Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015)

# Thompson Sampling for Budgeted Multi-armed Bandits

**Yingce Xia**[1,*], **Haifang Li**[2,*], **Tao Qin**[3], **Nenghai Yu**[1] and **Tie-Yan Liu**[3]
[1]University of Science and Technology of China, Hefei, China
[2] University of Chinese Academy of Sciences, Beijing, China [3]Microsoft Research, Beijing, China
yingce.xia@gmail.com, lihaifang@amss.ac.cn, {taoqin, tyliu}@microsoft.com, ynh@ustc.edu.cn

## Abstract

Thompson sampling is one of the earliest randomized algorithms for multi-armed bandits (MAB). In this paper, we extend the Thompson sampling to Budgeted MAB, where there is random cost for pulling an arm and the total cost is constrained by a budget. We start with the case of Bernoulli bandits, in which the random rewards (costs) of an arm are independently sampled from a Bernoulli distribution. To implement the Thompson sampling algorithm in this case, at each round, we sample two numbers from the posterior distributions of the reward and cost for each arm, obtain their ratio, select the arm with the maximum ratio, and then update the posterior distributions. We prove that the distribution-dependent regret bound of this algorithm is $O(\ln B)$, where $B$ denotes the budget. By introducing a Bernoulli trial, we further extend this algorithm to the setting that the rewards (costs) are drawn from general distributions, and prove that its regret bound remains almost the same. Our simulation results demonstrate the effectiveness of the proposed algorithm.

## 1 Introduction

The multi-armed bandit (MAB) problem, a classical sequential decision problem in an uncertain environment, has been widely studied in the literature [Lai and Robbins, 1985; Auer *et al.*, 2002]. Many real world applications can be modeled as MAB problems, such as news recommendation [Li *et al.*, 2010] and channel allocation [Gai *et al.*, 2010]. Previous studies on MAB can be classified into two categories: one focuses on designing algorithms to find a policy that can maximize the cumulative expected reward, such as UCB1 [Auer *et al.*, 2002], UCB-V [Audibert *et al.*, 2009], MOSS [yves Audibert and Bubeck, 2009], KL-UCB [Garivier and Cappé, 2011] and Bayes-UCB [Kaufmann *et al.*, 2012a]; the other aims at studying the sample complexity to reach a specific accuracy, such as [Bubeck *et al.*, 2009; Yu and Nikolova, 2013].

Recently, a new setting of MAB, called *budgeted MAB*, was proposed to model some new Internet applications, including online bidding optimization in sponsored search [Amin *et al.*, 2012; Tran-Thanh *et al.*, 2014] and on-spot instance bidding in cloud computing [Agmon Ben-Yehuda *et al.*, 2013; Ardagna *et al.*, 2011]. In budgeted MAB, pulling an arm receives both a random reward and a random cost, drawn from some unknown distributions. The player can keep pulling the arms until he/she runs out of budget $B$. A few algorithms have been proposed to solve the budgeted MAB problem. For example, in [Tran-Thanh *et al.*, 2010], an $\epsilon$-first algorithm was proposed which first spends $\epsilon B$ budget on pure explorations, and then keeps pulling the arm with the maximum empirical reward-to-cost ratio. It was proven that the $\epsilon$-first algorithm has a regret bound of $O(B^{\frac{2}{3}})$. KUBE [Tran-Thanh *et al.*, 2012] is another algorithm for budgeted MAB, which solves an integer linear program at each round, and then converts the solution to the probability of each arm to be pulled at the next round. A limitation of the $\epsilon$-first and KUBE algorithms lies in that they assume the cost of each arm to be deterministic and fixed, which narrows their application scopes. In [Ding *et al.*, 2013], the setting was considered that the cost of each arm is drawn from an unknown discrete distribution and two algorithms UCB-BV1/BV2 were designed. A limitation of these algorithms is that they require additional information about the minimum expected cost of all the arms, which is not available in some applications.

Thompson sampling [Thompson, 1933] is one of the earliest randomized algorithms for MAB, whose main idea is to choose an arm according to its posterior probability to be the best arm. In recent years, quite a lot of studies have been conducted on Thompson sampling, and good performances have been achieved in practical applications [Chapelle and Li, ]. It is proved in [Kaufmann *et al.*, 2012b] that Thompson sampling can reach the lower bound of regret given in [Lai and Robbins, 1985] for Bernoulli bandits. Furthermore, problem-independent regret bounds were derived in [Agrawal and Goyal, 2013] for Thompson sampling with Beta and Gaussian priors.

Inspired by the success of Thompson sampling in classical MAB, two natural questions arise regarding its extension to budgeted MAB problems: (i) How can we adjust Thompson sampling so as to handle budgeted MAB problems? (ii) What is the performance of Thompson sampling in theory and in

[*]This work was done when the first two authors were interns at Microsoft Research.

practice? In this paper, we try to provide answers to these two questions.

*Algorithm*: We propose a refined Thompson sampling algorithm that can be used to solve the budgeted MAB problems. While the optimal policy for budgeted MAB could be very complex (budgeted MAB can be viewed as a stochastic version of the knapsack problem in which the value and weight of the items are both stochastic), we prove that, when the reward and cost per pulling are supported in $[0, 1]$ and the budget is large, we can achieve the almost optimal reward by always pulling the optimal arm (associated with the maximum expected-reward-to-expected-cost ratio). With this guarantee, our proposed algorithm targets at pulling the optimal arm as frequently as possible. We start with Bernoulli bandits, in which the random rewards (costs) of an arm are independently sampled from a Bernoulli distribution. We design an algorithm which (1) uses beta distribution to model the priors of the expected reward and cost of each arm, and (2) at each round, samples two numbers from the posterior distributions of the reward and cost for each the arm, obtains their ratio, selects the arm with the maximum ratio, and then updates the posterior distributions. We further extend this algorithm to the setting that the rewards (costs) are drawn from general distributions by introducing Bernoulli trials.

*Theoretical analysis*: We prove that our proposed algorithm can achieve a distribution-dependent regret bound of $O(\ln B)$, with a tighter constant before $\ln B$ than existing algorithms (e.g., the two algorithms in [Ding *et al.*, 2013]). To obtain this regret bound, we first show that it suffices to bound the expected pulling times of all the suboptimal arms (whose expected-reward-to-expected-cost ratios are not maximum). To this end, for each suboptimal arm, we define two gaps, the $\delta$-ratio gap and the $\epsilon$-ratio gap, which compare its expected-reward-to-expected-cost ratio to that of the optimal arm. Then by introducing some intermediate events, we can decompose the expected pulling time of a suboptimal arm $i$ into several terms, each of which depends on only the reward or only the cost. After that, we can bound each term by the concentration inequalities and two gaps with careful derivations.

To our knowledge, it is the first time that Thompson sampling is applied to the budgeted MAB problem. We conduct a set of numerical simulations with different rewards/costs distributions and different number of arms. The simulation results demonstrate that our proposed algorithm is much better than several baseline algorithms.

## 2 Problem Formulation

In this section, we give a formal definition to the budgeted MAB problem.

In budgeted MAB, we consider a slot machine with $K$ arms ($K \geq 2$). [1] At round $t$, a player pulls an arm $i \in [K]$, receives a random reward $r_i(t)$, and pays a random cost $c_i(t)$ until he runs out of his budget $B$, which is a positive integer. Both the reward $r_i(t)$ and the cost $c_i(t)$ are supported on $[0, 1]$. For simplicity and following the practice in previous works, we make a few assumptions on the rewards and costs: (i) the rewards of an arm are independent of its costs; (ii) the rewards and costs of an arm are independent of other arms; (iii) the rewards and costs of the same arm at different rounds are independent and identically distributed.

We denote the expected reward and cost of arm $i$ as $\mu_i^r$ and $\mu_i^c$ respectively. W.l.o.g., we assume $\forall i \in [K]$, $\mu_i^r > 0$, $\mu_i^c > 0$, and $\arg\max_{i\in[K]} \frac{\mu_i^r}{\mu_i^c} = 1$. We name arm 1 as the optimal arm and the other arms as suboptimal arms.

Our goal is to design algorithms/policies for budgeted MAB with small pseudo-regret, which is defined as follows:

$$\text{Regret} = R^* - \mathbb{E}\sum_{t=1}^{T_B} r_t, \tag{1}$$

where $R^*$ is the expected reward of the optimal policy (the policy that can obtain the maximum expected reward given the reward and cost distributions of each arm), $r_t$ is the reward received by an algorithm at round $t$, $T_B$ is the stopping time of the algorithm, and the expectation is taken w.r.t. the randomness of the algorithm, the rewards (costs), and the stopping time.

Please note that it could be very complex to obtain the optimal policy for the budgeted MAB problem (under the condition that the reward and cost distributions of each arm are known). Even for its degenerated case, where the reward and cost of each arm are deterministic, the problem is known to be NP-hard (actually in this case the problem becomes an unbounded knapsack problem [Martello and Toth, 1990]). Therefore, generally speaking, it is hard to calculate $R^*$ in an exact manner.

However, we find that it is much easier to approximate the optimal policy and to upper bound $R^*$. Specifically, when the reward and cost per pulling are supported in $[0, 1]$ and $B$ is large, always pulling the optimal arm could be very close to the optimal policy. For Bernoulli bandits, since there is no time restrictions on pulling arms, one should try to always pull arm 1 so as to fully utilize the budget[2]. For the general bandits, the situation is a little more complicated and always pulling arm 1 will result in a suboptimiality of at most $2\frac{\mu_1^r}{\mu_1^c}$. These results are summarized in Lemma 1, together with upper bounds on $R^*$.

**Lemma 1** *When the reward and cost per pulling are supported in* $[0, 1]$*, for Bernoulli bandits, we have* $R^* = \frac{\mu_1^r}{\mu_1^c}B$ *and the optimal policy is exactly always pulling arm* 1*; for general bandits, we have* $R^* \leq \frac{\mu_1^r}{\mu_1^c}(B + 1)$*, and the suboptimality of always pulling arm* 1 *(as compared to the optimal policy) is at most* $2\frac{\mu_1^r}{\mu_1^c}$.

For any $i \geq 2$, define $T_i$ as the pulling time of arm $i$ when running out of budget. Denote the difference of the expected-reward-to-expected-cost ratio between the optimal arm 1 and

[1] Denote the set $\{1, 2, \cdots, K\}$ as $[K]$.

[2] This is inspired by the greedy heuristic for the knapsack problem [Fisher, 1980], i.e., at each round, one selects the item with the maximum value-to-weight ratio. Although there are many approximation algorithms for the knapsack problem like the total-value greedy heuristic [Kohli and Krishnamurti, 1992] and the FPTAS [Vazirani, 2001], under our budgeted MAB setting, we find that they will not bring much benefit on tightening the bound of $R^*$.

a suboptimal arm $i(\geq 2)$ as $\Delta_i$:

$$\Delta_i = \frac{\mu_1^r}{\mu_1^c} - \frac{\mu_i^r}{\mu_i^c}, \quad \forall i \geq 2. \tag{2}$$

Lemma 2 relates the regret to $T_i$ and $\Delta_i$ ($i \geq 2$). It is useful when we analyze the regret of a pulling algorithm.

**Lemma 2** *For Bernoulli bandits, we have*

$$\textit{Regret} = \sum_{i=2}^{K} \mu_i^c \Delta_i \mathbb{E}\{T_i\}. \tag{3}$$

*For general bandits, we have*

$$\textit{Regret} \leq 2\frac{\mu_1^r}{\mu_1^c} + \sum_{i=2}^{K} \mu_i^c \Delta_i \mathbb{E}\{T_i\}. \tag{4}$$

The intuition behind Lemma 2 is as follows. As aforementioned, for Bernoulli bandits, the optimal policy is to always pull arm 1. If one pulls a suboptimal arm $i$ ($> 1$) for $T_i$ times, then he/she will lose some rewards. Specifically, the expected budget spent on arm $i$ is $\mu_i^c T_i$, and if he/she spent such budget on the optimal arm 1, he/she can get $\mu_i^c \Delta_i T_i$ extra reward. For general bandits, always pulling arm 1 might not be optimal (see Lemma 1) – actually it leads to a regret at most $\frac{2\mu_1^r}{\mu_1^c}$. Therefore, we need to add an extra term $\frac{2\mu_1^r}{\mu_1^c}$ to the result for Bernoulli bandits.

## 3 Budgeted Thompson Sampling

In this section, we first show how Thompson sampling can be extended to handle budgeted MAB with Bernoulli distributions, and then generalize the setting to general distributions. For ease of reference, we call the corresponding algorithm *Budgeted Thompson Sampling (BTS)*.

First, the BTS algorithm for the budgeted Bernoulli bandits is shown in Algorithm 1. In the algorithm, $S_i^r(t)$ denotes the times that the player receives reward 1 from arm $i$ before (excluding) round $t$, $S_i^c(t)$ denotes the times that the player pays cost 1 for pulling arm $i$ before (excluding) round $t$, and $Beta(\cdot, \cdot)$ denotes the beta distribution. Please note that we use the beta distribution as a prior in Algorithm 1 because it is the conjugate distribution of the binomial distribution: If the prior is a $Beta(\alpha, \beta)$, after a Bernoulli experiment, the posterior distribution is either $Beta(\alpha+1, \beta)$ (if the trial is a success) or $Beta(\alpha, \beta+1)$ (if the trial is a failure).

In the original Thompson sampling algorithm, one draws a sample from the posterior beta distribution for the reward of each arm, pulls the arm with the maximum sampled reward, receives a reward, and then updates the reward distribution based on the received reward. In Algorithm 1, in addition to sampling rewards, we also sample costs for the arms at the same time, pull the arm with the maximum sampled reward-to-cost ratio, receive both the reward and cost, and then update the reward distribution and cost distribution.

As compared to KUBE [Tran-Thanh *et al.*, 2012], Algorithm 1 does not need to solve a complex integer linear program. As compared to the UCB-style algorithms like fractional KUBE [Tran-Thanh *et al.*, 2012] and UCB-BV1 [Ding *et al.*, 2013], Algorithm 1 does not need carefully designed confidence bounds. As can be seen, BTS only simply chooses one out of the $K$ arms according to their posterior probabilities to be the best arm, which is an intuitive, easy-to-implement, and efficient approach.

**Algorithm 1** Budgeted Thompson Sampling (BTS)

1: For each arm $i \in [K]$, set $S_i^r(1) \leftarrow 0$, $F_i^r(1) \leftarrow 0$, $S_i^c(1) \leftarrow 0$, and $F_i^c(1) \leftarrow 0$ ;
2: Set $B_1 \leftarrow B$; $t \leftarrow 1$;
3: **while** $B_t > 0$ **do**
4: For each arm $i \in [K]$, sample $\theta_i^r(t)$ from $Beta(S_i^r(t)+1, F_i^r(t)+1)$, and sample $\theta_i^c(t)$ from $Beta(S_i^c(t)+1, F_i^c(t)+1)$;
5: Pull arm $I_t = \arg\max_{i \in [K]} \frac{\theta_i^r(t)}{\theta_i^c(t)}$; receive reward $r_t$; pay cost $c_t$; update $B_{t+1} \leftarrow B_t - c_t$;
6: *For Bernoulli bandits*, $\tilde{r} \leftarrow r_t, \tilde{c} \leftarrow c_t$; *for general bandits*, sample $\tilde{r}$ from $\mathcal{B}(r_t)$ and sample $\tilde{c}$ from $\mathcal{B}(c_t)$;
7: $S_{I_t}^r(t+1) \leftarrow S_{I_t}^r(t) + \tilde{r}$; $F_{I_t}^r(t+1) \leftarrow F_{I_t}^r(t) + 1 - \tilde{r}$;
8: $S_{I_t}^c(t+1) \leftarrow S_{I_t}^c(t) + \tilde{c}$; $F_{I_t}^c(t+1) \leftarrow F_{I_t}^c(t) + 1 - \tilde{c}$;
9: $\forall j \neq I_t$, $S_j^r(t+1) \leftarrow S_j^r(t)$, $F_j^r(t+1) \leftarrow F_j^r(t)$, $S_j^c(t+1) \leftarrow S_j^c(t)$, $F_j^c(t+1) \leftarrow F_j^c(t)$;
10: Set $t \leftarrow t+1$.
11: **end while**

By leveraging the idea proposed in [Agrawal and Goyal, 2012], we can modify the BTS algorithm for Bernoulli bandits and make it work for bandits with general reward/cost distributions. In particular, with general distributions, the reward $r_t$ and cost $c_t$ (in Step 5) at round $t$ become real numbers in $[0, 1]$. We introduce a Bernoulli trial in Step 6: Set $\tilde{r} \leftarrow \mathcal{B}(r_t)$ and $\tilde{c} \leftarrow \mathcal{B}(c_t)$, in which $\mathcal{B}(r_t)$ is a Bernoulli test with success probability $r_t$ and so is $\mathcal{B}(c_t)$. Now $S_i^r(t)$ and $S_i^c(t)$ represent the number of success Bernoulli trials for the reward and cost respectively. Then we can use $\tilde{r}$ and $\tilde{c}$ to update $S_i^r(t)$ and $S_i^c(t)$ accordingly.

## 4 Regret Analysis

In this section, we analyze the regret of our proposed BTS algorithm. We start with Bernoulli bandits and then generalize the results to general bandits. Due to space restrictions, we only give proof sketches here. Although the proof sketches are self-contained, interested readers can find more details (as well as colored figures for experiments) in the online full version of this paper [Xia *et al.*, 2015].

In the classical MAB, the player only needs to explore the expected reward of each arm, however, in the budgeted MAB the player also needs to explore the expected cost simultaneously. Therefore, as compared with [Agrawal and Goyal, 2012], our regret analysis will heavily depends on some quantities related to the reward-to-cost ratio (such as the two gaps defined below).

For an arm $i(\geq 2)$ and a given $\gamma \in (0, 1)$,we define

$$\delta_i(\gamma) = \frac{\gamma \mu_i^c \Delta_i}{\frac{\mu_1^r}{\mu_1^c} + 1}, \quad \epsilon_i(\gamma) = \frac{(1-\gamma)\mu_1^c \Delta_i}{\frac{\mu_i^r}{\mu_i^c} + 1}.$$

It is easy to verify the following equation for any $i \geq 2$.

$$\frac{\mu_i^r + \delta_i(\gamma)}{\mu_i^c - \delta_i(\gamma)} = \frac{\mu_1^r - \epsilon_i(\gamma)}{\mu_1^c + \epsilon_i(\gamma)}$$

For ease of reference, $\forall i \geq 2$, we call $\delta_i(\gamma)$ the *$\delta$-ratio gap* between the optimal arm and a suboptimal arm $i$, and $\epsilon_i(\gamma)$ the *$\epsilon$-ratio gap*. In the remaining part of this section, we simply write $\epsilon_i(\gamma)$ as $\epsilon_i$ when the context is clear and there is no confusion.

The following theorem says that BTS achieves a regret bound of $O(\ln(B))$ for both Bernoulli and general bandits:

**Theorem 3** *$\forall \gamma \in (0,1)$, for both Bernoulli bandits and general bandits, the regret of the BTS algorithm can be upper bounded as below.*

$$\textit{Regret} \leq \sum_{i=2}^{K} \Big\{ \frac{2\ln B}{\gamma^2 \mu_i^c \Delta_i} \Big( \frac{\mu_1^r}{\mu_1^c} + 1 \Big)^2 + \Phi_i(\gamma) \Big\} + O\Big( \frac{K}{\gamma^2} \Big),$$

*in which $\Delta_i$ is defined in Eqn. (2) and $\Phi_i(\gamma)$ is defined as*

$$\begin{cases} O\Big( \frac{1}{\epsilon_i^4(\gamma)} \Big), & \text{if } \mu_1^c + \epsilon_i(\gamma) \geq 1; \\ O\Big( \frac{1}{\epsilon_i^6(\gamma)(1 - \mu_1^c - \epsilon_i(\gamma))} \Big), & \text{if } \mu_1^c + \epsilon_i(\gamma) < 1. \end{cases} \tag{5}$$

We first prove Theorem 3 holds for Bernoulli bandits in Section 4.1 and then extend the result for general bandits in Section 4.2 .

## 4.1 Analysis for Bernoulli Bandits

First, we describe the high-level idea of how to prove the theorem. According to Lemma 2, to upper bound the regret of BTS, it suffices to bound $\mathbb{E}\{T_i\}$ $\forall i \geq 2$. For a suboptimal arm $i$, $\mathbb{E}\{T_i\}$ can be decomposed into the sum of a constant and the probabilities of two kinds of events (see (6)). The first kind of event is related to the $\delta$-ratio gap $\delta_i(\gamma)$, and its probability can be bounded by leveraging concentrating inequalities and the relationship between the binomial distribution and the beta distribution. The second one is related to the $\epsilon$-ratio gap $\epsilon_i(\gamma)$, according to which the probability of the event related to arm $i$ can be converted to that related to the optimal arm 1. To bound the probability of the second kind of event, we need some complicated derivations, as shown in the later part of this subsection.

Then, we define some notations and intermediate variables, which will be used in the proof sketch.

$n_{i,t}$ denotes the pulling time of arm $i$ before (excluding) round $t$; $I_t$ denotes the arm pulled at round $t$; $\mathbf{1}\{\cdot\}$ is the indicator function; $\mu_{\min}^c = \min_{i \in [K]} \{\mu_i^c\}$; $H_{t-1}$ denotes the history until round $t-1$, including the arm pulled from round 1 to $t-1$, and the rewards/costs received at each round; $\theta_i(t)$ denotes the ratio $\frac{\theta_i^r(t)}{\theta_i^c(t)}$ $\forall i \in [K]$ where $\theta_i^r(t)$ and $\theta_i^c(t)$ are defined in Step 4 of Algorithm 1; $B_t$ denotes the budget left at the beginning of round $t$; $E_i^\theta(t)$ denotes the event that given $\gamma \in (0,1)$, $\theta_i(t) \leq \frac{\mu_i^r + \delta_i(\gamma)}{\mu_i^c - \delta_i(\gamma)}$ $\forall i > 1$; the probability $p_{i,t}$ denotes $\mathbb{P}\{\theta_1(t) > \frac{\mu_1^r - \epsilon_i(\gamma)}{\mu_1^c + \epsilon_i(\gamma)} | H_{t-1}, B_t > 0\}$ given $\gamma \in (0,1)$ $\forall i > 1$; $\overline{event}$ denotes the "$event$" does not hold.

After that, we give the proof sketch as follows, which can be partitioned into four steps.

***Step 1**: Decompose $\mathbb{E}\{T_i\}$ ($i > 1$).*

It can be shown that the pulling time of a suboptimal arm $i$ can be decomposed into three parts: a constant invariant to $t$ and the probabilities of two kinds of events:

$$\mathbb{E}\{T_i\} \leq \lceil L_i \rceil + \sum_{t=1}^{\infty} \mathbb{P}\{\overline{E_i^\theta(t)}, n_{i,t} \geq \lceil L_i \rceil, B_t > 0\} + \sum_{t=1}^{\infty} \mathbb{P}\{I_t = i, E_i^\theta(t), B_t > 0\}, \tag{6}$$

where $L_i = \frac{2\ln B}{\delta_i^2(\gamma)}$. Note that $L_i$ depends on $\gamma$. We omit the $\gamma$ when there is no confusion throughout the context. We then bound the probabilities of the two kinds of events in the next two steps.

***Step 2:** Bound $\sum_{t=1}^{\infty} \mathbb{P}\{\overline{E_i^\theta(t)}, n_{i,t} \geq \lceil L_i \rceil, B_t > 0\}$.*

Define two new events: $\forall i \geq 2$ and $t \geq 1$,

(I) $E_i^r(t) : \theta_i^r(t) \leq \mu_i^r + \delta_i(\gamma)$; (II) $E_i^c(t) : \theta_i^c(t) \geq \mu_i^c - \delta_i(\gamma)$.

If $\overline{E_i^\theta(t)}$ holds, at least one event of $\overline{E_i^r(t)}$ and $\overline{E_i^c(t)}$ holds. Therefore, we have

$$\mathbb{P}\{\overline{E_i^\theta(t)}, n_{i,t} \geq \lceil L_i \rceil | B_t > 0\} \leq \mathbb{P}\{\overline{E_i^r(t)}, n_{i,t} \geq \lceil L_i \rceil | B_t > 0\} + \mathbb{P}\{\overline{E_i^c(t)}, n_{i,t} \geq \lceil L_i \rceil | B_t > 0\}. \tag{7}$$

Intuitively, when $n_{i,t}$ is large enough, $\theta_i^r(t)$ and $\theta_i^c(t)$ should be very close to $\mu_i^r$ and $\mu_i^c$ respectively. Then, both $\overline{E_i^r(t)}$ and $\overline{E_i^c(t)}$ will be low-probability events. Mathematically, $\forall \gamma \in (0,1)$, the two terms in the right-hand side of (7) could be bounded as follows, by considering the relationship between the binomial distribution and the beta distribution.

$$\mathbb{P}\{\overline{E_i^r(t)}, n_{i,t} \geq \lceil L_i \rceil | B_t > 0\} \leq \frac{7}{B\delta_i^2(\gamma)}. \tag{8}$$

$$\mathbb{P}\{\overline{E_i^c(t)}, n_{i,t} \geq \lceil L_i \rceil | B_t > 0\} \leq \frac{28}{B\delta_i^2(\gamma)}. \tag{9}$$

As a result, we have

$$\mathbb{P}\{\overline{E_i^\theta(t)}, n_{i,t} \geq \lceil L_i \rceil | B_t > 0\} \leq \frac{35}{B\delta_i^2(\gamma)}.$$

One can also verify that $\sum_{t=1}^{\infty} \mathbb{P}\{B_t > 0\}$ is bounded by

$$\frac{1}{\mu_{\min}^c} \sum_{t=1}^{\infty} \sum_{i=1}^{K} \mathbb{E}\{c_i(t) \mathbf{1}\{I_t = i\} | B_t > 0\} \mathbb{P}\{B_t > 0\} \leq \frac{B}{\mu_{\min}^c}, \tag{10}$$

where $c_i(t)$ is the cost of arm $i$ at round $t$.

Therefore, we obtain that

$$\sum_{t=1}^{\infty} \mathbb{P}\{\overline{E_i^\theta(t)}, n_{i,t} \geq \lceil L_i \rceil, B_t > 0\} \leq \frac{35}{\delta_i^2(\gamma)\mu_{\min}^c}. \tag{11}$$

***Step 3**: Bound $\sum_{t=1}^{\infty} \mathbb{P}\{I_t = i, E_i^\theta(t), B_t > 0\}$.*

Let $\tau_k$ ($k \geq 0$) denote the round that arm 1 has been pulled for the $k$-th time and define $\tau_0 = 0$. $\forall i \geq 2$ and $\forall t \geq 1$, $p_{i,t}$ is only related to the pulling history of arm 1, thus $p_{i,t}$ will not change between $\tau_k + 1$ and $\tau_{k+1}$, $\forall k \geq 0$. With some derivations, we can get that

$$\sum_{t=1}^{\infty} \mathbb{P}\{I_t = i, E_i^\theta(t), B_t > 0\} \leq \sum_{k=0}^{\infty} \Big( \mathbb{E}\Big\{ \frac{1}{p_{i,\tau_k+1}} \Big\} - 1 \Big). \tag{12}$$

(12) bridges the probability of an event related to arm 1 and that related to arm $i$ ($i \geq 2$). To further decompose the r.h.s. of (12), define the following two probabilities which are related to the $\epsilon$-ratio gap between arm 1 and arm $i$:

$$p^r_{i,t} = \mathbb{P}\{\theta^r_1(t) \geq \mu^r_1 - \epsilon_i(\gamma)|H_{t-1}\},$$
$$p^c_{i,t} = \mathbb{P}\{\theta^c_1(t) \leq \mu^c_1 + \epsilon_i(\gamma)|H_{t-1}\}.$$

Since the reward of an arm is independent of its cost, we can verify $p_{i,t} \geq p^r_{i,t}p^c_{i,t}$ and then get

$$\mathbb{E}\Big\{\frac{1}{p_{i,\tau_k+1}}\Big\} \leq \mathbb{E}\Big\{\frac{1}{p^r_{i,\tau_k+1}}\Big\}\mathbb{E}\Big\{\frac{1}{p^c_{i,\tau_k+1}}\Big\}. \quad (13)$$

According to (12) and (13), $\sum_{t=1}^{\infty}\mathbb{P}\{I_t = i, E^\theta_i(t), B_t > 0\}$ can be bounded by the sum of the right-hand side of (13) over index $k$ from 0 to infinity, which is related to the pulling time of arm 1 and the $\epsilon$-ratio gap between arm 1 and arm $i$.

It is quite intuitive that when arm 1 is played for enough times, $\theta^r_1(t)$ and $\theta^c_1(t)$ will be very close to $\mu^r_1$ and $\mu^c_1$ respectively. That is, probabilities $p^r_{i,\tau_k+1}$ and $p^c_{i,\tau_k+1}$ will be close to 1, and so will their reciprocals. To mathematically characterize $p^r_{i,\tau_k+1}$ and $p^c_{i,\tau_k+1}$, we define some notations as follows, which are directly or indirectly related to the $\epsilon$-ratio gap: $y_i = \mu^r_1 - \epsilon_i$, $z_i = \mu^c_1 + \epsilon_i$, $R_{1,i} = \frac{\mu^r_1(1-y_i)}{y_i(1-\mu^r_1)}$, $R_{2,i} = \frac{\mu^c_1(1-z_i)}{z_i(1-\mu^c_1)}$, $D_{1,i} = y_i\ln(\frac{y_i}{\mu^r_1}) + (1-y_i)\ln(\frac{1-y_i}{1-\mu^r_1})$ and $D_{2,i} = z_i\ln(\frac{z_i}{\mu^c_1}) + (1-z_i)\ln(\frac{1-z_i}{1-\mu^c_1})$.

Based on the above notations and discussions, we can obtain the following results regarding the right-hand side of (13): $\forall i > 1$ and $k \geq 1$

$$\mathbb{E}\Big\{\frac{1}{p^r_{i,\tau_k+1}}\Big\} \leq 1 + \Theta\Big(\frac{3R_{1,i}e^{-D_{1,i}k}}{y_i(1-y_i)(k+1)(R_{1,i}-1)^2} + e^{-2\epsilon_i^2 k}$$
$$+ \frac{1+R_{1,i}}{1-y_i}e^{-D_{1,i}k} + e^{-\frac{1}{2}k\epsilon_i^2} + \frac{1}{\exp\{\frac{\epsilon_i^2k^2}{2(k+1)}\}-1}\Big); \quad (14)$$

If $z_i \geq 1$, $\mathbb{E}\{\frac{1}{p^c_{i,\tau_k+1}}\} = 1$; otherwise,

$$\mathbb{E}\Big\{\frac{1}{p^c_{i,\tau_k+1}}\Big\} \leq 1 + \Theta\Big(\frac{2e^{-D_{2,i}k}}{z_i(1-z_i)(1-R_{2,i})^2} + e^{-2\epsilon_i^2 k}$$
$$+ \frac{1}{z_iR_{2,i}}e^{-D_{2,i}k} + e^{-\frac{1}{2}\epsilon_i^2 k} + \frac{1}{\exp\{\frac{\epsilon_i^2k^2}{2(k+1)}\}-1}\Big). \quad (15)$$

Specifically, if $z_i \geq 1$, $\mathbb{E}[\frac{1}{p_{i,\tau_0+1}}] \leq \frac{1}{1-y_i}$; otherwise $\mathbb{E}[\frac{1}{p_{i,\tau_0+1}}] \leq \frac{1}{(1-y_i)z_i}$. The derivations of (14) and (15) need tight estimations of partial binomial sums and careful algebraic operations, which can be found in the online full version of this work [Xia *et al.*, 2015].

According to (12) and (13), to bound $\sum_{t=1}^{\infty}\mathbb{P}\{I_t = i, E^\theta_i(t), B_t > 0\}$, we only need to multiply each term in (14) by each one in (15), and sum up all the multiplicative terms over $k$ from 0 to $\infty$ except the constant 1. Using Taylor series expansion, we can verify that w.r.t. $\gamma$,

$$\frac{1}{D_{1,i}} = O\Big(\frac{1}{\epsilon_i^2(\gamma)}\Big), \frac{3R_{1,i}}{y_i(1-y_i)(R_{1,i}-1)^2} = O\Big(\frac{1}{\epsilon_i^2(\gamma)}\Big).$$

Therefore, if $\epsilon_i(\gamma) + \mu^c_1 \geq 1$, we have that w.r.t. $\gamma$,

$$\sum_{k=0}^{\infty}\Big(\mathbb{E}\Big\{\frac{1}{p_{i,\tau_k+1}}\Big\} - 1\Big) = O\Big(\frac{1}{\epsilon_i^4(\gamma)}\Big). \quad (16)$$

Similarly, if $\epsilon_i(\gamma) + \mu^c_1 < 1$, we can obtain that w.r.t $\gamma$,

$$\sum_{k=0}^{\infty}\Big(\mathbb{E}\Big\{\frac{1}{p_{i,\tau_k+1}}\Big\} - 1\Big) = O\Big(\frac{1}{(1-\mu^c_1-\epsilon_i(\gamma))\epsilon_i^6(\gamma)}\Big). \quad (17)$$

Note that the constants in the $O(\cdot)$ of (16) and (17) do not depend on $B$ (but depend on $\mu^r_i$ and $\mu^c_i$ $\forall i \in [K]$).

***Step 4**: Bound $\mathbb{E}\{T_i\}$ $\forall i \geq 2$ for Bernoulli bandits.*

Combining (6), (11), (16) and (17), we can get the following result:

$$\mathbb{E}\{T_i\} \leq 1 + \frac{2\ln B}{\delta_i^2(\gamma)} + \frac{35}{\delta_i^2(\gamma)\mu^c_{\min}} + \Phi_i(\gamma)$$
$$\leq 1 + \frac{2\ln B}{\gamma^2(\mu^c_i\Delta_i)^2}\Big(\frac{\mu^r_1}{\mu^c_1}+1\Big)^2 + O\Big(\frac{1}{\gamma^2}\Big) + \Phi_i(\gamma), \quad (18)$$

in which $\Delta_i$ is defined in (2) and $\Phi_i(\gamma)$ is defined in (5).

According to the Eqn. (3) of Lemma 2, we can eventually obtain the regret bound of Budgeted Thompson Sampling as shown in Theorem 3 by first multiplying $\mu^c_i\Delta_i$ on the right of (18) and then summing over $i$ from 2 to $K$.

## 4.2 Analysis for General Bandits

The regret bound we obtained for Bernoulli bandits in the previous subsection also works for general bandits, as shown in Theorem 3.

The result for general bandits is a little surprising since the problem of general bandits seems more difficult than the Bernoulli bandit problem, and one may expect a slightly looser asymptotic regret bound. The reason why we can retain the same regret bound lies in the Bernoulli trials of the general bandits. Intuitively, the Bernoulli trials can be seen as the intermediate that can transform the general bandits to Bernoulli bandits while keeping the expected reward and cost of each arm unchanged. Therefore, when $B$ is large, there should not be too many differences in the regret bound between the Bernoulli bandits and general bandits.

Specifically, similar to the case of Bernoulli bandits, in order to bound the regret of the BTS algorithm for the general bandits, we only need to bound $\mathbb{E}\{T_i\}$ (according to inequality (4)). To bound $\mathbb{E}\{T_i\}$, we also need four steps similar to those described in the previous subsection. In addition, we need one extra step which is related to the Bernoulli trials. Details are described as below.

*$\mathcal{S}0$: Obtain the success probabilities of the Bernoulli trials.* Denote the reward and cost of arm $i$ at round $t$ as $r_i(t)$ and $c_i(t)$ respectively. Denote the Bernoulli trial results of arm $i$ at round $t$ as $\tilde{r}_i(t)$ (for reward) and $\tilde{c}_i(t)$ (for cost). We need to prove $\mathbb{P}\{\tilde{r}_i(t) = 1\} = \mu^r_i$ and $\mathbb{P}\{\tilde{c}_i(t) = 1\} = \mu^c_i$, which is straightforward:

$$\mathbb{P}\{\tilde{r}_i(t) = 1\} = \mathbb{E}\{\mathbb{E}[\mathbf{1}\{\tilde{r}_i(t) = 1\}|r_i(t)]\} = \mathbb{E}[r_i(t)] = \mu^r_i,$$
$$\mathbb{P}\{\tilde{c}_i(t) = 1\} = \mathbb{E}\{\mathbb{E}[\mathbf{1}\{\tilde{c}_i(t) = 1\}|c_i(t)]\} = \mathbb{E}[c_i(t)] = \mu^c_i.$$

*$\mathcal{S}1$: Decompose $\mathbb{E}\{T_i\}$*: This step is the same as Step 1 in the Bernoulli bandit case. For the general bandit case, $\mathbb{E}\{T_i\}$ can also be bounded by inequality (6).

*$\mathcal{S}2$: Bound $\sum_{t=1}^{\infty}\mathbb{P}\{\overline{E^\theta_i(t)}, n_{i,t} \geq \lceil L_i\rceil, B_t > 0\}$.* $\mathcal{S}2$ is almost the same as Step 2 in the proof for Bernoulli bandits but contains some minor changes. For the general bandits, we have $c_i(t) \in [0,1]$ rather than $c_i(t) \in \{0,1\}$. Then we

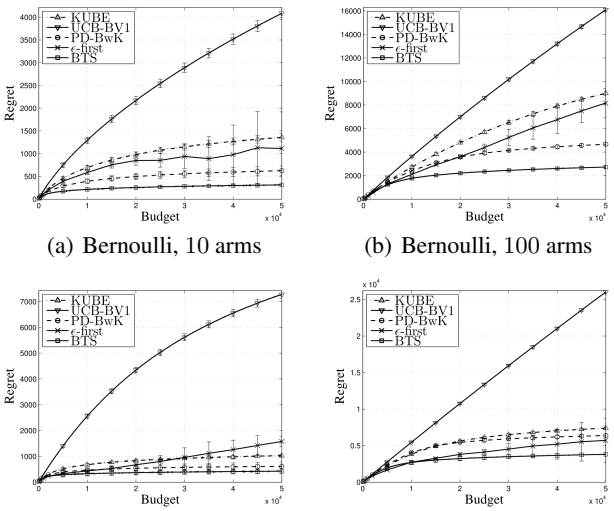
have $\sum_{t=1}^{\infty} \mathbb{P}\{B_t > 0\} \leq \frac{B+1}{\mu^c_{\min}}$, and can get a similar result to (11).

*$\mathcal{S}3$: Bound* $\sum_{t=1}^{\infty} \mathbb{P}\{I_t = i, E_i^{\theta}(t), B_t > 0\}$. Since we have already got the success probabilities of the Bernoulli trials, this step is the same as Step 3 for the Bernoulli bandits.

*$\mathcal{S}4$:* Substituting the results of $\mathcal{S}2$ and $\mathcal{S}3$ into the corresponding terms in (6), we can get an upper bound of $\mathbb{E}\{T_i\}$ for the general bandits. Then according to (4), for general bandits, the results in Theorem 3 can be eventually obtained.

The classical MAB problem in [Auer *et al.*, 2002] can be regarded as a special case of the budgeted MAB problem by setting $c_i(t) = 1\ \forall i \in [K], t \geq 1$, and $B$ is the total pulling time. According to [Lai and Robbins, 1985], we can verify the order of the distribution-dependent regret bound of the budgeted MAB problem is at least $O(\ln B)$. Comparing with the two algorithms in [Ding *et al.*, 2013], we have the following results:

**Remark 4** *By setting $\gamma = \frac{1}{\sqrt{2}}$ in Theorem 3, we can see that BTS gets a tighter asymptotic regret bound in terms of the constants before $\ln B$ than the two algorithms proposed in [Ding* et al.*, 2013].*




(a) Bernoulli, 10 arms (b) Bernoulli, 100 arms

(c) Multinomial, 10 arms (d) Multinomial, 100 arms

Figure 1: Regrets under different bandit settings

## 5 Numerical Simulations

In addition to the theoretical analysis of the BTS algorithm, we are also interested in its empirical performance. We conduct a set of experiments to test the empirical performance of BTS algorithm and present the results in this section.

For comparison purpose, we implement four baseline algorithms: (1) the $\epsilon$-first algorithm [Tran-Thanh *et al.*, 2010] with $\epsilon = 0.1$; (2) a variant of the PD-BwK algorithm [Badanidiyuru *et al.*, 2013]: at each round, pull the arm with the maximum $\frac{\min\{\overline{r}_{i,t}+\varphi(\overline{r}_{i,t},n_{i,t}),1\}}{\max\{\overline{c}_{i,t}-\varphi(\overline{c}_{i,t},n_{i,t}),0\}}$, in which $\overline{r}_{i,t}$ ($\overline{c}_{i,t}$) is the average reward (cost) of arm $i$ before round $t$, $\varphi(x, N) = \sqrt{\frac{\nu x}{N}} + \frac{\nu}{N}$ and $\nu = 0.25 \log(BK)$; (3) the UCB-BV1 algorithm [Ding *et al.*, 2013]; (4) a variant of the KUBE algorithm [Tran-Thanh *et al.*, 2012]: at round $t$, pull the arm with the maximum ratio $\left(\overline{r}_{i,t} + \sqrt{\frac{2 \ln t}{n_{i,t}}}\right) / \overline{c}_{i,t}$. $\epsilon$-first and PD-BwK need to know $B$ in advance, and thus we try several budgets as $\{100, 200, 500, 1K, 2K, 5K, 10K, 15K, 20K, \cdots, 50K\}$. BTS, UCB-BV1 and KUBE do not need to know $B$ in advance, and thus by setting $B = 50K$ we can get their empirical regrets for every budget smaller than $50K$.

We simulate bandits with two different distributions: one is the Bernoulli distribution (simple), and the other is the multinomial distribution (complex). Their parameters are randomly chosen. For each distribution, we simulate a 10-armed case and a 100-armed case. We then independently run the experiments for 500 times and report the average performance of each algorithm.

The average regret and the standard deviation of each algorithm over 500 random runs are shown in Figure 1. From the figure we have the following observations:

First, for both the Bernoulli distribution and the multinomial distribution, and for both the 10-arm case and 100-arm case, our proposed BTS algorithm has clear advantage over the baseline methods: It achieves the lowest regrets. Furthermore, the standard deviation of the regrets of BTS over 500 runs is small, indicating that its performance is very stable across different random run of the experiments.

Second, as the number of arms increases (from 10 to 100), the regrets of all the algorithms increase, given the same budget. This is easy to understand because more budget is required to make good explorations on more arms.

Third, the standard deviation of the regrets of the $\epsilon$-first algorithm is much larger than the other algorithms, which shows that $\epsilon$-first is not stable under certain circumstances. Take the 10-armed Bernoulli bandit for example: when $B = 50K$, during the 500 random runs, there are 13 runs that $\epsilon$-first cannot identify the optimal arm. The average regret over the 13 runs is 4630. However, over the other 487 runs, the average regret of $\epsilon$-first is 1019.9. Therefore, the standard derivation of $\epsilon$-first is large. In comparison, the BTS algorithm is much more stable.

Overall speaking, the simulation results demonstrate the effectiveness of our proposed Budgeted Thompson Sampling algorithm.

## 6 Conclusion and Future work

In this paper, we have extended the Thompson sampling algorithm to the budgeted MAB problems. We have proved that our proposed algorithm has a distribution-dependent regret bound of $O(\ln B)$. We have also demonstrated its empirical effectiveness using several numerical simulations.

For future work, we plan to investigate the following aspects: (1) We will study the distribution-free regret bound of Budgeted Thompson Sampling. (2) We will try other priors (e.g., the Gaussian prior) to see whether a better regret bound and empirical performance can be achieved in this way. (3) We will study the setting that the reward and the cost are correlated (e.g., an arm with higher reward is very likely to have higher cost).

## Acknowledgments

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

## References

[Agmon Ben-Yehuda *et al.*, 2013] Orna Agmon Ben-Yehuda, Muli Ben-Yehuda, Assaf Schuster, and Dan Tsafrir. Deconstructing amazon ec2 spot instance pricing. *ACM Transactions on Economics and Computation*, 1(3):16, 2013.

[Agrawal and Goyal, 2012] Shipra Agrawal and Navin Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *Proceedings of the 25th Annual Conference on Learning Theory (COLT)*, June 2012.

[Agrawal and Goyal, 2013] Shipra Agrawal and Navin Goyal. Further optimal regret bounds for thompson sampling. *Sixteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, April 2013.

[Amin *et al.*, 2012] Kareem Amin, Michael Kearns, Peter Key, and Anton Schwaighofer. Budget optimization for sponsored search:censored learning in mdps. In *Uncertainty in Artificial Intelligence (UAI)*. Uncertainty in Artificial Intelligence (UAI), August 2012.

[Ardagna *et al.*, 2011] Danilo Ardagna, Barbara Panicucci, and Mauro Passacantando. A game theoretic formulation of the service provisioning problem in cloud systems. In *Proceedings of the 20th international conference on World wide web*, pages 177–186. ACM, 2011.

[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, April 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.

[Badanidiyuru *et al.*, 2013] Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. Bandits with knapsacks. In *FOCS*, pages 207–216. IEEE, 2013.

[Bubeck *et al.*, 2009] Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. Pure exploration in multi-armed bandits problems. In *Algorithmic Learning Theory*, pages 23–37. Springer, 2009.

[Chapelle and Li, ] Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. In *NIPS (2011)*.

[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 *AAAI*, 2013.

[Fisher, 1980] Marshall L Fisher. Worst-case analysis of heuristic algorithms. *Management Science*, 26(1):1–17, 1980.

[Gai *et al.*, 2010] Yi Gai, Bhaskar Krishnamachari, and Rahul Jain. Learning multiuser channel allocations in cognitive radio networks: A combinatorial multi-armed bandit formulation. In *New Frontiers in Dynamic Spectrum, 2010 IEEE Symposium on*, pages 1–9. IEEE, 2010.

[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.

[Kaufmann *et al.*, 2012a] Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On bayesian upper confidence bounds for bandit problems. In *AISTATS*, pages 592–600, 2012.

[Kaufmann *et al.*, 2012b] Emilie Kaufmann, Nathaniel Korda, and Rémi Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In *Algorithmic Learning Theory*, pages 199–213. Springer, 2012.

[Kohli and Krishnamurti, 1992] Rajeev Kohli and Ramesh Krishnamurti. A total-value greedy heuristic for the integer knapsack problem. *Operations research letters*, 12(2):65–71, 1992.

[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.

[Martello and Toth, 1990] Silvano Martello and Paolo Toth. *Knapsack problems: algorithms and computer implementations*. John Wiley & Sons, Inc., 1990.

[Thompson, 1933] William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, pages 285–294, 1933.

[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*, April 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 Stavrogiannis, Victor Naroditskiy, Valentin Robu, Nicholas R Jennings, and Peter Key. Efficient regret bounds for online bid optimisation in budget-limited sponsored search auctions. In *UAI*. AUAI, July 2014.

[Vazirani, 2001] Vijay V Vazirani. *Approximation algorithms*. Springer Science & Business Media, 2001.

[Xia *et al.*, 2015] Yingce Xia, Haifang Li, Tao Qin, Nenghai Yu, and Tie-Yan Liu. Thompson sampling for budgeted multi-armed bandits. *arXiv preprint*, 2015.

[Yu and Nikolova, 2013] Jia Yuan Yu and Evdokia Nikolova. Sample complexity of risk-averse bandit-arm selection. In *IJCAI*, pages 2576–2582. AAAI Press, 2013.

[yves Audibert and Bubeck, 2009] Jean yves Audibert and Sbastien Bubeck. Minimax policies for adversarial and stochastic bandits. In *COLT*, pages 773–818, 2009.