the following change-of-measure identity holds:[8]

$$\mathbb{P}_{M'}\{A\} = \mathbb{E}_M\{\mathbb{I}\{A\}\exp\{-\mathscr{L}(\Gamma_i)\}\}. \tag{63}$$

According to (62), we know that $\mathbb{P}_{M'}(\xi) = o((B/L)^{a-1})$, where

$$\xi = \Big\{\Gamma_i < \frac{(1-\rho)\ln(B/L)}{\tilde{D}_i} \text{ and } \mathscr{L}(\Gamma_i) \le (1-a)\ln(B/L)\Big\}. \tag{64}$$

Given $n_1, n_2, \cdots, n_K$ s.t. $\sum_{j=1}^K n_j = B$, $0\le n_j \le B/L\ \forall j\in[K]$, we have

$$\begin{aligned}
&\mathbb{P}_{M'}\{\Gamma_1=n_1,\Gamma_2=n_2,\cdots,\Gamma_K=n_K,\mathscr{L}(\Gamma_i)\le(1-a)\ln(B/L)\} \\
=&\mathbb{E}_M\{\mathbb{I}\{\Gamma_1=n_1,\Gamma_2=n_2,\cdots,\Gamma_K=n_K,\mathscr{L}(\Gamma_i)\le(1-a)\ln(B/L)\}\exp\{-\mathscr{L}(\Gamma_i)\}\} \\
\ge&\exp(-(1-a)\ln(B/L))\mathbb{P}_M\{\Gamma_1=n_1,\Gamma_2=n_2,\cdots,\Gamma_K=n_K,\mathscr{L}(\Gamma_i)\le(1-a)\ln(B/L)\}.
\end{aligned} \tag{65}$$

Since

$$\xi = \bigcup_{\sum_{j=1}^K n_j=B;\Gamma_i<\frac{(1-\rho)\ln(B/L)}{\tilde{D}_i}} \Big\{\Gamma_1=n_1,\Gamma_2=n_2,\cdots,\Gamma_K=n_K,\mathscr{L}(\Gamma_i)\le(1-a)\ln(B/L)\Big\}, \tag{66}$$

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

$$\mathbb{P}_M(\xi) \le (B/L)^{1-a}\mathbb{P}_{M'}(\xi)\to 0 \text{ as } B\to\infty. \tag{67}$$

By the strong law of large numbers, $\frac{\mathscr{L}(m)}{m}\to\tilde{D}_i>0$, therefore, according to Lemma 10.5 in [Bubeck, 2010], as $m\to\infty$

$$\max_{l\le m}\frac{\mathscr{L}(l)}{m}\to\tilde{D}_i \qquad \text{a.s. } [\mathbb{P}_M]. \tag{68}$$

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

$$\mathbb{P}_M\Big\{\mathscr{L}(l)>(1-a)\ln(B/L) \text{ for some } l<\frac{(1-\rho)\ln(B/L)}{\tilde{D}_i}\Big\}\to 0 \tag{69}$$

as $B\to\infty$. Therefore,

$$\begin{aligned}
\lim_{B\to\infty}\mathbb{P}_M\Big\{\Gamma_i<\frac{(1-\rho)\ln(B/L)}{\tilde{D}_i}\Big\} &= \lim_{B\to\infty}\mathbb{P}_M\Big\{\Gamma_i<\frac{(1-\rho)\ln(B/L)}{\tilde{D}_i},\mathscr{L}(\Gamma_i)>(1-a)\ln(B/L)\Big\} \\
&\quad + \lim_{B\to\infty}\mathbb{P}_M\{\xi\} = \lim_{B\to\infty}\mathbb{P}_M\{\xi\} = 0.
\end{aligned}$$

[8] Please refer to Eqn. (2.10) of [Bubeck and Cesa-Bianchi, 2012], or Eqn. (2.6) of [Lai and Robbins, 1985] for more details. We have already assumed that the rewards and costs are independent, i.e., their joint distributions can be written in the product form. Therefore, we can safely use (63).

As a result, we have

$$\lim_{B\to\infty}\mathbb{P}_M\Big\{\Gamma_i<\frac{1-\rho}{1+\rho}\frac{\ln(B/L)}{\mathcal{L}_i^*}\Big\}\le\lim_{B\to\infty}\mathbb{P}_M\Big\{\Gamma_i<\frac{(1-\rho)\ln(B/L)}{\tilde{D}_i}\Big\}=0. \tag{70}$$

Therefore, by replacing $\frac{1-\rho}{1+\rho}$ with $1-\epsilon$, we can obtain Theorem 4. By letting $\epsilon\to 0$, we get the second claim.

Next we will prove the regret stated in the words below Theorem 4. The proof consists of three steps:
(S1) We will prove that *Regret* $\ge\sum_{I\in\mathcal{C}_s}\mathbb{E}\{\Gamma_I(B/L)\}\Delta^I-2L\varrho_L^*$.
Proof: Obviously, the expected reward of $\mathcal{M}_g$ is a lower bound for R^*. Similar to the derivations in Section E.1, we have

$$\begin{aligned}
\text{Regret}&\ge(B-L)\varrho_L^*-\mathbb{E}\sum_{t=1}^{\infty}\sum_{I\in\mathcal{C}_L^K}\Sigma(r(t),I)\mathbb{I}\{B_t\ge 0,I_t=I\}\ge(B-L)\varrho_L^*-\mathbb{E}\sum_{t=1}^{\infty}\sum_{I\in\mathcal{C}_L^K}\Sigma(r(t),I)\mathbb{I}\{B_{t-1}\ge 0,I_t=I\}\\
&\ge(B-L)\varrho_L^*-\mathbb{E}\sum_{t=1}^{\infty}\sum_{I\in\mathcal{C}_L^K}\Sigma(\mu^r,I)\mathbb{I}\{B_{t-1}\ge 0,I_t=I\}\\
&\ge\mathbb{E}\sum_{t=1}^{\infty}\sum_{I\in\mathcal{C}_L^K}\varrho_L^*\Sigma(c(t),I)\mathbb{I}\{B_{t-1}\ge 0,I_t=I\}-2L\varrho_L^*-\mathbb{E}\sum_{t=1}^{\infty}\sum_{I\in\mathcal{C}_L^K}\Sigma(\mu^r,I)\mathbb{I}\{B_{t-1}\ge 0,I_t=I\}\quad\triangleright\text{ By (25)}\\
&=\mathbb{E}\sum_{t=1}^{\infty}\sum_{I\in\mathcal{C}_s}\Delta^I\mathbb{I}\{B_t\ge 0,I_t=I\}-2L\varrho_L^*\\
&\ge\mathbb{E}\sum_{t=1}^{B/L}\sum_{I\in\mathcal{C}_s}\Delta^I\mathbb{I}\{I_t=I\}-2L\varrho_L^*\quad\triangleright\text{ The budget cannot run out at the first }B/L\text{ rounds.}\\
&\ge\sum_{I\in\mathcal{C}_s}\mathbb{E}\{\Gamma_I(B/L)\}\Delta^I-2L\varrho_L^*.\quad\triangleright\ \Gamma_I(T)\text{ represent the pulling number of super arm }I\text{ at the first }T\text{ rounds.}\quad\square
\end{aligned} \tag{71}$$

(S2) We will prove that $\sum_{I\in\mathcal{C}_s}\mathbb{E}\{\Gamma_I(B/L)\}\Delta^I\ge\sum_{i\notin I_*}\mathbb{E}\{\Gamma_i(B/L)\}\Delta^i_{\min}$.
Proof. For any $I\in\mathcal{C}_s$, denote the element in I as $I=\{i_1^*,i_2^*,\cdots,i_m^*,i_1,i_2,\cdots,i_n\}$, where $i_k^*\in I_*$ for any $k\in[m]$, $i_k\notin I_*$ for any $k\in[n]$ and $m+n=L$. Assume $I_*\backslash I=\{i_{m+1}^*,i_{m+2}^*,\cdots,i_{m+n}^*\}$. Temporally define the following super arms: $S_k^I=I_*\cup\{i_k\}\backslash\{i_{m+k}^*\}$ for any $k\in[n]$. Since

$$\begin{aligned}
\Delta^I&=\Big[\sum_{k=1}^{m}\mu_{i_k^*}^c+\sum_{k=1}^{n}\mu_{i_k}^c\Big]\varrho_L^*-\Big[\sum_{k=1}^{m}\mu_{i_k^*}^r+\sum_{k=1}^{n}\mu_{i_k}^r\Big]\\
\sum_{k=1}^{n}\Delta^{S_k^I}&=\Big[\sum_{k=1}^{n}\mu_{i_k}^c+\sum_{k=1}^{m}\mu_{i_k^*}^c+(n-1)\Sigma(\mu^c,I_*)\Big]\varrho_L^*-\Big[\sum_{k=1}^{n}\mu_{i_k}^r+\sum_{k=1}^{m}\mu_{i_k^*}^r+(n-1)\Sigma(\mu^r,I_*)\Big],
\end{aligned}$$

we can verify that $\Delta^I=\sum_{k=1}^{n}\Delta^{S_k^I}$. Please also note that for each super arm S_k^I, there is only one element in $I\backslash I_*$. Accordingly,

$$\sum_{I\in\mathcal{C}_s}\Gamma_I(B/L)\Delta^I=\sum_{I\in\mathcal{C}_s}\Gamma_I(B/L)\sum_{k=1}^{|I\backslash I_*|}\Delta^{S_k^I}\ge\sum_{I\in\mathcal{C}_s}\Gamma_I(B/L)\sum_{i\in I\backslash I_*}\Delta^i_{\min}=\sum_{I\in\mathcal{C}_s}\sum_{i\notin I_*}\Gamma_I(B/L)\Delta^i_{\min}\mathbb{I}\{i\in I\}=\sum_{i\notin I_*}\Gamma_i(B/L)\Delta^i_{\min}.$$

(S3) According to (S1) and (S2), we have that the regret is lower bound by $\Omega(\sum_{i\notin I_*}\Gamma_i(B/L)\Delta^i_{\min})$, which can be further written as $\Omega(\sum_{i\notin I_*}(\Delta^i_{\min}/\mathcal{L}_i^*)\ln(B/L))$.

G Omitted Proofs for Single-Play BMAB

G.1 Proof of the Example 5

By setting $\kappa=2$, the coefficient of arm i before $\ln B$ in Theorem 3 is

$$(\varrho_L^*+1)^2L^2(\sqrt{\kappa}+1)^2(2/\Delta^i_{\min}-1/\Delta^i_{\max})\le(\sqrt{2}+1)^2(\frac{1-p}{p}+1)^2L^2\frac{2}{\Delta^i_{\min}}\le\frac{12L^2}{p^2\Delta^i_{\min}}. \tag{72}$$

For the lower bound, we know that $\gamma=1$ belongs to the constraint in (19) for any $i\notin I_*$, because

$$\mu_i^r+\delta^i_{\min}(1)\le 1-p+\frac{p/2}{\varrho_L^*+1}<1-p+\frac{p}{2}=1-\frac{p}{2}<1;\text{ and }\mu_i^c-\delta^i_{\min}(1)\ge p-\frac{p/2}{\varrho_L^*+1}>p-\frac{p}{2}>0.$$

Thus, according to the fact that

$$2(x-y)^2 \le kl(x,y) \le \frac{(x-y)^2}{y(1-y)}, \quad \text{for any } x,y \in (0,1), \tag{73}$$

we have

$$\begin{aligned} &kl(\mu_i^r, \mu_i^r + \delta_{\min}^i(1)) + kl(\mu_i^c, \mu_i^c - \delta_{\min}^i(1)) \\ \le &\frac{[\delta_{\min}^i(1)]^2}{(\mu_i^r + \delta_{\min}^i(1))[1-(\mu_i^r + \delta_{\min}^i(1))]} + \frac{[\delta_{\min}^i(1)]^2}{(\mu_i^c - \delta_{\min}^i(1))[1-(\mu_i^c - \delta_{\min}^i(1))]} = \frac{4[\delta_{\min}^i(1)]^2}{p^2}. \end{aligned} \tag{74}$$

As a result, we have that

$$\sum_{i\notin I_*} \frac{\Delta_{\min}^i}{\mathcal{L}_i^*} \ge \sum_{i\notin I_*} \frac{p^2\Delta_{\min}^i}{4[\delta_{\min}^i(1)]^2} = \sum_{i\notin I_*} \frac{p^2(\varrho_L^*+1)^2}{4\Delta_{\min}^i} \ge \sum_{i\notin I_*} \frac{p^2}{4\Delta_{\min}^i}.$$

G.2 Discussion for SP-BMAB

Discussions We discuss the upper bound in Corollary 6 and lower bound Corollary 7 like those for MP-BMAB by the examples. By setting $\kappa = 2$, the regret of MRCB for SP-BMAB is still $O(\ln B)$. For ease of reference, denote the upper bound and lower bound as $O(\sum_{i\notin I_*} o_i \ln B)$ and $\Omega(\sum_{i\notin I_*} \omega_i \ln B)$ respectively. Similar to the UCB-based policies for conventional MABs (without budget constraints), our MRCB cannot match the lower bound perfectly, i.e., $o_i > \omega_i$.

Like Example 5, the following example shows that o in the upper bound of MRCB and ω in the lower bound share similar trends.

Example 10 *We study the relationship between the regret and the ratio gap Δ^i. Suppose $p \in (0, 0.5)$ and consider a Bernoulli bandit with $\mu_i^r, \mu_i^c \in [p, 1-p]$, $\Delta^i < p/2$, for all $i \in [K]$. In this case, we have that $o_i = 1/(p^2\Delta^i)$ and $\omega_i = p^2/(\Delta^i)$. That is, the coefficients of $\ln B$ in both the upper and lower bounds of the regret are linear to $\sum_{i\notin I_*}(1/\Delta^i)$.*

Proof that MRCB has smaller regret bound than UCB-BV1 The constant in the regret bound of UCB-BV1 [Ding *et al.*, 2013] before $\ln B$ is at least:

$$\varrho_1^* \sum_{i\neq i_*} \Big(\frac{2+\frac{2}{\mu_{\min}^c} + \tilde{\Delta}_i}{\tilde{\Delta}_i \mu_{\min}^c}\Big)^2 + \sum_{i:\mu_i^r < \mu_{i_*}^r} (\mu_{i_*}^r - \mu_i^r)\Big(\frac{2+\frac{2}{\mu_{\min}^c} + \tilde{\Delta}_i}{\tilde{\Delta}_i \mu_{\min}^c}\Big), \tag{75}$$

where $\tilde{\Delta}_i = \varrho_1^* - \frac{\mu_i^r}{\mu_i^c} > 0$ for any $i \neq i_*$.

Both the two terms in (75) are larger than zero. We have the following facts:

1. We can find the proper $\kappa > 1$ s.t. $(2+\frac{2}{\mu_{\min}^c} + \tilde{\Delta}_i)^2 > (1+\kappa)^2(1+\varrho_1^*)^2$, since $\frac{1}{\mu_{\min}^c} > \varrho_1^*$ and $\tilde{\Delta}_i > 0$.
2. $\frac{\varrho_1^*}{(\mu_{\min}^c \tilde{\Delta}_i)^2} > \frac{1}{\Delta^i}$.

Therefore, by choosing proper κ, the regret bound of MRCB outperforms UCB-BV1.

Similar discussions could be applied to the UCB-BV2 in [Ding *et al.*, 2013].

H Supplemental for Experiments

H.1 Parameters for the Distributions

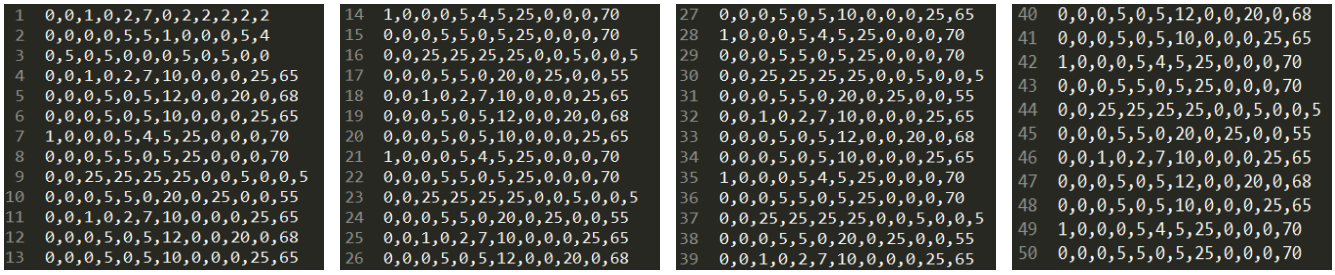
The parameters for 50-armed distributions in Section 6 are shown in Figure 2 and 3. For the 10-armed setting, we use the first 10 arms. We generate random variables using *std::discrete_distribution* (for multinomial distributions) and *std::_Beta_distribution* (for beta distributions). Each line in Figure 2 and 3 represent the parameters for an arm.

For multinomial distributions, each line in Figure 2 is of the form ArmID $(\alpha_0, \alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5)$, which represents that:

$$\mathbb{P}\{r_{\text{ArmID}}(t) = 0.2j\} = \frac{\alpha_j}{\sum_{k=0}^5 \alpha_k}; \quad \mathbb{P}\{c_{\text{ArmID}}(t) = 0.2j\} = \frac{\beta_j}{\sum_{k=0}^5 \beta_k}, \quad \forall j \in \{0,1,2,3,4,5\}. \tag{76}$$

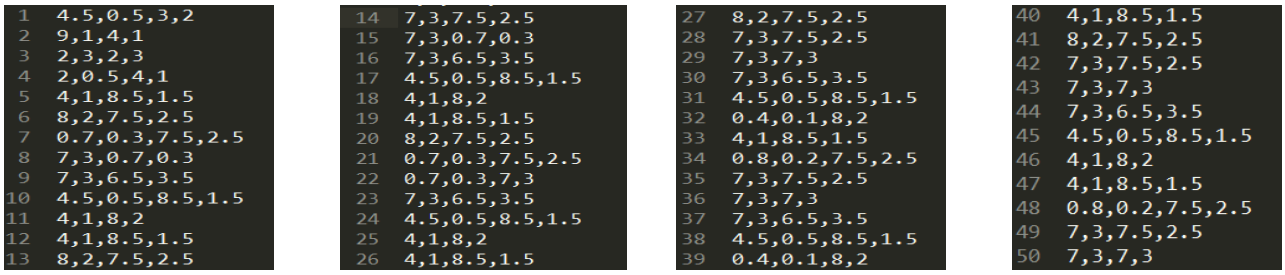
For beta distributions, each line in Figure 2 is of the form ArmID $(\alpha_r, \beta_r, \alpha_c, \beta_c)$, which represents that the reward distribution of ArmID is Beta(α_r,β_r) and the cost distribution of ArmID is Beta(α_c,β_c).

H.2 Additional Experiments



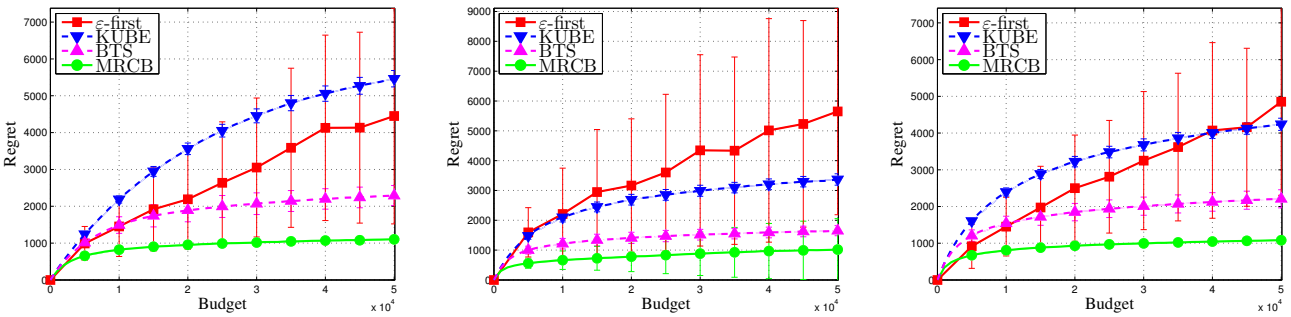
(a) Arm 1 to Arm 13 (b) Arm 14 to Arm 26 (c) Arm 27 to Arm 39 (d) Arm 40 to Arm 50

Figure 2: Parameters for the Multinomial Distribution

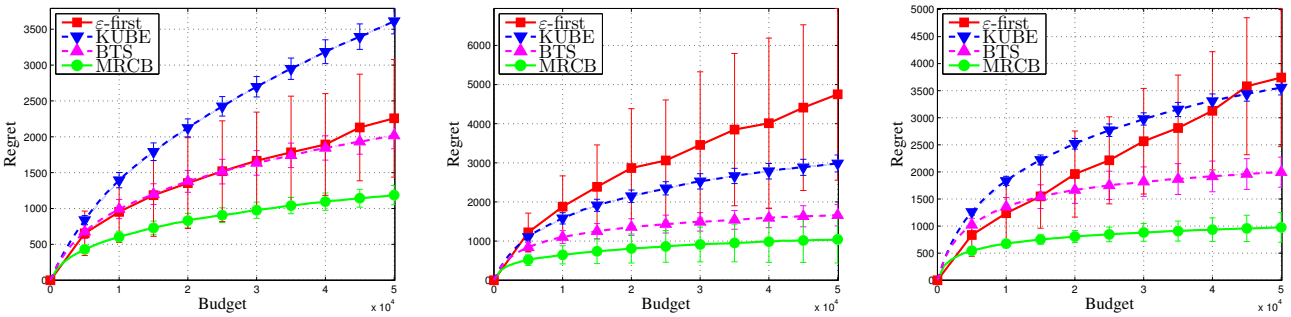


(a) Arm 1 to Arm 13 (b) Arm 14 to Arm 26 (c) Arm 27 to Arm 39 (d) Arm 40 to Arm 50

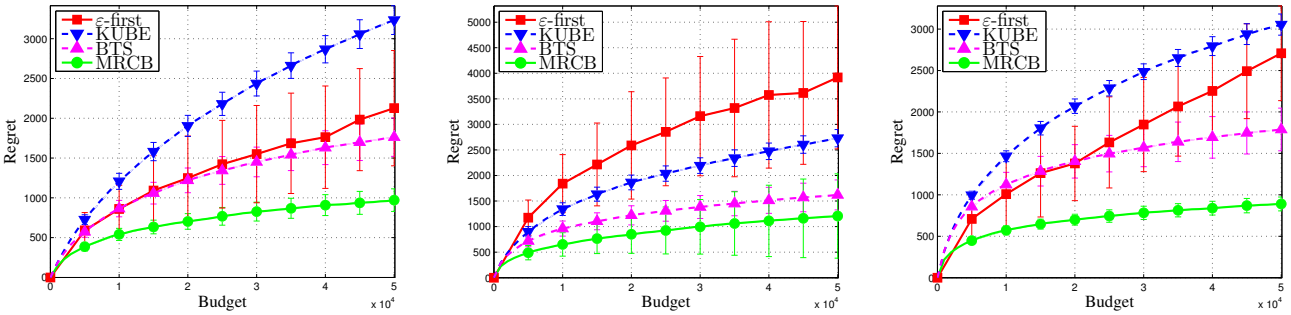
Figure 3: Parameters for the Beta Distribution



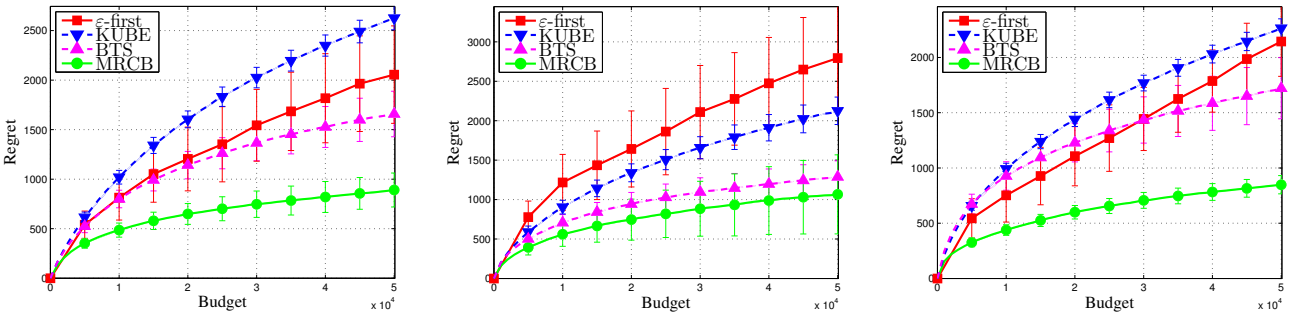
(a) Multinomial, $K = 50, L = 1$ (b) Bernoulli, $K = 50, L = 1$ (c) Beta, $K = 50, L = 1$



(d) Multinomial, $K = 50, L = 3$ (e) Bernoulli, $K = 50, L = 3$ (f) Beta, $K = 50, L = 3$



(g) Multinomial, $K = 50, L = 5$ (h) Bernoulli, $K = 50, L = 5$ (i) Beta, $K = 50, L = 5$



(j) Multinomial, $K = 50, L = 10$ (k) Bernoulli, $K = 50, L = 10$ (l) Beta, $K = 50, L = 10$

Figure 4: We carry out additional experiments. The parameters are randomly generated. Each experiment is individually repeated for 100 times. The average regrets and standard derivations are reported. We can see that our MRCB performs the best.