Budgeted Multi-Armed Bandits with Multiple Plays*

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.

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

Let I_* denote the super arm defined as follows:

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

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

[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.

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.

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 $\mu_i^r$
    and expected cost $\mu_i^c$; 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[2] 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.

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

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 $\mathcal{O}(a,b,L)$

```
Input: Vectors $a$ and $b$ with $a_i > 0, b_i \geq 0\ \forall i \in [K]$;
  $L \in [K]$;
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 $\mathrm{argmax}_i(a_i/b_i)$ for any $i \in [K]$
  directly; Else, go to the next step;
Solve the LP problem marked with $(\triangle)$ by Interior Point
  Method. Denote the solution as $y^*$ and $z^*$.
  $\max\ a^T y$ s.t. $\sum_{i=1}^K y_i - Lz = 0$; $b^T y = 1$;        $(\triangle)$
                 $z \geq 0$; $0 \leq y_i \leq z\ \forall i \in [K]$;
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*[3].

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[4] 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[5] with the i-th element $\tilde{\mu}_i^r(t)$, and so for $\tilde{\mu}^c(t)$.

Algorithm 3: Multiple Ratio Confidence Bound (MRCB)

```
Input: hyper parameter $\kappa > 0$, the budget $B$; $L \in [K]$;
for $t \to 1 : \lceil K/L \rceil$ do
    Pull arms $\{(([(t-1)L + j - 1] \bmod K) + 1 | j \in [L]\}$;
for $t \to \lceil K/L \rceil + 1 : \infty$ do
    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 $i$;
    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

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

[2] All the appendices are included in the online full version of this work, which is at http://goo.gl/ewyX9c.

[3] 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".

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

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

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

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 $\varphi_\iota = (\varrho^*_L + 1)^2 L^2 (\sqrt{\kappa} + 1)^2 \sum_{i\in\mathcal{B}} (2/\Delta^i_{\min} - 1/\Delta^i_{\max})$, $\varphi_s = 2L \sum_{i\in\mathcal{B}} \Delta^i_{\max}$, *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 $(\sum_{k\in I} \mu^c_k)[\varrho^*_L - (\sum_{k\in I} \mu^r_k)/(\sum_{k\in I} \mu^c_k)]$, $\mathcal{N}_I$ 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 $\sum_{k\in I} \mu^c_k$; 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).

$$\text{(a)}\, \delta^{i,j}(\gamma) = \frac{\Delta^{i,j}}{\gamma\varrho^*_L + 1}; \text{(b)}\, \varrho^*_L = \frac{(\sum_{k\in S(i,j)} \mu^r_k) + \delta^{i,j}(\gamma)}{(\sum_{k\in S(i,j)} \mu^c_k) - \gamma\delta^{i,j}(\gamma)}. \tag{11}$$

In the analysis of the upper bound of the regret, we only need to consider the case of [6] $\gamma = 1$. 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:

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

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

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

[6] 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. □

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 $\mathit{\Gamma}_k^w(T)$ denote the pulling number of arm $k \in [K]$ in the first T rounds, and $\mathit{\Gamma}_I^w(T)$ for super arm $I \in \mathcal{C}_L^K$, where $T \in \mathbb{Z}_+$. If $\sum_{I\in\mathcal{C}_s} \mathbb{E}\{\mathit{\Gamma}_I^w(T)\} = o(T^a)$ 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:

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

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., $\sum_{I\in\mathcal{C}_s} \mathbb{E}\{\mathit{\Gamma}_I^w(T)\} = o(T^a)$ holds for any $a \in (0,1)$ and $T \in \mathbb{Z}_+$), we have that for any $i \notin I_*$ and $\epsilon > 0$,*

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

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

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:

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

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., $\sum_{i\neq i_*} \mathbb{E}\{\mathit{\Gamma}_i^w(T)\} = o(T^a)$ holds for any $a \in (0,1)$ and $T \in \mathbb{Z}_+$), we have that for any $i \neq i_*$ and $\epsilon > 0$,*

$\lim_{B\to\infty} \mathbb{P}\{\mathit{\Gamma}_i^w(B) \geq [(1-\epsilon)/\mathcal{L}_i^{**}]\ln B\} = 1$; *consequently,* $\liminf_{B\to\infty} \mathbb{E}[\mathit{\Gamma}_i^w(B)]/[\ln B] \geq 1/\mathcal{L}_i^{**}$.

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.

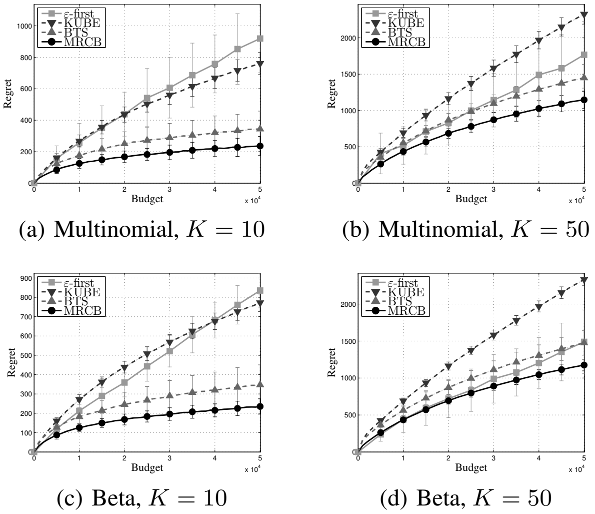
MRCB has a hyper parameter κ. We searched the κ in the set $\{2^{-10}, 2^{-7}, 2^{-4}, 2^{1}\}$ and found that $\kappa = 2^{-4}$ worked well for most cases. Therefore, we fix 2^{-4} in the following experiments. Though asymptotically MRCB enjoys log-linear regret when $\kappa > 1$, it is not good to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

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	- - -

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.



(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 *FOCS*. IEEE, 2013.

[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.

[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.

[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.

[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 *KDD*. 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.