JMLR: Workshop and Conference Proceedings 45:317–332, 2015 ACML 2015

# Budgeted Bandit Problems with Continuous Random Costs*

**Yingce Xia** YINGCE.XIA@GMAIL.COM
*University of Science and Technology of China, Hefei, 230027, P.R. China*

**Wenkui Ding** DINGWENKUI@GMAIL.COM
**Xu-Dong Zhang** ZHANGXD@TSINGHUA.EDU.CN
*Tsinghua University, Beijing, 100084, P.R.China*

**Nenghai Yu** YNH@USTC.EDU.CN
*University of Science and Technology of China, Hefei, 230027, P.R. China*

**Tao Qin** TAOQIN@MICROSOFT.COM
*Microsoft Research, Beijing, 100080, P.R. China*

**Editor:** Geoffrey Holmes and Tie-Yan Liu

## Abstract

We study the budgeted bandit problem, where each arm is associated with both a reward and a cost. In a budgeted bandit problem, the objective is to design an arm pulling algorithm in order to maximize the total reward before the budget runs out. In this work, we study both multi-armed bandits and linear bandits, and focus on the setting with continuous random costs. We propose an upper confidence bound based algorithm for multi-armed bandits and a confidence ball based algorithm for linear bandits, and prove logarithmic regret bounds for both algorithms. We conduct simulations on the proposed algorithms, which verify the effectiveness of our proposed algorithms.

**Keywords:** Budgeted Multi-armed Bandit, Budgeted Linear Bandit, Online Learning

## 1. Introduction

Bandit problems are typical examples of sequential decision making problems in an uncertain environment. Many different kinds of bandit problems have been studied in the literature, including multi-armed bandits (MAB) and linear bandits. In a multi-armed bandit problem, an agent faces a slot machine with $K$ arms, each of which has an unknown reward distribution (throughout this work, we restrict our attention to stochastic bandits); at each round $t$, he/she needs to pull one arm from the $K$ candidates and will receive a random reward drawn from the unknown distribution associated with the pulled arm. The goal of the agent is to sequentially pull arms so as to maximize the total reward. In a linear bandit problem, the set of $K$ arms is replaced by a compact set of $d$-dimensional vectors (Dani et al., 2008), and the expected reward of an arm linearly depends on its vector representation. The key of bandit algorithms is to make a trade-off between exploration and exploitation: the agent should simultaneously consider pulling the best arm based on the information collected in the past (exploitation) and trying other arms to collect useful information for future pulling (exploration).

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

In this paper, we consider the budgeted bandit problem, in which pulling an arm is costly and the objective is to maximize the total reward subject to a budget constraint on the total cost. A variety of Internet applications can fall into this problem. For example, both the real-time bidding problem in online advertising (Chakrabarti et al., 2008) and the cloud service provisioning problem in cloud computing (Ardagna et al., 2011) can be regarded as budgeted bandits. In fact, this problem has been studied in the literature, with different settings (Tran-Thanh et al., 2010, 2012; Ding et al., 2013). For example, Tran-Thanh et al. (2010) study budgeted multi-armed bandits with fixed costs and propose an epsilon-first algorithm with regret bound $O(B^{\frac{2}{3}})$ where $B$ is the budget. Tran-Thanh et al. (2012) further derive an improved regret bound of $O(\log B)$ for the same setting. Ding et al. (2013) study multi-armed bandits with discrete random costs and propose two algorithms based on upper confidence bound (UCB) whose regret bounds are both $O(\log B)$. Different from these works, in this paper, we study budgeted bandits with continuous costs. Many real world applications can be modeled as budgeted bandits with continuous costs. An example is bid optimization in sponsored search: the per-click payment of a bid (the cost of an arm) is allowed to have six digits after the radix point, which is more convenient to be modeled as a continuous variable. Another example is virtual instance selection in cloud computing. When a user has a certain job to run in the cloud, he/she can select different type of virtual instances [1]; the running time of his/her job on a selected instance is a continuous variable and so for the payment.

We consider two kinds of budgeted bandits, multi-armed bandits and linear bandits.

For the budgeted MAB problem with continuous random costs, we propose a UCB-based algorithm called Budget-UCB. The algorithm can be regarded as an extension of the UCB-BV1 algorithm proposed in (Ding et al., 2013), however, its theoretic analysis is far different from that of UCB-BV1. Actually the discrete costs are very critical to UCB-BV1 because the analysis of its regret bound is based on induction which only works for a well-ordered set (the set of discrete costs is well ordered). In contrast, the interval $[0, 1]$ of the continuous costs is not well-ordered at all. As a result, with the techniques used in (Ding et al., 2013), UCB-BV1 cannot achieve a meaningful regret bound in our setting. To tackle this challenge, we make modifications to UCB-BV1 and explore new proof techniques so as to obtain a nearly optimal distribution dependent regret bound of $O(\log B)$.

For the budgeted linear bandit problem, all the arms constitute a subspace of $d$-dimensional space $\mathbb{R}^d$, and the expected reward and cost of an arm are fixed but unknown linear functions of the vector representation of the arm. For this setting, we generalize the algorithms proposed by Dani et al. (2008) and Abbasi-Yadkori et al. (2011), and obtain a new algorithm called Budget-CB. In each step of the algorithms proposed in (Dani et al., 2008) and (Abbasi-Yadkori et al., 2011), a confidence ball is constructed, which contains, with high probability, the parameter of the fixed and unknown linear function associated with the expected rewards. Since they do not take costs into consideration, they tend to pull arms with large rewards and do not work well for our setting. Consider a simple example with two arms: the expected reward and cost for the first arm are both 1, while the expected reward and cost for the second arm are 0.5 and 0.1 respectively. For this example, the algorithms proposed in (Dani et al., 2008) tend to pull the first arm frequently. However, it

1. Amazon EC2 Instances: https://aws.amazon.com/ec2/instance-types/?nc1=h_ls.

is clear that a much better policy is to pull the second arm more frequently. To tackle this challenge, we construct two confidence balls, one for the reward and the other for the cost, and then optimize the reward-to-cost ratio in these two confidence balls. We prove that by doing so, we can achieve a distribution dependent regret bound of polylog($B$).

The proofs for the regret bounds of the above two algorithms are technical: We need to bridge the relationship between the expected pulling time of suboptimal arms, bound the expected pulling time for suboptimal arms, and deal with the randomness of the stopping time. We also need to carefully characterize the conditional independence between rewards, costs, the arm pulled by the algorithm, and the stopping time (as random variables) in the regret analysis. We believe that our proof techniques can be applied to more general settings and can be of independent interest to the community.

To summarize, we have two main contributions: (1) To the best of our knowledge, this is the first work that obtains an $O(\log(B))$ distribution dependent regret bound of UCB based algorithms for budgeted multi-armed bandits with continuous random costs; (2) It is the first time that budgeted linear bandits are investigated and a distribution dependent regret bound of order polylog($B$) is derived.

The remaining of the paper is organized as follows. Related work is discussed in Section 2. The algorithm for budgeted multi-armed bandits with continuous random costs is introduced and analyzed in Section 3. The algorithm for budgeted linear bandits is described and analyzed in Section 4. Experimental results are given in Section 5. Section 6 concludes the paper and discusses possible further research directions.

## 2. Related Work

While most existing works on bandit problems (Agrawal et al., 1988; Auer et al., 2002; György et al., 2007; Dani et al., 2008; Abbasi-Yadkori et al., 2011; Wang et al., 2009; Kleinberg, 2004; Li et al., 2010) do not consider budget constraints, which does not hold in many real-world applications, there have been some attempts that take budget constraint into consideration (Audibert et al., 2010; Bubeck et al., 2009; Guha and Munagala, 2007; Tran-Thanh et al., 2010, 2012; Badanidiyuru et al., 2013). Roughly speaking, these related works can be classified into two categories.

In the first category, the cost of pulling an arm is fixed and no exploration for costs is needed. In the pure exploration problem studied by Audibert et al. (2010) and Bubeck et al. (2009), the exploration phase and exploitation phase are separated. In the exploration phase, it is assumed that pulling each arm has a unit cost, and the budget is imposed on the total number of pulls. After the exploration, the agent is asked to choose the best arm according to certain criteria, and this arm will always be used in the future exploitation phase (which is not associated with a cost any more). Guha and Munagala (2007) study how to minimize the budget given a confidence level for the optimality of the selected arm. Tran-Thanh et al. (2010) propose $\epsilon$-first methods, which use a fixed proportion ($\epsilon$) of the budget for exploration, and the remaining proportion for exploitation. Tran-Thanh et al. (2012) propose a UCB style algorithm for the regret minimization problem of fixed cost budgeted bandits.

In the second category, the cost of pulling an arm is a discrete random variable, instead of a fixed value. Due to the uncertainty of costs, both the exploitation phase and exploration

phase should take the costs into account. Ding et al. (2013) develop two UCB based algorithms to solve this problem and obtain nearly optimal regret bounds.

In this paper, we study a more general setting for UCB based algorithms, in which the cost of pulling an arm is random and continuous. We call it budgeted bandits with continuous random costs, which covers previous work with UCB based algorithms (both fixed costs (Tran-Thanh et al., 2012) and discrete random costs (Ding et al., 2013)) as its special cases. We also propose the budgeted linear bandits in this work. A closely related work is (Badanidiyuru et al., 2013), which also studies bandit problems with continuous costs. The differences between (Badanidiyuru et al., 2013) and our work lie in two aspects: (1) We study both multi-armed bandits and linear bandits, while Badanidiyuru et al. (2013) only consider multi-armed bandits; (2) Our analysis focuses on distribution dependent regret bounds, while Badanidiyuru et al. (2013) focus on distribution independent bounds.

## 3. Budgeted Multi-armed Bandits with Continuous Random Costs

### 3.1. Problem Formulation of Budgeted Multi-armed Bandits

In the budgeted multi-armed bandit (MAB) problems, the bandit is associated with a finite set of arms denoted by $\{1, 2, \cdots, K\}$ ($K \geq 2$). For ease of reference, denote the set $\{1, 2, \cdots, K\}$ as $[K]$. The agent is asked to pull one of the $K$ arms at each round $t$. If arm $i$ is pulled, the agent will receive a reward $r_{i,t}$ and a cost $c_{i,t}$. We assume that the pairs $\{(r_{i,t}, c_{i,t})\}_{t=1}^{\infty}$ are independently and identically drawn from an unknown continuous distribution,[2] and both rewards and costs take real values from $[0, 1]$.

We use $B$ to denote the budget, which is a known parameter and will constrain the total number of pulls, i.e., the stopping time of the pulling procedure. The stopping time $T_{a,B}$ of a pulling algorithm $a$ is a random variable depending on $B$, and can be mathematically formulated by $\sum_{t=1}^{T_{a,B}-1} c_{a_t,t} \leq B < \sum_{t=1}^{T_{a,B}} c_{a_t,t}$. The total reward collected up to time $T_{a,B}$ by the pulling algorithm $a$ is defined as $R_a = \sum_{t=1}^{T_{a,B}} r_{a_t,t}$. Let $R^*$ denote the optimal expected total reward when the distribution of rewards and costs are known. We use *expected regret* to evaluate the performance of the algorithm which is defined as below,

$$\text{Regret} \stackrel{\text{def}}{=} R^* - \mathbb{E}[R_a] = R^* - \mathbb{E}\Big[\sum_{t=1}^{T_{a,B}} r_{a_t,t}\Big], \tag{1}$$

where the $\mathbb{E}$ is taken over the randomness of rewards, costs, and the pulling algorithm.

### 3.2. Algorithm

Our proposed Budget-UCB algorithm is shown in Algorithm 1. In the algorithm, for any arm $i$, $n_{i,t}$, $\bar{r}_{i,t}$ and $\bar{c}_{i,t}$ denote the pulling times, average reward and average cost of arm $i$ before round $t$. Mathematically,

$$n_{i,t} = \sum_{s=1}^{t-1} \mathbf{1}(a_s = i), \quad \bar{r}_{i,t} = \frac{\sum_{s=1}^{t-1} r_{i,s}\mathbf{1}(a_s = i)}{n_{i,t}}, \quad \bar{c}_{i,t} = \frac{\sum_{s=1}^{t-1} c_{i,s}\mathbf{1}(a_s = i)}{n_{i,t}},$$

2. Note that we only assume that the reward-cost pairs at different rounds are independent, and do not assume that the cost and reward at the same round are independent. In some applications the cost of pulling an arm may be correlated to the reward.

where $\mathbf{1}(\cdot)$ is the indicator function. The parameter $\lambda$ is a positive lower bound of the expected costs of all arms.[3] In Algorithm 1, the first term on the right side of Eqn. (2) is an exploitation term, which is the average reward-to-cost ratio. This term forces the algorithm to choose those arms with higher marginal rewards. The second term is an exploration term: it is proportional to $\epsilon_{i,t}$ thus prefers the arms pulled less frequently in the past (this is in the same spirit with the exploration term in UCB1 (Auer et al., 2002)), and inversely proportional to $\bar{c}_{i,t}$ thus prefers the arms with lower costs (intuitively, due to the limited budget, the arms with lower costs are worth exploring because they consume less). The third term is a hybrid term that performs joint exploitation and exploration.

**Algorithm 1** The Budget-UCB Algorithm (Input: $\lambda$)

**Initialization:** Pull each arm $i$ once in the first $K$ steps, set $t = K$.

1: **while** $\sum_{s=1}^{t} c_{a_s,s} \leq B$ **do**

2: Set $t = t + 1$.

3: Define $\epsilon_{i,t} = \sqrt{\frac{2\log(t-1)}{n_{i,t}}}$. Calculate the index $D_{i,t}$ of each arm $i$ as follows.

$$D_{i,t} = \frac{\bar{r}_{i,t}}{\bar{c}_{i,t}} + \frac{\epsilon_{i,t}}{\bar{c}_{i,t}} + \frac{\epsilon_{i,t}}{\bar{c}_{i,t}} \frac{\min\{\bar{r}_{i,t} + \epsilon_{i,t}, 1\}}{\max\{\bar{c}_{i,t} - \epsilon_{i,t}, \lambda\}}; \tag{2}$$

4: Pull the arm $a_t$ with the largest index: $a_t \in \arg\max_i D_{i,t}$.

5: **end while**

**Return:** The accumulate reward $\sum_{s=1}^{t} r_{a_s,s}$.

### 3.3. Regret Analysis

We first declare the notations that will be frequently used throughout this work: (1) $\mu_i^r$ and $\mu_i^c$ denote the expected reward and cost of arm $i$ respectively. (2) $i^*$ denotes the arm with the largest expected reward to expected cost ratio, i.e., $i^* = \arg\max_i \frac{\mu_i^r}{\mu_i^c}$. We name arm $i^*$ as the optimal arm. Without loss of generality, we assume there is a unique optimal arm.[4]. The others arms are called suboptimal arms. (3) For any $i \neq i^*$, $\Delta_i = \frac{\mu_{i^*}^r}{\mu_{i^*}^c} - \frac{\mu_i^r}{\mu_i^c}$. Obviously, $\Delta_i > 0$. (4) $B_t$ denotes the budget left at the beginning of round $t$. (5) $n_{i,T_{a,B}}$ denotes the pulling time of arm $i$ when the algorithm $a$ stops. (6) $\mu_{\min}^c$ denotes $\min_{i\in[K]}\{\mu_i^c\}$.

The regret analysis of the Budget-UCB algorithm is different from previous works (Tran-Thanh et al., 2010; Ding et al., 2013). This is because we consider continuous costs and the main proof technique in (Ding et al., 2013), the induction methodology designed for discrete costs, cannot be directly applied. Therefore, we first propose a framework that relates the regret with the pulling time of each suboptimal arm.

3. Note that to use this algorithm, one needs to know $\lambda$ in advance, which is a lower bound of the expected costs of all the arms. In some applications, the expected costs can be known and it is easy to set $\lambda$. In some other applications, the costs correspond to monetary payments and one can set $\lambda$ to be the minimum unit of the currency.

4. If there are multiple optimal arms (denote them as $\mathcal{O}$), we can randomly pick an $i^* \in \mathcal{O}$ as the "optimal" arm. We have $\Delta_i = 0\ \forall i \in \mathcal{O}$, which will not contribute to upper bound of the regret in Lemma 1.

**Lemma 1** *The expected regret of any algorithm $a$ (denoted as* $\mathrm{Regret}(a)$*) is upper bounded by:* $\mathrm{Regret}(a) \leq \sum_{i\neq i^*} \mu_i^c \Delta_i \mathbb{E}[n_{i,T_{a,B}}] + \mu_{i^*}^r/\mu_{i^*}^c$.

**Proof** The proof of Lemma 1 consists of two steps.
($\mathcal{S}1$) We will prove that $R^* \leq (B+1)\mu_{i^*}^r/\mu_{i^*}^c$.
*Proof*: Let $a^*$ denote the optimal pulling algorithm. We can obtain that

$$
\begin{aligned}
R^* &= \mathbb{E}\Big[\sum_{t=1}^{\infty} r_{a_t^*,t}\mathbf{1}(B_t \geq 0)\Big] = \sum_{t=1}^{\infty}\sum_{i=1}^{K} \mathbb{E}\big[r_{i,t}|a_t^*=i, B_t\geq 0\big]\mathbb{P}(a_t^*=i, B_t\geq 0) \\
&\leq \Big(\sum_{t=1}^{\infty}\sum_{i=1}^{K} \mathbb{E}\big[c_{i,t}|a_t^*=i, B_t\geq 0\big]\mathbb{P}(a_t^*=i, B_t\geq 0)\Big)\frac{\mu_{i^*}^r}{\mu_{i^*}^c} = \mathbb{E}\Big[\sum_{t=1}^{\infty} c_{a_t^*,t}\mathbf{1}(B_t\geq 0)\Big]\frac{\mu_{i^*}^r}{\mu_{i^*}^c} \qquad (3) \\
&= \mathbb{E}\Big[\sum_{t=1}^{T_{a^*,B}} c_{a_t^*,t}\Big]\frac{\mu_{i^*}^r}{\mu_{i^*}^c} \leq \frac{(B+1)\mu_{i^*}^r}{\mu_{i^*}^c}.
\end{aligned}
$$

The inequality in (3) holds because: (i) $a_t^*$ and $B_t$ are only related to the pulling history until round $t-1$. Thus, $\mathbb{E}\big[r_{i,t}|a_t^*=i, B_t\geq 0\big] = \mu_i^r \leq \mu_i^c\mu_{i^*}^r/\mu_{i^*}^c$. (ii) Similar to the discussion in (i), we can obtain that $\mathbb{E}\big[c_{i,t}|a_t^*=i, B_t\geq 0\big] = \mu_i^c$. □
($\mathcal{S}2$) We will prove Lemma 1 in this step. According to ($\mathcal{S}1$), the optimal reward can be upper bounded as $R^* \leq \frac{(B+1)\mu_{i^*}^r}{\mu_{i^*}^c} < \mathbb{E}\Big[\sum_{t=1}^{T_{a,B}} c_{a_t,t}+1\Big]\frac{\mu_{i^*}^r}{\mu_{i^*}^c}$. Accordingly, for any algorithm $a$, the regret can be bounded as

$$
\begin{aligned}
\mathrm{Regret}(a) &\leq \mathbb{E}\Big[\sum_{t=1}^{T_{a,B}} c_{a_t,t}+1\Big]\frac{\mu_{i^*}^r}{\mu_{i^*}^c} - \mathbb{E}\Big[\sum_{t=1}^{T_{a,B}} r_{a_t,t}\Big] = \mathbb{E}\Big[\sum_{t=1}^{T_{a,B}} \frac{\mu_{i^*}^r}{\mu_{i^*}^c}c_{a_t,t} - r_{a_t,t}\Big] + \frac{\mu_{i^*}^r}{\mu_{i^*}^c} \\
&= \mathbb{E}\Big[\sum_{t=1}^{\infty}\sum_{i=1}^{K}\Big(\frac{\mu_{i^*}^r}{\mu_{i^*}^c}c_{i,t} - r_{i,t}\Big)\mathbf{1}(a_t=i, B_t\geq 0)\Big] + \frac{\mu_{i^*}^r}{\mu_{i^*}^c} \\
&= \sum_{i=1}^{K}\sum_{t=1}^{\infty}\mathbb{E}\Big[\Big(\frac{\mu_{i^*}^r}{\mu_{i^*}^c}c_{i,t} - r_{i,t}\Big)\Big|a_t=i, B_t\geq 0\Big]\mathbb{P}(a_t=i, B_t\geq 0) + \frac{\mu_{i^*}^r}{\mu_{i^*}^c} \\
&= \sum_{t=1}^{\infty}\sum_{i\neq i^*}\mu_i^c\Delta_i\mathbb{P}(a_t=i, B_t\geq 0) + \frac{\mu_{i^*}^r}{\mu_{i^*}^c} = \sum_{i\neq i^*}\mu_i^c\Delta_i\mathbb{E}[n_{i,T_{a,B}}] + \frac{\mu_{i^*}^r}{\mu_{i^*}^c}.
\end{aligned}
$$

Thus we reach the conclusion in Lemma 1. ■

Then we only need to focus on bounding the expected pulling time of each suboptimal arm. We introduce two notations $T_0$ and $N$ as follows: (both $T_0$ and $N$ are deterministic)

$$
T_0 = \Big\lfloor \frac{2B}{\mu_{\min}^c}\Big\rfloor; \quad N = 8\Big(\frac{1+\frac{1}{\lambda}+\frac{\Delta_i}{2}}{\Delta_i\mu_i^c}\Big)^2\log T_0. \qquad (4)
$$

Without loss of generality, we assume $N > 1$, which is easy to be satisfied with $B \geq 1$ given the costs are upper bounded by 1. We can verify that $n_{i,T_{a,B}}$ can be decomposed as

$$
n_{i,T_{a,B}} = \sum_{t=1}^{\infty}\mathbf{1}\{a_t=i, B_t\geq 0\} \leq 1+N+\sum_{t=K+1}^{T_0}\mathbf{1}\{a_t=i, n_{i,t}\geq N\} + \sum_{t=T_0+1}^{\infty}\mathbf{1}\{B_t\geq 0\}. \qquad (5)
$$

For ease of reference, for any $t \geq 1$ and a given $i \neq i^*$, denote the event $\{a_t=i, n_{i,t}\geq N\}$ as $E_t^0$. The expectations of the last two terms in (5) will be bounded in ($\mathcal{S}1$) and ($\mathcal{S}2$).

($\mathcal{S}$1): *Bound* $\mathbb{E}[\sum_{t=K+1}^{T_0}\mathbf{1}\{E_t^0\}]$. Define $H_{i,t}=\frac{\epsilon_{i,t}}{\bar{c}_{i,t}}+\frac{\epsilon_{i,t}}{\bar{c}_{i,t}}\frac{\min\{\bar{r}_{i,t}+\epsilon_{i,t},1\}}{\max\{\bar{c}_{i,t}-\epsilon_{i,t},\lambda\}}$. If $E_t^0$ happens, at least one of the following four events happens.

$$E_t^1:|\bar{c}_{i,t}-\mu_i^c|\geq\epsilon_{i,t};\qquad E_t^2:\frac{\bar{r}_{i,t}}{\bar{c}_{i,t}}\geq\frac{\mu_i^r}{\mu_i^c}+H_{i,t},|\bar{c}_{i,t}-\mu_i^c|<\epsilon_{i,t};$$

$$E_t^3:\frac{\bar{r}_{i^*,t}}{\bar{c}_{i^*,t}}\leq\frac{\mu_{i^*}^r}{\mu_{i^*}^c}-H_{i^*,t},,|\bar{c}_{i,t}-\mu_i^c|<\epsilon_{i,t};\qquad E_t^4:\frac{\mu_{i^*}^r}{\mu_{i^*}^c}<\frac{\mu_i^r}{\mu_i^c}+2H_{i,t},|\bar{c}_{i,t}-\mu_i^c|<\epsilon_{i,t},n_{i,t}\geq N.$$

This is because: (i) One of $E_t^1$ and $\overline{E_t^1}$ must hold; (ii) Conditioned on $\overline{E_t^1}$, at least one of the following three events must hold: (ii-a) $\frac{\bar{r}_{i,t}}{\bar{c}_{i,t}}\geq\frac{\mu_i^r}{\mu_i^c}+H_{i,t}$; (ii-b) $\frac{\bar{r}_{i^*,t}}{\bar{c}_{i^*,t}}\leq\frac{\mu_{i^*}^r}{\mu_{i^*}^c}-H_{i^*,t}$; (ii-c) $\frac{\mu_{i^*}^r}{\mu_{i^*}^c}<\frac{\mu_i^r}{\mu_i^c}+2H_{i,t}$. Thus, $\mathbb{E}[\sum_{t=K+1}^{T_0}\mathbf{1}\{E_t^0\}]\leq\sum_{j=1}^4\sum_{t=K+1}^{T_0}\mathbb{P}\{E_t^j\}$. We will bound the sum of the four probabilities from ($\mathcal{S}$1-1) to ($\mathcal{S}$1-4).
($\mathcal{S}$1-1)*Bound* $\sum_{t=K+1}^{T_0}\mathbb{P}\{E_t^1\}$: This step depends on the following lemma:

**Lemma 2** *For Budget-UCB, we have*

$$\mathbb{P}(|\bar{r}_{i,t}-\mu_i^r|\geq\epsilon_{i,t})\leq 2(t-1)^{-3},\quad \mathbb{P}(|\bar{c}_{i,t}-\mu_i^c|\geq\epsilon_{i,t})\leq 2(t-1)^{-3}. \tag{6}$$

**Proof** We adapt the proof of Theorem 1 in (Auer et al., 2002). Denote $\bar{X}_{i,n}$ as the average reward of $n$ independent plays of arm $i$. We have

$$\mathbb{P}(|\bar{r}_{i,t}-\mu_i^r|\geq\epsilon_{i,t})=\sum_{n=1}^{t-1}\mathbb{P}(|\bar{r}_{i,t}-\mu_i^r|\geq\epsilon_{i,t},n_{i,t}=n)=\sum_{n=1}^{t-1}\mathbb{P}\Big(|\bar{r}_{i,t}-\mu_i^r|\geq\sqrt{\frac{2\log(t-1)}{n_{i,t}}},n_{i,t}=n\Big)$$
$$\leq\sum_{n=1}^{t-1}\mathbb{P}\Big(|\bar{X}_{i,n}-\mu_i^r|\geq\sqrt{\frac{2\log(t-1)}{n}}\Big)\leq^{\triangle}\sum_{n=1}^{t-1}2(t-1)^{-4}=2(t-1)^{-3},$$

where the inequality marked with $^{\triangle}$ comes from the conventional Chernoff-Hoeffding inequality. Similarly, the second inequality of (6) can be obtained. ■

By Lemma 2, we have that $\sum_{t=K+1}^{T_0}\mathbb{P}\{E_t^1\}\leq 2\sum_{t=3}^{\infty}1/(t-1)^3\leq 1$.
($\mathcal{S}$1-2)*Bound* $\sum_{t=K+1}^{T_0}\mathbb{P}\{E_t^2\}$. If event $E_t^2$ happens, we have $|\bar{r}_{i,t}-\mu_i^r|\geq\epsilon_{i,t}$. Otherwise, we have $|\bar{r}_{i,t}-\mu_i^r|<\epsilon_{i,t}$, which implies that $\mu_i^r<\bar{r}_{i,t}+\epsilon_{i,t}$. One can also verify that $E_t^2$ happens means that $\mu_i^c>\bar{c}_{i,t}-\epsilon_{i,t}$. Therefore, we can obtain

$$\frac{\bar{r}_{i,t}}{\bar{c}_{i,t}}-\frac{\mu_i^r}{\mu_i^c}=\frac{(\bar{r}_{i,t}-\mu_i^r)\mu_i^c+(\mu_i^c-\bar{c}_{i,t})\mu_i^r}{\bar{c}_{i,t}\mu_i^c}<\frac{\epsilon_{i,t}}{\bar{c}_{i,t}}+\frac{\epsilon_{i,t}\mu_i^r}{\bar{c}_{i,t}\mu_i^c}<\frac{\epsilon_{i,t}}{\bar{c}_{i,t}}+\frac{\epsilon_{i,t}}{\bar{c}_{i,t}}\frac{\min\{\bar{r}_{i,t}+\epsilon_{i,t},1\}}{\max\{\bar{c}_{i,t}-\epsilon_{i,t},\lambda\}}=H_{i,t}.$$

By Lemma 2, we have $\sum_{t=K+1}^{T_0}\mathbb{P}\{E_t^2\}\leq\sum_{t=3}^{\infty}\mathbb{P}(|\bar{r}_{i,t}-\mu_i^r|\geq\epsilon_{i,t})\leq 2\sum_{t=3}^{\infty}(t-1)^{-3}=1$.
($\mathcal{S}$1-3)*Bound* $\sum_{t=K+1}^{T_0}\mathbb{P}\{E_t^3\}$. We can verify that if the event $E_t^3$ holds, we have $|\bar{r}_{i^*,t}-\mu_{i^*}^r|\geq\epsilon_{i^*,t}$. Leveraging Lemma 2 again, we have $\sum_{t=K+1}^{T_0}\mathbb{P}\{E_t^3\}\leq 1$.
($\mathcal{S}$1-4)*Bound* $\sum_{t=K+1}^{T_0}\mathbb{P}\{E_t^4\}$. For event $E_t^4$, given $|\bar{c}_{i,t}-\mu_i^c|<\epsilon_{i,t}$ and $n_{i,t}\geq N$, it can be verified that for any $t\leq T_0$, $\frac{\mu_{i^*}^r}{\mu_{i^*}^c}\geq\frac{\mu_i^r}{\mu_i^c}+2H_{i,t}$ (i.e., $\mathbf{1}(E_t^4)=0$) as follows:

$$\epsilon_{i,t}=\sqrt{\frac{2\log(t-1)}{n_{i,t}}}<\sqrt{\frac{2\log(T_0)}{N}}=\frac{\Delta_i\mu_i^c}{2+\frac{2}{\lambda}+\Delta_i}\Rightarrow\epsilon_{i,t}\left(1+\frac{1}{\lambda}+\frac{\Delta_i}{2}\right)\leq\frac{\Delta_i}{2}\mu_i^c$$
$$\Rightarrow\quad\epsilon_{i,t}\left(1+\frac{1}{\lambda}\right)\leq\frac{\Delta_i}{2}(\mu_i^c-\epsilon_{i,t})\leq\frac{\Delta_i}{2}\bar{c}_{i,t}\Rightarrow H_{i,t}\leq\frac{\epsilon_{i,t}}{\bar{c}_{i,t}}\left(1+\frac{1}{\lambda}\right)\leq\frac{\Delta_i}{2}=\frac{1}{2}\left(\frac{\mu_{i^*}^r}{\mu_{i^*}^c}-\frac{\mu_i^r}{\mu_i^c}\right).$$

According to ($\mathcal{S}$1-1) to ($\mathcal{S}$1-4), we can obtain that $\mathbb{E}[\sum_{t=K+1}^{T_0}\mathbf{1}\{E_t^0\}] \leq 3$.
($\mathcal{S}$2): *Bound* $\sum_{t=T_0+1}^{\infty}\mathbb{P}\{B_t \geq 0\}$. We use two sub steps to bound it.
($\mathcal{S}$2-1) *We prove that for any* $t \geq T_0$, $\mathbb{P}\{B_{t+1} \geq 0\} \leq \exp\big\{-2(B-t\mu_{\min}^c)^2/t\big\}$.
*Proof.* Denote the cost at round $s \in \{1, 2, \cdots, t\}$ as $c(s)$. We have

$$\mathbb{E}[c(1)] + \cdots + \mathbb{E}[c(t)] \geq t\mu_{\min}^c \geq \lfloor \frac{2B}{\mu_{\min}^c} \rfloor \mu_{\min}^c \geq 2B - \mu_{\min}^c > B. \tag{7}$$

Accordingly, we have that

$$\begin{aligned}
&\mathbb{P}\{B_{t+1} \geq 0\} = \mathbb{P}\{c(1) + \cdots + c(t) \leq B\} \\
\leq &\mathbb{P}\{c(1) + \cdots + c(t) - \mathbb{E}[c(1)] - \cdots - \mathbb{E}[c(t)] \leq B - \mathbb{E}[c(1)] - \cdots - \mathbb{E}[c(t)]\} \\
\leq &\exp\{-\frac{2(B - \mathbb{E}[c(1)] - \cdots - \mathbb{E}[c(t)])^2}{t}\} \qquad \text{(according to Hoeffding inequality)} \\
\leq &\exp\Big\{-\frac{2(B-t\mu_{\min}^c)^2}{t}\Big\}. \quad \text{(according to (7))} \ . \qquad \square
\end{aligned}$$

($\mathcal{S}$2-2) *Bound* $\sum_{t=T_0+1}^{\infty}\mathbb{P}\{B_t \geq 0\}$ according to ($\mathcal{S}$2-1). Define $\mathcal{T}(B)$ as follows:

$$\mathcal{T}(B) = \exp\Big\{-\frac{\mu_{\min}^c(B-\mu_{\min}^c)^2}{B}\Big\} + (\frac{3B}{\mu_{\min}^c})\exp\Big\{-\frac{B\mu_{\min}^c}{2}\Big\} + \frac{1}{(\mu_{\min}^c)^2}\exp\{1 - 2B\mu_{\min}^c\}. \tag{8}$$

Please note that $\mathcal{T}(B)$ tends to zero as $B$ tends to infinity. Then we have

$$\begin{aligned}
&\sum_{t=T_0+1}^{\infty}\mathbb{P}\{B_t \geq 0\} \leq \exp\Big\{-\frac{\mu_{\min}^c(B-\mu_{\min}^c)^2}{B}\Big\} + \sum_{t=T_0+1}^{\infty}\mathbb{P}\{B_{t+1} \geq 0\} \\
\leq &\exp\Big\{-\frac{\mu_{\min}^c(B-\mu_{\min}^c)^2}{B}\Big\} + \sum_{l=0}^{\infty}\exp\Big\{-\frac{2(B+l\mu_{\min}^c)^2}{\frac{2B}{\mu_{\min}^c}+l}\Big\} \\
\leq &\exp\Big\{-\frac{\mu_{\min}^c(B-\mu_{\min}^c)^2}{B}\Big\} + (\lfloor\frac{2B}{\mu_{\min}^c}\rfloor + 1)\exp\Big\{-\frac{B\mu_{\min}^c}{2}\Big\} + \sum_{l=T_0+1}^{\infty}\exp\{-l(\mu_{\min}^c)^2\} \leq \mathcal{T}(B).
\end{aligned}$$

Therefore, according to Lemma 1, (5), ($\mathcal{S}$1) and ($\mathcal{S}$2), we come to the following theorem:

**Theorem 3** *The regret of Budget-UCB is at most*

$$\sum_{i\neq i^*}\frac{8\left(1+\frac{1}{\lambda}+\frac{\Delta_i}{2}\right)^2}{\Delta_i\mu_i^c}\log\frac{2B}{\mu_{\min}^c} + (3+\mathcal{T}(B))\sum_{i\neq i^*}^{K}\Delta_i\mu_i^c + \frac{\mu_{i^*}^r}{\mu_{i^*}^c}, \tag{9}$$

*where* $\mathcal{T}(B)$ *is defined in* (8).

Please note that $\mathcal{T}(B)$ tends to zero as $B$ tends to infinity. We make the following discussions about the theorem.

- This theorem shows that the asymptotic regret bound of the proposed algorithm is of $O(\log B)$ when $B$ is sufficiently large. More precisely, given a bandit (i.e., the parameters $\{\mu_i^r, \mu_i^c\}_{i=1}^K$ are fixed), the regret of the algorithm grows in the logarithmic order of the budget $B$ when $B$ goes to infinity.

- When $B$ is not very large, other terms in the regret bound may play a more important role than $\log(B)$. We conducted a set of numerical experiments to evaluate how the algorithm performs when $B$ is not very large. The experimental results are reported in Section 5.
- Badanidiyuru et al. (2013) propose an algorithm whose regret is

$$O(\sqrt{K\mathrm{OPT}} + \mathrm{OPT}\sqrt{K/B}),$$

where OPT is the expected total reward of the optimal policy. Since OPT is linear order of $B$, the regret of their algorithm is upper bounded by $O(\sqrt{B})$. However, this is not to say that our algorithm is better than theirs, because our $O(\log B)$ bound is distribution dependent, which is incomparable with their distribution independent bound.

## 4. Budgeted Linear Bandits with Continuous Random Costs

### 4.1. Problem Formulation of Budgeted Linear Bandits

The linear bandit (Dani et al., 2008; Abbasi-Yadkori et al., 2011) problem is a more complex and practical one than the conventional multi-armed bandits . We extend the budgeted multi-armed bandits to the linear setting and proposed the *budgeted linear bandits*. In budgeted linear bandits, the set of arms is represented by a $d$-dimensional compact set $\mathcal{K} \subseteq \mathbb{R}^d$ $(d \geq 1)$. An agent chooses an arm $x_t \in \mathcal{K}$ at round $t$, then results in a random reward $r_t(x_t) \in [0,1]$ and a random cost $c_t(x_t) \in [0,1]$. For any arm $x \in \mathcal{K}$, the reward at round $t$, the cost at round $t$, the expected reward and cost are of the following linear forms:

$$r_t(x) = x^\top \mu^r + \eta_t^r,\ c_t(x) = x^\top \mu^c + \eta_t^c,\ \mathbb{E}[r_t(x)] = x^\top \mu^r,\ \mathbb{E}[c_t(x)] = x^\top \mu^c, \tag{10}$$

where $\mu^r, \mu^c \in \mathbb{R}^d$ are unknown parameters, and $\mu^r$ and $\mu^c$ are bounded, i.e., $\|\mu^r\|_2 \leq S, \|\mu^c\|_2 \leq S$, $S > 0$ and $S$ is known in advance. $x^\top$ is the transpose of $x$ and $\eta_t^r$, $\eta_t^c$ are C-sub-Gaussion.[5] For any $x \in \mathcal{K}, \|x\|_2 \leq L$.[6] We further assume that the expected costs of all the arms are larger than a positive parameter $\lambda > 0$ which is known in advance and smaller than 1. The expected reward of all the arms are no less than zero. Mathematically, for any $x \in \mathcal{K}$, $x^\top \mu^r \in (0,1)$ and $x^\top \mu^c \in [\lambda, 1)$. This assumption is very natural since in practice whatever non-trivial action an agent takes, he/she needs to afford a certain non-zero cost. This assumption means that the zero action $x = 0$ is not in the compact set $\mathcal{K}$. Taking the action $x = 0$ will lead to both zero expected reward and zero expected cost and it is ignored in our problem.

Similar to the conventional budgeted MAB, we also define an optimal arm $x^*$, which is $\arg\max_{x \in \mathcal{K}} \frac{x^\top \mu^r}{x^\top \mu^c}$ and $\gamma$ denotes the corresponding ratio, i.e., $\frac{(x^*)^\top \mu^r}{(x^*)^\top \mu^c}$. We require that there is a positive gap $\Delta_{\min}$ between the optimal and sub-optimal arms, i.e., $\Delta_{\min} = \frac{(x^*)^\top \mu^r}{(x^*)^\top \mu^c} - \sup_{x \in \mathcal{K} \setminus \{x^*\}} \frac{x^\top \mu^r}{x^\top \mu^c}$ and $\Delta_{\min} > 0$. Please note such the $\Delta_{\min}$ also exists in (Dani et al., 2008) (the $\Delta$ in (Dani et al., 2008)). On the other hand, $\Delta_{\max}$ denotes $\frac{(x^*)^\top \mu^r}{(x^*)^\top \mu^c} - \inf_{x \in \mathcal{K} \setminus \{x^*\}} \frac{x^\top \mu^r}{x^\top \mu^c}$.

5. We adapt the definition of C-sub-Gaussion from (Abbasi-Yadkori et al., 2011), which is: For any $\xi \in \mathbb{R}$, $\mathbb{E}[e^{\xi \eta_t} | x_1, x_2, \cdots, x_t, \eta_1, \cdots, \eta_{t-1}] \leq \exp\{\frac{\xi^2 C^2}{2}\}$
6. This implies that for any $t$, $\|x_t\|_2 \leq L$.

Adapting the stopping time and the regret for conventional budgeted bandits, for any algorithm $a$, the stopping time $T_{a,B}$ and the regret can be defined as follows:

$$\sum_{t=1}^{T_{a,B}-1} c_t(x_t) \le B < \sum_{t=1}^{T_{a,B}} c_t(x_t); \quad \text{Regret}(a) = R^* - \mathbb{E}\Big[\sum_{t=1}^{T_{a,B}} r_t(x_t)\Big], \tag{11}$$

where $R^*$ is the total expected reward of the optimal algorithm when $\mu^r$ and $\mu^c$ are known.

### 4.2. Algorithm

Our proposed algorithm for budgeted linear bandits is shown in Algorithm 2. We call it Budget-CB since confidence balls (CB) play an important role in the algorithm. The algorithm follows the framework *Optimism in the Face of Uncertainty principle for Linear bandits (OFUL)* (Abbasi-Yadkori et al., 2011). In Algorithm 2, $\|x\|_A$ denotes $\sqrt{x^\top Ax}$ where $A$ is positive definite and $I_d$ is the $d$-dimensional identity matrix. To run Algorithm 2, we need two hyper parameters: $\delta \in (0,1)$ and $\rho > 0$.

**Algorithm 2** The Budget-CB Algorithm (Input: $\lambda$, hyper parameters $\rho > 0$, $\delta \in (0,1)$)

**Initialization:** Pull an arm randomly in the first step, set $t = 1$.

1: **while** $\sum_{s=1}^{t} c_s(x_s) \le B$ **do**

2: Set $t = t + 1$.

3: Define $\epsilon(t) = C\sqrt{d\log\left(1 + tL^2/(\rho d)\right) + 2\log(1/\delta)} + S\sqrt{\rho}$.
Calculate $A_t = \rho I_d + \sum_{s=1}^{t-1} x_s x_s^\top$, $\bar{\mu}_t^r = A_t^{-1}\sum_{s=1}^{t-1} r_s(x_s)x_s$, $\bar{\mu}_t^c = A_t^{-1}\sum_{s=1}^{t-1} c_s(x_s)x_s$.
Construct confidence balls as follows:
$B_t^r = \{\nu^r; \|\nu^r - \bar{\mu}_t^r\|_{A_t} \le \epsilon(t)\}$, $B_t^c = \{\nu^c; \|\nu^c - \bar{\mu}_t^c\|_{A_t} \le \epsilon(t)\}$;

4: Pull one arm $x_t \in \arg\max_{x\in\mathcal{K}, \nu^r\in B_t^r, \nu^c \in B_t^c} \frac{\min\{x^\top \nu^r, 1\}}{\max\{x^\top \nu^c, \lambda\}}$.

5: **end while**

**Return:** The accumulate reward $\sum_{s=1}^{t} r_s(x_s)$.

At each round, we update the estimated reward vector $\bar{\mu}_t^r$ and the estimated cost vector $\bar{\mu}_t^c$ according to the Step 3 of the algorithm, which are in fact the results of minimizing the regularized square loss on the past decisions $x_t$ and observations $(r_s(x_s), c_s(x_s))$. In Step 3, we also construct confidence balls $B_t^r$ and $B_t^c$ centered at $\bar{\mu}_t^r$ and $\bar{\mu}_t^c$ respectively. The radius of the balls is $\epsilon(t)$ which controls the degree of exploration. The decision of $x_t$ is made by jointly maximizing the reward-to-cost ratio among all the possible $x$, $\nu_r$ and $\nu_c$ in $\mathcal{K}$, $B_t^r$ and $B_t^c$.[7]

### 4.3. Regret Analysis

In this subsection, we upper bound the regret of Budget-CB algorithm. Similarly, we first give the framework of analyzing budgeted linear bandit. Let $\tau_a$ denote the pulling time of all the suboptimal arms when the algorithm $a$ stops. The analysis of the regret bound of Budget-CB depends on the following lemma.

7. The computation of $x_t$ is similar to that in (Dani et al., 2008): if the decision set is small, we can enumerate all choices; for some other special cases, such as $\mathcal{K}$ is a polytope, the optimization problem is indeed NP-hard (Sahni, 1974), the exact computation is not computationally practical, and we can adopt local search methods to compute it.

**Lemma 4** *For any algorithm $a$, the regret is at most $\Delta_{\max}\mathbb{E}\{\tau_a\}+\gamma$.*

**Proof** $(\mathcal{S}1)$ We will prove that $R^* \le (B+1)\gamma$.
*Proof of $(\mathcal{S}1)$.* Let $a^*$ denote the optimal policy. Denote $\mathbb{P}\{B_t \ge 0, \cup_{i=1}^d \{x_t^i \le x^i\}\}$ as $F_t^B(x)$, where $x_t^i$ is the $i$-th dimension of $x_t$, and so is $x^i$. Again, we can safely write $\mathbb{E}[r_t(x)|B_t \ge 0, x_t = x]$ and $\mathbb{E}[c_t(x)|B_t \ge 0, x_t = x]$ as $\mathbb{E}[r_t(x)]$ and $\mathbb{E}[c_t(x)]$ respectively, since the event $\{B_t \ge 0, x_t = x\}$ is only related to the history until round $t-1$, while $r_t(x)$ and $c_t(x)$ are independent of the history. Similar to the $(\mathcal{S}1)$ of the proof of Lemma 1,

$$\mathbb{E}\Big[\sum_{t=1}^{T_{a^*,B}} r_t(x_t)\Big] = \sum_{t=1}^{\infty}\int_{\mathcal{K}} \mathbb{E}[r_t(x)]\mathrm{d}F_t^B(x) \le \Big(\sum_{t=1}^{\infty}\int_{\mathcal{K}} \mathbb{E}[c_t(x)]\mathrm{d}F_t^B(x)\Big)\gamma$$
$$= \mathbb{E}\Big[\sum_{t=1}^{\infty} c_t(x_t)\mathbf{1}(B_t \ge 0)\Big]\gamma \le (B+1)\gamma. \qquad \square$$

$(\mathcal{S}2)$ We will prove Lemma 4 according to the results in $(\mathcal{S}1)$. We can verify that $R^* \le \Big(\mathbb{E}\Big[\sum_{t=1}^{T_{a,B}} c_t(x_t)\Big]+1\Big)\gamma$. Thus, the regret can be upper bounded as follows:

$$\Big(\mathbb{E}\Big[\sum_{t=1}^{T_{a,B}} c_t(x_t)\Big]+1\Big)\gamma - \mathbb{E}\Big[\sum_{t=1}^{T_{a,B}} r_t(x_t)\Big] = \mathbb{E}\sum_{t=1}^{T_{a,B}}[\gamma c_t(x_t) - r_t(x_t)] + \gamma$$
$$= \sum_{t=1}^{\infty}\mathbb{E}[(\gamma c_t(x_t) - r_t(x_t))\mathbf{1}(B_t \ge 0)] + \gamma = \sum_{t=1}^{\infty}\int_{\mathcal{K}} \mathbb{E}\left[\gamma c_t(x) - r_t(x)\right]\mathrm{d}F_t^B(x) + \gamma$$
$$= \sum_{t=1}^{\infty}\int_{\mathcal{K}} \left(\gamma x^\top \mu^c - x^\top \mu^r\right)\mathrm{d}F_t^B(x) + \gamma \le \sum_{t=1}^{\infty}\int_{\mathcal{K}\setminus\{x^*\}} x^\top \mu^c \Delta_{\max}\mathrm{d}F_t^B(x) + \gamma$$
$$\le \sum_{t=1}^{\infty}\int_{\mathcal{K}\setminus\{x^*\}} \Delta_{\max}\mathrm{d}F_t^B(x) + \gamma = \Delta_{\max}\mathbb{E}[\tau_a] + \gamma.$$

Therefore, we can get Lemma 4. ■

To get $\mathbb{E}\{\tau_a\}$, we also need to divide the pulling rounds before round $\tau_0$ and after round $\tau_0$, where $\tau_0$ denotes $\lfloor \frac{2B}{\lambda} \rfloor$. Similar to the $(\mathcal{S}2)$ for conventional budgeted multi-armed bandits, we can get that $\mathbb{E}[\sum_{t=\tau_0+1}^{\infty} \mathbf{1}\{a_t \ne x^*, B_t \ge 0\}] \le \mathcal{L}(B)$, where

$$\mathcal{L}(B) = \exp\Big\{-\frac{\lambda(B-\lambda)^2}{B}\Big\} + \frac{3B}{\lambda}\exp\Big\{-\frac{B\lambda}{2}\Big\} + \frac{1}{\lambda^2}\exp\{1-2B\lambda\}. \tag{12}$$

Then, we focus on bounding the expected pulling time of suboptimal arms before the $\tau_0$-th round, which is denoted as $\mathbb{E}\{\tau'\}$. The proof consists of three steps.
$(\mathcal{S}1)$ *We will prove that the two parameter vectors $\mu^r$ and $\mu^c$ lie in the confidence balls with high probabilities.*

**Lemma 5** *For the Budget-CB algorithm, with probability at least $1-\delta$, for any $t \ge 1$, $\mu^r \in B_t^r, \mu^c \in B_t^c$.*

**Proof** Define the following notations: $\eta^r = (\eta_1^r, \cdots, \eta_t^r)^\top$, $\eta^c = (\eta_1^c, \cdots, \eta_t^c)^\top$, $X = (x_1, \cdots, x_t)^\top$. According to Theorem 1 in (Abbasi-Yadkori et al., 2011), for any $\delta > 0$ and $t$, with probability at least $1-\delta$,

$$\|X^\top \eta^r\|_{A_t^{-1}} \le C\sqrt{2\log\left(\frac{det(A_t)^{1/2}det(\rho I_d)^{-1/2}}{\delta}\right)}. \tag{13}$$

Since $\bar{\mu}_t^r = (X^\top X + \rho I)^{-1} X^\top (X\mu^r + \eta^r) = (X^\top X + \rho I)^{-1} X^\top \eta^r + \mu^r - \rho (X^\top X + \rho I)^{-1} \mu^r$, for any $x \in \mathbb{R}^d$, we have $x^\top \bar{\mu}_t^r - x^\top \mu^r = x^\top (X^\top X + \rho I)^{-1} X^\top \eta^r - \rho x^\top (X^\top X + \rho I)^{-1} \mu^r = \langle x, X^\top \eta^r \rangle_{A_t^{-1}} - \rho \langle x, \mu^r \rangle_{A_t^{-1}}$, where $\langle x, y \rangle_M = x^\top M y$ and $M$ is positive definite. According to Lemma 10 in (Abbasi-Yadkori et al., 2011), we have $det(A_t) \leq (\rho + \frac{tL^2}{d})^d$. As a result, (13) can be further bounded as

$$\|X^\top \eta^r\|_{A_t^{-1}} \leq C \sqrt{d \log\Big(1 + \frac{tL^2}{\rho d}\Big) + 2 \log \frac{1}{\delta}}.$$

When the above event holds, using Cauchy-Schwarz inequality, we can obtain

$$\begin{aligned} x^\top \bar{\mu}_t^r - x^\top \mu^r \leq \|x\|_{A_t^{-1}} \Big( \|X^\top \eta^r\|_{A_t^{-1}} + \rho \|\mu^r\|_{A_t^{-1}} \Big) \leq \|x\|_{A_t^{-1}} \Big( \|X^\top \eta^r\|_{A_t^{-1}} + \sqrt{\rho} \|\mu^r\|_2 \Big) \\ \leq \|x\|_{A_t^{-1}} \Big( C \sqrt{d \log\Big(1 + \frac{tL^2}{\rho d}\Big) + 2 \log \frac{1}{\delta}} + \sqrt{\rho} S \Big) = \|x\|_{A_t^{-1}} \epsilon(t), \end{aligned}$$

where the first inequality holds because $\|\mu^r\|^2_{A_t^{-1}} \leq 1/\lambda_{\min}(A_t) \|\mu^r\|_2^2 \leq 1/\rho \|\mu^r\|_2^2$ and $\lambda_{\min}(A_t)$ is the minimum eigenvalue of $A_t$, which is certainly no smaller than $\rho$. Finally, by setting $x = A_t(\bar{\mu}_t^r - \mu^r)$, we get $\|\bar{\mu}_t^r - \mu^r\|_{A_t} \leq \epsilon(t)$.

Similarly, for the cost of the arms, we have $x^\top \bar{\mu}_t^c - x^\top \mu^c \leq \|x\|_{A_t^{-1}} \epsilon(t)$. Then by setting $x = A_t(\bar{\mu}_t^c - \mu^c)$, we get $\|\bar{\mu}_t^c - \mu^c\|_{A_t} \leq \epsilon(t)$. ■

($\mathcal{S}2$) *We will give a high probability bound of $\tau'$.*

**Lemma 6** *For Budget-CB, if $\rho \geq \max\{1, L^2\}$, with probability at least $1 - \delta$, $\tau'$ is upper bounded by*

$$\tau' \leq \frac{32d}{\lambda^4 \Delta_{\min}^2} \Big( C \sqrt{d \log\Big(1 + \frac{\tau_0 L^2}{\rho d}\Big) + 2 \log \frac{1}{\delta}} + \sqrt{\rho} S \Big)^2 \log\Big(1 + \frac{\tau_0 L^2}{d\rho}\Big). \tag{14}$$

**Proof** Define the following two formulas in (15). It is easy to see that $\varphi(t) \geq \lambda^2 \Delta_{\min}$ when $x_t \neq x^*$.

$$\text{(a) } \varphi(t) = \langle x^*, \mu^r \rangle \langle x_t, \mu^c \rangle - \langle x_t, \mu^r \rangle \langle x^*, \mu^c \rangle; \text{ (b) } (\tilde{\mu}^r, \tilde{\mu}^c) = \arg \max_{\nu^r \in B_t^r; \mu^c \in B_t^c} \frac{\min\{\langle x_t, \nu^r \rangle, 1\}}{\max\{\langle x_t, \nu^c \rangle, \lambda\}}. \tag{15}$$

At round $t$, since arm $x_t$ is pulled, we have

$$\frac{\min\{\langle x_t, \tilde{\mu}^r \rangle, 1\}}{\max\{\langle x_t, \tilde{\mu}^c \rangle, \lambda\}} \geq \frac{\min\{\langle x^*, \mu^r \rangle, 1\}}{\max\{\langle x^*, \mu^c \rangle, \lambda\}} = \frac{\langle x^*, \mu^r \rangle}{\langle x^*, \mu^c \rangle}, \tag{16}$$

which shows that $\langle x^*, \mu^r \rangle \langle x_t, \tilde{\mu}^c \rangle \leq \langle x^*, \mu^c \rangle \langle x_t, \tilde{\mu}^r \rangle$. Given $\|\bar{\mu}_t^r - \mu^r\|_{A_t} \leq \epsilon(t)$ $\|\bar{\mu}_t^c - \mu^c\|_{A_t} \leq \epsilon(t)$, we can obtain that

$$\begin{aligned} \varphi(t) = \langle x^*, \mu^r \rangle \langle x_t, \mu^c \rangle - \langle x_t, \mu^r \rangle \langle x^*, \mu^c \rangle \leq \langle x^*, \mu^r \rangle \langle x_t, \mu^c - \tilde{\mu}^c \rangle - \langle x^*, \mu^c \rangle \langle x_t, \mu^r - \tilde{\mu}^r \rangle \\ = \langle x^*, \mu^r \rangle \langle x_t, \mu^c - \bar{\mu}_t^c \rangle + \langle x^*, \mu^r \rangle \langle x_t, \bar{\mu}_t^c - \tilde{\mu}^c \rangle - \langle x^*, \mu^c \rangle \langle x_t, \mu^r - \bar{\mu}_t^r \rangle - \langle x^*, \mu^c \rangle \langle x_t, \bar{\mu}_t^r - \tilde{\mu}^r \rangle \\ \leq 4 \epsilon(t) \|x_t\|_{A_{t-1}^{-1}}. \end{aligned} \tag{17}$$

By Lemma 11 in (Abbasi-Yadkori et al., 2011), we get that if $\rho \geq \max\{1, L^2\}$, with probability at least $1 - \delta$,

$$\tau' \lambda^4 \Delta_{\min}^2 \leq \sum_{t=1}^{\tau_0} \varphi(t)^2 \leq 32d \Big( C \sqrt{d \log\Big(1 + \frac{\tau_0 L^2}{\rho d}\Big) + 2 \log \frac{1}{\delta}} + \sqrt{\rho} S \Big)^2 \log\Big(1 + \frac{\tau_0 L^2}{d\rho}\Big). \tag{18}$$

Then we can get Lemma 6. ■

($\mathcal{S}3$) *Give the expected regret of Budget-CB.* According to Lemma 6, we can get that

$$\mathbb{E}\{\tau'\} \leq (1-\delta)\Big\{\frac{32d}{\lambda^4\Delta_{\min}^2}\Big(C\sqrt{d\log\Big(1+\frac{\tau_0 L^2}{\rho d}\Big)+2\log\frac{1}{\delta}}+\sqrt{\rho}S\Big)^2\log\Big(1+\frac{\tau_0 L^2}{d\rho}\Big)\Big\}+\delta\tau_0. \quad (19)$$

The expected pulling time of suboptimal arms is bounded by $\mathbb{E}\{\tau'\}+\mathcal{L}(B)$. Therefore, by setting $\delta=\frac{1}{B}$ and by Lemma 4, we can get that

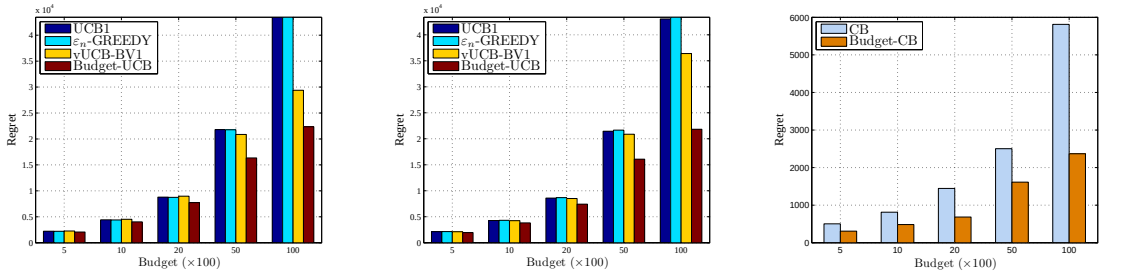
**Theorem 7** *For any $\rho \geq \max\{1, L^2\}$, the expected regret of Budget-CB is at most*

$$\frac{32d\Delta_{\max}}{\lambda^4\Delta_{\min}^2}\Big(C\sqrt{d\log\Big(1+\frac{2BL^2}{\lambda\rho d}\Big)+2\log B}+\sqrt{\rho}S\Big)^2\log\Big(1+\frac{2BL^2}{\lambda d\rho}\Big)+(\frac{2}{\lambda}+\mathcal{L}(B))\Delta_{\max}+\gamma.$$

Theorem 7 shows that the asymptotic regret of Budget-CB is $O\Big(\frac{d^2\Delta_{\max}}{\Delta_{\min}^2}\log^2(B)\Big)$, which is a polylog(B) regret bound.

## 5. Experiments

In this section, we report our experimental results on the performance of the proposed algorithms. First, we simulate a budgeted multi-armed bandit as follows. (1) A bandit




($a$) Continuous costs ($b$) Discrete costs ($c$) Linear bandit problems

Figure 1: Experimental results of the MAB

with 100 arms is created. (2) The reward/cost of each pulling of an arm is sampled from beta distributions. The two parameters of each beta distribution are uniformly sampled from $[1, 5]$. (3) The budget is chosen from the set $\{500, 1000, 2000, 5000, 10000\}$. For each value of budget, all the algorithms are run for 100 times and their average regrets are examined. (4) For comparison purpose, UCB1 and the $\varepsilon_n$-GREEDY algorithm (Auer et al., 2002) are implemented as baselines. We also propose a variant of UCB-BV1 as baseline[8] (Ding et al., 2013): we only need to replace the $D_{i,t}$ in (2) of Algorithm 1 with Eqn. (20-a). For ease of reference, denote the variant as vUCB-BV1.

$$\text{(20-a) } D_{i,t}=\frac{\bar{r}_{i,t}}{\bar{c}_{i,t}}+1.5(1+\frac{1}{\lambda})\epsilon_{i,t}; \quad \text{(20-b) } D_{i,t}=\frac{\bar{r}_{i,t}}{\bar{c}_{i,t}}+\frac{(1+\frac{1}{\lambda})\epsilon_{i,t}}{\lambda-\epsilon_{i,t}}. \quad (20)$$

8. We find that if we directly run the original UCB-BV1 (by replacing the $D_{i,t}$ in (2) of Algorithm 1 with Eqn. (20-b)), when the $B$ is not sufficiently large, the exploration term in Eqn. (20-b)) is too large. As a result, UCB-BV1 wastes much budget on suboptimal arms and does not achieve good results.

The results of the simulation are shown in Figure 1($a$), from which we have the following observations. (1) UCB1 and the $\varepsilon_n$-greedy algorithm perform the worst. The reason is that these two algorithms are not designed for the budgeted bandit problems and do not take budget constraints into consideration: they tend to pull arms with large empirical rewards which may also have large costs and therefore run out of budgets quickly. To verify this, we have listed the mean rewards ($\bar{r}$), mean costs ($\bar{c}$), ratio of mean reward over mean cost ($\bar{r}/\bar{c}$), the stopping time ($\tau$) and the pulling percentage of the optimal (%opt) of each algorithm across the 100 runs in Table 1. It is clear that UCB1 and $\varepsilon_n$-GREEDY have larger average rewards but much smaller stopping time than the other two algorithms. (2) Budget-UCB outperforms the vUCB-BV1 algorithm, since it can find the optimal arm with fewer pullings and then focus on pulling the optimal arm, which can be reflected in the pulling percentage of the optimal arm.

Table 1: Statistics of the Continuous Costs Bandit

| Algorithm | $\bar{r}$ | $\bar{c}$ | $\bar{r}/\bar{c}$ | $\tau$ | %opt |
|---|---|---|---|---|---|
| UCB1 | 0.861 | 0.868 | 0.991 | 11518.7 | 1.01 |
| $\varepsilon_n$-GREEDY | 0.966 | 0.976 | 0.990 | 10248.1 | $< 1$ |
| vUCB-BV1 | 0.743 | 0.310 | 2.396 | 32258.9 | 72.04 |
| Budget-UCB | 0.759 | 0.245 | 3.100 | 40872.4 | 80.03 |

We also test how our proposed Budget-UCB algorithm performs when the costs take discrete values (which is a special case of continuous costs). For this purpose, we change the costs to be discrete by sampling the cost of each pulling of an arm from a Bernoulli distribution. In addition, we change the distribution of the reward of each arm to be Bernoulli too. The parameters of the Bernoulli distributions are also randomly chosen from $(0, 1)$. The results are shown in Figure 1($b$), from which we can see that the Budget-UCB also achieves the lowest empirical regrets among the four algorithms. Similar statistical results like that for continuous cost bandits can be found at Table 2.

Table 2: Statistics of the Discrete Random Costs Bandit

| Algorithm | $\bar{r}$ | $\bar{c}$ | $\bar{r}/\bar{c}$ | $\tau$ | (%)opt |
|---|---|---|---|---|---|
| UCB | 0.789 | 0.786 | 1.027 | 13018.8 | 1.088 |
| $\epsilon_n$-GREEDY | 0.874 | 0.879 | 0.995 | 11374.9 | 0.132 |
| vUCB-BV1 | 0.451 | 0.781 | 1.692 | 21673.6 | 45.714 |
| Budget-UCB | 0.785 | 0.249 | 3.149 | 40101.3 | 79.845 |

Second, we simulate a linear bandit as follows. (1) The set of arms is a ten-dimensional polyhedron in the Euclidean space. (2) The reward/cost of each pulling of an arm is sampled from a truncated Gaussian distribution supported on $[0, 1]$; each dimension of $\mu^r$ and $\mu^c$ (the true reward/cost vector) is sampled from a uniform distribution. (3) For comparison purpose, the CB algorithm (Abbasi-Yadkori et al., 2011) (designed for classical linear bandit problems without budget constraint) is implemented as a baseline. (4) The budget is set

as $500, 1000, 2000, 5000, 10000$. For each value of budget, both algorithms are run for 100 times and their average regrets are examined.

The results of the simulation is summarized in Figure 1($c$).

From Figure 1($c$), we can see that Budget-CB outperforms CB. The reason is that CB tends to pull arms with large empirical rewards and thus most of the budget is not spent on the optimal arm (which has large reward-cost ratio but not necessarily large reward). This verifies the necessity of designing a specific algorithm for the linear bandits with continuous random costs, rather than simply adopting an algorithm designed for other settings.

To sum up, the experimental results show that our algorithms are better than baselines for budgeted bandit problems with continuous random costs.

## 6. Conclusions and Future Work

In this paper, we have studied both budgeted multi-armed bandits and budgeted linear bandits with continuous random costs. We have designed two algorithms and proved their (poly) logarithmic regret bounds with respect to the budget.

For future work, we plan to investigate the following aspects. First, we have assumed that the rewards and costs of different arms are independent of each other in this work. We will investigate the case that the rewards and costs of different arms are correlated. Second, we will investigate how to solve real-world bandit applications with more complex constraints (e.g., with both budget constraint and time constraint). Third, we will explore the distribution-free regret bound.

## References

Yasin Abbasi-Yadkori, Csaba Szepesvári, and David Tax. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320, 2011.

Rajeev Agrawal, MV Hedge, and Demosthenis Teneketzis. Asymptotically efficient adaptive allocation rules for the multiarmed bandit problem with switching cost. *IEEE Transactions on Automatic Control*, 33(10):899–906, 1988.

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.

Jean-Yves Audibert, Sébastien Bubeck, and Rémi Munos. Best arm identification in multi-armed bandits. In *Annual Conference on Learning Theory*, 2010.

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

Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. Bandits with knapsacks. In *IEEE Symposium on Foundations of Computer Science*, pages 207–216. IEEE, 2013.

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

Deepayan Chakrabarti, Ravi Kumar, Filip Radlinski, and Eli Upfal. Mortal multi-armed bandits. In *Advances in Neural Information Processing Systems*, pages 273–280, 2008.

Varsha Dani, Thomas P Hayes, and Sham M Kakade. Stochastic linear optimization under bandit feedback. In *Annual Conference on Learning Theory*, pages 355–366, 2008.

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

Sudipto Guha and Kamesh Munagala. Approximation algorithms for budgeted learning problems. In *Annual ACM Symposium on Theory of Computing*, pages 104–113. ACM, 2007.

András György, Levente Kocsis, Ivett Szabó, and Csaba Szepesvári. Continuous time associative bandit problems. In *International Joint Conference on Artificial Intelligence*, pages 830–835, 2007.

Robert Kleinberg. Nearly tight bounds for the continuum-armed bandit problem. In *Advances in Neural Information Processing Systems*, pages 697–704, 2004.

Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670. ACM, 2010.

Sartaj Sahni. Computationally related problems. *SIAM Journal on Computing*, 3(4):262–279, 1974.

Long Tran-Thanh, Archie Chapman, Munoz De Cote Enrique, Alex Rogers, and Nicholas R. Jennings. Epsilon–first policies for budget–limited multi-armed bandits. In *AAAI Conference on Artificial Intelligence*, pages 1211–1216, 2010.

Long Tran-Thanh, Archie Chapman, Alex Rogers, and Nicholas R. Jennings. Knapsack based optimal policies for budget–limited multi–armed bandits. In *AAAI Conference on Artificial Intelligence*, pages 1134–1140, 2012.

Yizao Wang, Jean-Yves Audibert, and Rémi Munos. Algorithms for infinitely many-armed bandits. In *Advances in Neural Information Processing Systems*, pages 1729–1736, 2009.