Journal of Artificial Intelligence Research 82 (2025) 1943-1959 Submitted 03/2024; published 03/2025

# A New Regret-analysis Framework for Budgeted Multi-Armed Bandits

**Evan Yifan Xu** XYF@SEU.EDU.CN
*Southeast University, No.2 , Sipailou,*
*Nanjing, Jiangsu 210096 CN*

**Pan Xu** PXU@NJIT.EDU
*New Jersey Institute of Technology,*
*323 Dr Martin Luther King Jr Blvd,*
*Newark, NJ 07102 USA*

## Abstract

We consider two versions of the (stochastic) budgeted Multi-Armed Bandit problem. The first one was introduced by Tran-Thanh *et al.* (AAAI, 2012): Pulling each arm incurs a fixed deterministic cost and yields a random reward *i.i.d.* sampled from an unknown distribution (prior free). We have a global budget $B$ and aim to devise a strategy to maximize the expected total reward. The second one was introduced by Ding *et al.* (AAAI, 2013): It has the same setting as before except costs of each arm are *i.i.d.* samples from an unknown distribution (and independent from its rewards). We propose a new *budget-based* regret-analysis framework and design two simple algorithms to illustrate the power of our framework. Our regret bounds for both problems not only match the optimal bound of $O(\ln B)$ but also significantly reduce the dependence on other input parameters (assumed constants), compared with the two studies of Tran-Thanh *et al.* (AAAI, 2012) and Ding *et al.* (AAAI, 2013) where both utilized a *time-based* framework. Extensive experimental results show the effectiveness and computation efficiency of our proposed algorithms and confirm our theoretical predictions.

## 1. Introduction

Multi-Armed Bandit (MAB) models are frequently employed in various real-world scenarios to manage the balance between exploration and exploitation, *e.g.,* crowdsourcing (Singla & Krause, 2013; Singla et al., 2015; Sankararaman et al., 2020; Gao et al., 2020), structured interviewing (Schumann et al., 2019a, 2019b), online advertising (Tran-Thanh et al., 2014; Jaillet et al., 2017), dynamic pricing (Babaioff et al., 2015; Singla et al., 2015) and task offloading in edge computing (Wang et al., 2022). Several variants of the MAB models have been extensively studied in different application domains. Examples include budgeted MAB (Tran-Thanh et al., 2012a; Ding et al., 2013): a global fixed budget constraint is considered during the decision process; bandits with knapsacks (Badanidiyuru et al., 2018a; Immorlica et al., 2019): one or more limited-supply resources are consumed in each round; the sleeping bandits (Kleinberg et al., 2010; Li et al., 2019): arms could sometimes be unavailable; and fair bandits (Li et al., 2019; Patil et al., 2020; Huang et al., 2020; Steiger et al., 2022): fairness constraints are enforced such that each arm is pulled at least a pre-specified fraction of times.

In this paper, we consider two versions of (stochastic) budgeted MAB, namely budgeted MAB with fixed costs and variable costs, respectively. These two models are inspired by a variety of practical applications. For instance, in UAV networks, actions such as video recording or hovering consume energy, thereby limiting the number of actions based on the UAV's battery capacity. This situation aligns with the budgeted MAB model with fixed costs. For another example, in real-time bidding within ad exchanges, the cost of selecting an arm depends on user behavior and competition from other bidders, making it more appropriately modeled as a random variable rather than a fixed value. The following sections provide detailed discussions of the budgeted MAB models with fixed and variable costs.

**Budgeted MAB with Fixed Costs** (BMAB-FC). BMAB-FC is first introduced by (Tran-Thanh et al., 2012a) with the following setting. We have a set $J$ of $K$ arms and a budget $B$. Each arm $j \in J$ has a deterministic cost $c_j$. During every time (or round),[1] pulling an arm $j$ will incur a cost $c_j$ and yield a random reward $V_j \in [0, 1]$, and rewards from pulling $j$ over different time are assumed *i.i.d.* samples from an unknown distribution with (unknown) mean $v_j \in [0, 1]$. The goal is to devise a pulling strategy such that the expected total rewards is maximized subject to the budget constraint. Note that in (Tran-Thanh et al., 2012a), it is assumed that $\lambda \doteq \min_j c_j \geq 1$ and the pulling process should stop whenever the remaining budget is less than $\lambda$.

**Budgeted MAB with Variable Costs** (BMAB-VC). The work of (Ding et al., 2013) proposed BMAB-VC which has a similar setting to BMAB-FC. The key difference is that every time pulling an arm $j$, we see a random cost $C_j \in [0, 1]$ and a random reward $V_j \in [0, 1]$. The random costs and rewards over different time are assumed *i.i.d.* samples from two respective unknown distributions with (unknown) means $c_j \in [0, 1]$ and $v_j \in [0, 1]$. The random costs and rewards are assumed independent from each other over all arms and all rounds. The work of (Ding et al., 2013) considered both settings when $\lambda \doteq \min_j c_j$ is either unknown or known as a prior, while the latter is the focus of this paper. Note that here $\lambda \leq 1$, which differs from the previous case. The pulling process will stop if a pulled arm has a realized cost exceeding the remaining budget.

Throughout this paper, we assume a sufficiently large budget ($B \gg 1$), as also assumed in previous works (Ding et al., 2013; Tran-Thanh et al., 2012a). Specifically, we consider $B > \sum_{j \in J} c_j$ in BMAB-FC and $B > K$ in BMAB-VC, ensuring that the budget $B$ allows for pulling each arm at least once. According to (Lai & Robbins, 1985; Auer, 2002), the regret bounds for BMAB-FC and BMAB-VC are lower bounded by $\Omega(\ln B)$. The algorithms proposed in (Ding et al., 2013; Tran-Thanh et al., 2012a) have been shown to achieve asymptotically optimal bounds of $O(\ln B)$. In this paper, we aim to design simple and effective algorithms for BMAB-FC and BMAB-VC that further improve upon the regret bounds in (Ding et al., 2013; Tran-Thanh et al., 2012a) by reducing the constant factors in front of the dominant $\ln B$ term in the total regret. Detailed discussions can be found in Section 3.

[1] The two terms "time" and "round" are used interchangeably throughout this paper.

Table 1: A glossary of notations.

| | |
|---|---|
| $(a)_+ \doteq \max(a, 0)$ | Max between a generic real number $a$ and 0 |
| $c_j$ | Cost or the cost mean of arm $j$ |
| $v_j$ | Reward mean of arm $j$ |
| $K = \|J\|$ | Number of all arms |
| $\lambda \doteq \min_j c_j$ | Min of the cost mean over all arms |
| $\rho_j = v_j/c_j$ | Density of arm $j$ |
| $j^*(*) = \text{argmax}_j \rho_j$ | Index of arm having the highest density |
| $v_*$ | Reward mean of arm $j^*$ |
| $c_*$ | Cost or the cost mean of arm $j^*$ |
| $d_j = \rho_* - \rho_j$ | Density gap between arm $j$ and an optimal $j^*$ |
| $\alpha_j = c_j \cdot d_j$ | Product of the cost mean and density gap on arm $j$ |
| $d = \min_{j:d_j>0} d_j$ | Min density gap over all non-optimal arms |
| $J_t$ | Set of arms with costs within the remaining budget at time $t$ |

## 2. Main Techniques

For each $j \in J$, let $\rho_j \doteq v_j/c_j$ be the density of arm $j$. Let $j^* = \text{argmax}_j \rho_j$ and we write $*$ short for $j^*$ when the context is clear. Throughout this paper, we use OPT and $\mathbb{E}[\text{OPT}]$ to denote the optimal strategy and the corresponding expected rewards achieved, respectively. The same for ALG (a generic algorithm). The below lemma gives an upper bound for OPT.

**Lemma 1.** *For* BMAB-FC*, we have* $\text{OPT} \leq B\rho_*$*; and for* BMAB-VC*, we have* $\text{OPT} \leq (B+1)\rho_*$.

The first claim is explicitly stated in the proof of Theorem 1 on page 15 of the full version (Tran-Thanh et al., 2012b), while the second claim is proved in Lemma 1 of the work (Ding et al., 2013).[2]

**A Time-based Analysis Framework**. Both of studies (Ding et al., 2013; Tran-Thanh et al., 2012a) apply a time-based framework to analyze the regret. Consider a given algorithm ALG. Let $T$ be the random number of rounds pulled by ALG and $j(t)$ be the choice of ALG at time $t$. They tried to decompose the total regret $R(\text{ALG}) = \mathbb{E}[\text{OPT}] - \mathbb{E}[\text{ALG}]$ as follows.

$$R(\text{ALG}) \leq B\rho_* - \sum_{t=1}^{T} \mathbb{E}[v_{j(t)}] \tag{1}$$

$$= \underbrace{\mathbb{E}_T[B\rho_* - Tv_*]}_{R^a(\text{ALG})} + \underbrace{\mathbb{E}_T\left[Tv_* - \sum_{t=1}^{T} \mathbb{E}[v_{j(t)}]\right]}_{R^b(\text{ALG})}. \tag{2}$$

[2] Note that for BMAB-VC, the upper bound of $(B+1)\rho_*$ is asymptotically tight. Consider a simple case of BMAB-VC where the budget is an integer $B \in \mathbb{Z}_+$, and there is a single arm with a Bernoulli random cost having a mean of $\epsilon > 0$ and a deterministic reward of $\epsilon$, with $\rho_* = 1$. In this scenario, any OPT will continue pulling the arm an expected $(B+1)/\epsilon - 1$ times, resulting in an expected reward of $B + 1 - \epsilon$.

Under the premise that OPT will always play the optimal arm $j^*$, we can interpret the two parts as follows: $R^a(\text{ALG})$ is the expected total regret over those missing rounds played in OPT but not in ALG; $R^b(\text{ALG})$ is the expected regret over rounds when ALG played on non-optimal arms.

**A Budget-based Analysis Framework**. In contrast, we propose a budget-based framework to analyze the regret. Consider a given algorithm ALG. Let $B(\text{ALG})$ be the random number of budgets used by ALG. For each $j \in J$, let $N_j$ be the random number of pulls on arm $j$, and $V_j, C_j$ be the corresponding total (random) rewards and costs involved on $j$ in ALG. Under the larget budget assumption ($B \gg 1$), we can show that

$$\mathbb{E}[V_j] = \mathbb{E}[N_j] \cdot v_j, \mathbb{E}[C_j] = \mathbb{E}[N_j] \cdot c_j. \tag{3}$$

For any $j$, let $d_j \doteq \rho_* - \rho_j$ be the density gap between $j$ and $j^*$. We try to decompose the total regret as follows. Consider, for example, the case of BMAB-FC.

$$\begin{aligned}
&\mathbb{E}[\text{OPT}] - \mathbb{E}[\text{ALG}] && (4)\\
&\le \mathbb{E}\big[B\rho_* - B(\text{ALG})\rho_* + B(\text{ALG})\rho_* - \sum_j V_j\big] && (5)\\
&= \mathbb{E}\big[B\rho_* - B(\text{ALG})\rho_* + \sum_j C_j\rho_* - \sum_j V_j\big] && (6)\\
&= \mathbb{E}[B\rho_* - B(\text{ALG})\rho_*] + \sum_j \mathbb{E}[N_j](c_j \cdot \rho_* - v_j) && (7)\\
&= \underbrace{\mathbb{E}[B\rho_* - B(\text{ALG})\rho_*]}_{R^a(\text{ALG})} + \underbrace{\sum_{j:d_j>0} \mathbb{E}[N_j] \cdot (c_j \cdot d_j)}_{R^b(\text{ALG})}. && (8)
\end{aligned}$$

Here we can interpret $R^a(\text{ALG})$ as the expected regret over those budgets wasted in ALG, and $R^b(\text{ALG})$ as the expected regret over budgets misused on non-optimal arms.

Under the time-based analysis framework, the analyses of $R^a(\text{ALG})$ and $R^b(\text{ALG})$ are equally technical. In contrast, in the budget-based framework, only the analysis of $R^b(\text{ALG})$ presents a significant challenge, while that of $R^a(\text{ALG})$ is relatively straightforward. This shift allows us to focus on upper bounding $R^b(\text{ALG})$—the total expected regret arising from misallocated budgets on suboptimal arms. As a result, the budget-based approach nearly halves the analytical effort compared to the time-based approach and reduces potential regret gaps between the target algorithm and the optimal one. This simplifies the regret analysis while delivering improved regret bounds.

## 3. Main Contributions

We acknowledge that the budget-based analysis framework has already been used *inexplicitly* in several prior studies, *e.g.,* (Xia et al., 2015; Rangi et al., 2019a), though none of them states it in a formal way. Part of our main contributions lies in that we formalize the budget-based framework for a general setting of budgeted MAB. In addition, we propose two simple UCB-based algorithms to exemplify the power of the framework. Specifically,

we propose a UCB-based algorithm UF for BMAB-FC and a UCB-and-LCB-based algorithm ULV for BMAB-VC (Section 4). By using the unifying budget-based framework, we show that the two algorithms achieve a regret bound of $O(\ln B)$ for BMAB-FC and BMAB-VC, respectively (Sections 5 and 6). Both of our algorithms and the analyses are featured by their simplicity and generality. Moreover, our regret bounds significantly reduce the dependence on input parameters as shown in the two previous works (Tran-Thanh et al., 2012a) and (Ding et al., 2013), where both utilized the time-based analysis framework. Our improvement in the regret bound highlights the superiority of the budget-based analysis framework over the time-based version. Table 1 gives a glossary of notations used throughout this paper. For a generic real number $a$, let $(a)_+ \doteq \max(a, 0)$.

### 3.1 Regret Bounds for BMAB-FC

Recall that in BMAB-FC, (a) each arm $j$ has a deterministic cost $c_j$ and $\lambda = \min_j c_j \geq 1$; (b) $d_j = \rho_* - \rho_j < \rho_* \leq 1$.

**Theorem 1.** *For* BMAB-FC*, the algorithm* UF *achieves a regret bound at most*

$$\sum_{j:d_j>0} \Big(\frac{8\ln B}{c_j d_j} + c_j + 1\Big) + 1 + \frac{2}{B}.$$

**Comparison with the Work (Tran-Thanh et al., 2012a)**. Our setting of BMAB-FC is the same as (Tran-Thanh et al., 2012a), which gave an algorithm[3] achieving a regret bound of $\ln B \cdot H$, where $H \doteq \frac{8}{d^2}\sum_j \Big(\frac{(c_j - c_*)_+}{c_*} + (v_* - v_j)_+\Big)$ with $d = \min_{j:d_j>0} d_j$ being the minimum non-zero density gap. Here we focus only on the *dominant part involving* $\ln B$ *and skip the rest*. Note that $H = \Omega(K/d^2)$ since for each $j \neq j^*$, either $c_j > c_*$ or $v_* > v_j$. As shown in Theorem 1, our coefficient of $\ln B$ is $H' \doteq \sum_{j:d_j>0} \frac{8}{c_j d_j} \leq \frac{8K}{d}$ since $c_j \geq 1$. By comparing $H$ and $H'$, we see that our regret reduces the dependence on $d$ from $d^{-2}$ to $d^{-1}$. What's more, we completely remove the dependence on gaps of cost and reward means between optimal and non-optimal arms in the dominant part termed by $H$. Note that the gap of cost means, $\sum_j \frac{(c_j - c_*)_+}{c_*}$, can be arbitrarily large, since no upper bound is assumed on $c_j$.

### 3.2 Regret Bounds for BMAB-VC

Recall that in BMAB-VC, each arm $j$ has a random cost with an unknown mean $c_j$ such that $\lambda \leq c_j \leq 1$. Let $\alpha_j \doteq c_j \cdot d_j$ for all $j \in J$.

**Theorem 2.** *For* BMAB-VC*, the algorithm* ULV *achieves a regret bound at most*

$$\sum_{j:d_j>0} \Big(8\ln B \cdot \frac{(2+\alpha_j)^2}{\alpha_j c_j^2} + 1\Big) + \Big(2 + \frac{4}{B\lambda}\Big) \cdot \rho_*.$$

**Comparison with the Work (Ding et al., 2013)**. The work of (Ding et al., 2013) considered BMAB-VC under the same setting with us. They gave an algorithm achieving

[3] Actually, the work of (Tran-Thanh et al., 2012a) gave two algorithms but here we select the one with a better regret bound.

a regret bound of $\ln B \cdot (H_1 + H_2)$, where $H_1 \doteq \lambda^{-2} \sum_{j:d_j>0} \big(1 + 2/d_j + 2/(\lambda d_j)\big)^2 \rho_*$ and $H_2 \doteq \lambda^{-2} \sum_{j:d_j>0} \big(1 + 2/d_j + 2/(\lambda d_j)\big)^2 \cdot \max(v_* - v_j, 0)$. Again, we focus only on the dominant part involving $\ln B$. Theorem 2 suggests that our dominant term is $\ln B \cdot H'$, where $H' \doteq \sum_{j:d_j>0} 8(2+\alpha_j)^2/(\alpha_j c_j^2)$. Lemma 2 implies our leading coefficient $H'$ is at most a constant factor of 8 of $H_1$, and in the special case when all non-optimal arm has a density gap $\epsilon$ from the optimal, our coefficient $H'$ improves the dependence of $H_1$ on $\epsilon$ and $\lambda$ from $\epsilon^{-2}\lambda^{-4}$ to $\epsilon^{-1}\lambda^{-3}$. Furthermore, our regret removes the direct dependence on the gaps of reward means between optimal and non-optimal arms as $H_2$ did.

**Lemma 2.** *(1) $H' \leq 8H_1$. (2) Suppose $d_j = \epsilon \ll 1$ for all $d_j > 0$. We have $H' = O(\frac{K}{\epsilon\lambda^3})$ and $H_1 = \Omega(\frac{K}{\epsilon^2\lambda^4})$.*

*Proof.* We show (1) first.

$$H_1 = \lambda^{-2} \sum_{j:d_j>0} \big(1 + 2/d_j + 2/(\lambda d_j)\big)^2 \rho_* \tag{9}$$

$$> \lambda^{-2} \sum_{j:d_j>0} \big(\rho_* + \frac{1}{\lambda}\frac{\rho_*}{d_j}\frac{4}{\lambda d_j} + \frac{4\rho^*}{\lambda d_j}\big) \tag{10}$$

$$> \sum_{j:d_j>0} \frac{1}{c_j^2}\big(\alpha_j + \frac{4}{\alpha_j\lambda} + \frac{4}{\lambda}\big) \tag{11}$$

$$> \sum_{j:d_j>0} c_j^{-2}(\alpha_j + 4/\alpha_j + 4) = H'/8. \tag{12}$$

Inequality (11) is due to facts that $\lambda \leq c_j$, $\rho_* > \rho_* - \rho_j = d_j \geq c_j d_j = \alpha_j$, $\lambda d_j \leq c_j d_j = \alpha_j$. Inequality (12) is due to $\lambda \leq 1$.

Now we show part (2). When $d_j = \epsilon$ for all $d_j > 0$, we see $H' \leq \sum_{j:d_j>0} 8(2+\epsilon)^2/(\epsilon\lambda^3) = O(K/(\epsilon\lambda^3))$, since $\alpha_j = c_j d_j \leq \epsilon$ and $\alpha_j \geq \lambda\epsilon$. In contrast, $H_1 \geq \lambda^{-2} \sum_{j:d_j>0} 4\rho_*/(\epsilon^2\lambda^2) = \Omega(K/(\epsilon^2\lambda^4))$. □

## 4. Main Algorithms

Here are a few notations used in our algorithms. Consider a given time $t$. Let $N_{j,t}$ be the (random) number of pulls on arm $j$ before $t$, $V_{j,t}$ be the empirical estimate (sample average) of rewards on $j$ by $t$ and for BMAB-VC only, let $C_{j,t}$ be the empirical estimate of cost on $j$ by $t$. Define the upper and lower confidence bounds of rewards as $\mathsf{UV}_{j,t} = V_{j,t} + \sqrt{2\ln B/N_{j,t}}$ and $\mathsf{LV}_{j,t} = V_{j,t} - \sqrt{2\ln B/N_{j,t}}$. Similarly, for random costs in BMAB-VC, we define $\mathsf{UC}_{j,t} = C_{j,t} + \sqrt{2\ln B/N_{j,t}}$ and $\mathsf{LC}_{j,t} = C_{j,t} - \sqrt{2\ln B/N_{j,t}}$.[4]

Our main algorithms, UF and ULV, are formally stated as below.

**Remarks**. (1) Note that under the large-budget assumption, we have $B > \sum_{j\in J} c_j$ in BMAB-FC and $B > K$ in BMAB-VC. In other words, budget $B$ will afford to pull each arm

[4]The expressions for the upper and lower confidence bounds on rewards and costs follow the approach outlined by (Tran-Thanh et al., 2012a); see expression (7) in that paper, which defines the upper confidence bound on an arm' density. This is a common practice in regret-bound analysis.

**Algorithm 1:** A UCB-based Algorithm for BMAB-FC (UF).

1 **Initial.**: Pull each arm once during the first $K$ steps;
2 **While** $J_t \neq \emptyset$ **do**: Pull the arm $j(t) = \text{argmax}_{j \in J_t} \mathsf{UV}_{j,t}/c_j$, where $J_t$ is the set of arms whose costs are no larger than the current remaining budget at $t$.

**Algorithm 2:** A UCB-and-LCB-based Algorithm for BMAB-VC (ULV).

1 **Initial.**: Pull each arm once during the first $K$ steps;
2 **While the current remaining budget is affordable to pull** $j(t)$ **do**: Pull the arm $j(t) = \text{argmax}_j \mathsf{UV}_{j,t}/\max(\mathsf{LC}_{j,t}, \lambda)$.

at least once surely. (2) Now, we justify Equation (3) holds for both of the above algorithms. Consider a given $j$ in UF. Recall that $N_j$ and $V_j$ be the respective total (random) numbers of pulls and rewards on $j$. Observe that $V_j$ is a sum of $N_j$ *i.i.d.* random samples each with mean $v_j$. What's more, $N_j$ qualifies as a *stopping* time according to (Chewi, 2020): for a given set of cost realizations over all other arms $j' \neq j$ and over all time, we claim that the event $(N_j \leq N)$ occurs or not is completely determined by the first $N$ cost samples on $j$. Thus, by applying Wald's equation, we get $\mathbb{E}[V_j] = \mathbb{E}[N_j] \cdot v_j$. We can argue in the same way for ULV.

## 5. Regret Analysis for BMAB-FC

We prove the main Theorem 1 here. Throughout this paper, we denote $[n] = \{1, 2, \ldots, n\}$ for a generic integer $n$. Observe that UF will surely terminate after at most $B$ rounds since $\lambda = \min_j c_j \geq 1$. This suggests that $N_{j,t} \leq B$ with probability one for all $j$ and $t$. We define the clean event $\mathsf{CE}$ as follows: for all $j \in J$ and all possible integers $N_{j,t} \in [B]$, $v_j \in (\mathsf{LV}_{j,t}, \mathsf{UV}_{j,t})$.

**Lemma 3.** $\Pr[\neg\mathsf{CE}] \leq 2B^{-2}$.

*Proof.* Consider a given $j \in J$ and $N_{j,t} \in [B]$. Observe that

$$\Pr[v_j \notin (\mathsf{LV}_{j,t}, \mathsf{UV}_{j,t})] \tag{13}$$

$$= \Pr\left[|V_{j,t} - v_j| \geq \sqrt{2 \ln B / N_{j,t}}\right] \leq 2B^{-4}. \tag{14}$$

Inequality (14) is due to Hoeffding's inequality. By taking union bound over all possible $j \in J$ and $N_{j,t} \in [B]$, we have $\Pr[\neg\mathsf{CE}] \leq K \cdot B \cdot (2B^{-4}) < 2B^{-2}$. Note that by our inexplicitly assumption in BMAB-FC, we have $B > \sum_j c_j \geq K\lambda \geq K$. Thus, we establish our claim. □

Now assume $\mathsf{CE}$ happens and we apply the budget-based framework to analyze the regret of UF. Let $\mathbb{E}[R^a(\text{UF})|\mathsf{CE}]$ and $\mathbb{E}[R^b(\text{UF})|\mathsf{CE}]$ be the first and second parts of regret of UF under $\mathsf{CE}$ shown in (8).

**Lemma 4.** $\mathbb{E}[R^a(\text{UF})|\mathsf{CE}] < \lambda \cdot \rho_* \leq 1$.

*Proof.* Let $B(\text{UF})$ be budget used in UF. By the stopping rule of UF, $B - B(\text{UF}) < \lambda$. Thus, $R^a(\text{UF}) = (B - B(\text{UF}))\rho_* < \lambda\rho_* \leq v_* \leq 1$. □

**Lemma 5.** $\mathbb{E}[R^b(\text{UF})|\mathsf{CE}] \leq \sum_{j:d_j>0} \left(\frac{8\ln B}{c_j d_j} + c_j + 1\right)$.

*Proof.* All over this proof, we assume $\mathsf{CE}$ occurs by default. Consider a given $j$ with $d_j > 0$ and a given optimal arm $j^*$[5]. Let $N_j$ be the total number of pulls on $j$ and decompose $N_j = N_j^a + N_j^b$, where $N_j^a$ be the number of pulls on $j$ during time $t$ with $J_t \ni j^*$ and $N_j^b$ be that during the time $t$ when $j^* \notin J_t$. Observe that $j^* \notin J_t$ suggests that the remaining budget is less than $c_*$ and thus, $N_j^b < c_*/c_j$. Now we focus on analyzing $N_j^a$. Let $T$ be the index of the last time we pull $j$ with $B_T \geq c_*$. Since $J_t \ni j^*$, we have that

$$\frac{v_*}{c_*} < \frac{\mathsf{UV}_{*,T}}{c_*} \leq \frac{\mathsf{UV}_{j,T}}{c_j} < \frac{v_j + 2\sqrt{2\ln B/N_{j,T}}}{c_j}. \tag{15}$$

The first and third inequalities are due to the clean event $\mathsf{CE}$; the second one follows from the rule of UF. Note that $d_j = v_*/c_* - v_j/c_j$. Subtracting the term $v_j/c_j$ from both sides of Inequality (15), we get $N_{j,T} < 8\ln B/(c_j d_j)^2$. Thus,

$$\begin{aligned} N_j \cdot (c_j d_j) &= (N_j^a + N_j^b)(c_j d_j) \\ &\leq (N_{j,T} + 1 + c_*/c_j)(c_j d_j) \\ &< 8\ln B/(c_j d_j) + c_j d_j + c_* d_j \\ &\leq 8\ln B/(c_j d_j) + c_j + 1. \end{aligned}$$

The last inequality above is due to facts $d_j < \rho_* \leq 1$ and $c_* d_j < c_*\rho_* = v_* \leq 1$. By the definition of $R^b(\text{UF})$, we establish our claim. □

*Proof of the Main Theorem 1.* Let $R(\text{UF})$ be the total expected regret. We have

$$\begin{aligned} R(\text{UF}) &\leq \mathbb{E}[R(\text{UF})|\mathsf{CE}] + \mathbb{E}[R(\text{UF})|\neg\mathsf{CE}]\Pr[\neg\mathsf{CE}] \\ &\leq \mathbb{E}[R^a(\text{UF})|\mathsf{CE}] + \mathbb{E}[R^b(\text{UF})|\mathsf{CE}] + (B\rho_*)(2B^{-2}) \quad \text{(due to Lemma 1)} \\ &\leq \textstyle\sum_{j:d_j>0} \left(8\ln B/(c_j d_j) + c_j + 1\right) + 1 + 1/(2B). \end{aligned}$$

□

## 6. Regret Analysis for BMAB-VC

We prove the main Theorem 2 here. The whole proofs are essentially the same as those shown in Section 5. Let $T$ be the random number of rounds played by ULV before termination. Recall that $\lambda \leq c_j \leq 1$ for all $j$. Thus, we have $\mathbb{E}[T] \leq B/\lambda$. Consider a given $T$ and we define the conditional clean event $(\mathsf{CE}|T)$ as follows: for all $j \in J$ and all possible integers $N_{j,t} \in [T]$, $v_j \in (\mathsf{LV}_{j,t}, \mathsf{UV}_{j,t})$ and $c_j \in (\mathsf{LC}_{j,t}, \mathsf{UC}_{j,t})$.

**Lemma 6.** $\Pr[\neg\mathsf{CE}|T] \leq 4 \cdot T \cdot B^{-3}$.

[5] There might be multiple optimal arms, and in that case we break ties arbitrarily.

*Proof.* Consider a given $j \in J$ and $N_{j,t} \in [T]$. By applying Hoeffding's inequality, we have $\Pr[v_j \notin (\mathsf{LV}_{j,t}, \mathsf{UV}_{j,t})] = \Pr\left[|V_{j,t} - v_j| \geq \sqrt{2\ln B/N_{j,t}}\right] \leq 2B^{-4}$. Similarly, we have $\Pr[c_j \notin (\mathsf{LC}_{j,t}, \mathsf{UC}_{j,t})] \leq 2B^{-4}$. Applying union bounds over all possible $j \in J$ and $N_{j,t} \in [T]$, we have $\Pr[\neg\mathsf{CE}|T] \leq K \cdot T \cdot (4B^{-4}) < 4TB^{-3}$. □

Consider a given $T$ and assume $\mathsf{CE}$ happens. Now we apply the budget-based framework to analyze the regret of ULV. Let $\mathbb{E}[R^b(\mathrm{UF})|\mathsf{CE}, T]$ be the second part of regret of ULV shown in (8) under a given $T$ and $\mathsf{CE}$. Recall that $\alpha_j = c_j d_j$ for every $j$.

**Lemma 7.**
$$\mathbb{E}[R^b(\mathrm{ULV})|\mathsf{CE}, T] \leq \textstyle\sum_{j:d_j>0} 8\ln B \cdot (2+\alpha_j)^2/(\alpha_j c_j^2).$$

*Proof.* Consider a given $j$ with $d_j > 0$. Let $N_{j,T}$ be the number of rounds we play $j$, and let $t \leq T$ be the last round we play on $j$. Thus, we have that $\mathsf{UV}_{j,t}/\max(\mathsf{LC}_{j,t}, \lambda) \geq \mathsf{UV}_{*,t}/\max(\mathsf{LC}_{*,t}, \lambda)$. Consider **Case 1** when $c_j - 2\sqrt{2\ln B/N_{j,t}} \leq 0$. This suggests that $N_{j,t} \leq 8\ln B/c_j^2$. Consider **Case 2** when $c_j - 2\sqrt{2\ln B/N_{j,t}} > 0$. Under $\mathsf{CE}$, we have

$$\frac{v_j + 2\sqrt{2\ln B/N_{j,t}}}{c_j - 2\sqrt{2\ln B/N_{j,t}}} > \frac{\mathsf{UV}_{j,t}}{\max(\mathsf{LC}_{j,t}, \lambda)} \tag{16}$$

$$\geq \frac{\mathsf{UV}_{*,t}}{\max(\mathsf{LC}_{*,t}, \lambda)} > \frac{v_*}{c_*}. \tag{17}$$

Inequality (16) is partially due to $c_j - 2\sqrt{2\ln B/N_{j,t}} < \mathsf{LC}_{j,t}$ under $\mathsf{CE}$; The second inequality on (17) is partially due to $c_* \geq \lambda$ and $c_* > \mathsf{LC}_{*,t}$ under $\mathsf{CE}$. Let $\epsilon \doteq 2\sqrt{2\ln B/N_{j,t}}$. Observe that

$$\frac{v_j+\epsilon}{c_j-\epsilon} > \frac{v_*}{c_*} \Rightarrow \frac{v_j+\epsilon}{c_j-\epsilon} - \frac{v_j}{c_j} > \frac{v_*}{c_*} - \frac{v_j}{c_j} = d_j$$

$$\Rightarrow \epsilon > \frac{d_j c_j(c_j-\epsilon)}{v_j + c_j} \geq d_j c_j(c_j-\epsilon)/2 \Rightarrow \epsilon > \frac{d_j c_j^2}{2 + d_j c_j}.$$

Substituting $\epsilon = 2\sqrt{2\ln B/N_{j,t}}$ to the above inequality, we have $N_{j,t} \leq 8\ln B \cdot (2+\alpha_j)^2/(\alpha_j c_j)^2$, where $\alpha_j = d_j \cdot c_j$. Observe that $8\ln B \cdot (2+\alpha_j)^2/(\alpha_j c_j)^2 = (8\ln B/c_j^2) \cdot \left(1 + 2/\alpha_j\right)^2 > 8\ln B/c_j^2$, where the latter is the upper bound for $N_{j,t}$ in **Case 1**. Summarizing the analysis for the two cases, we have

$$\begin{aligned}\mathbb{E}[N_{j,T}|\mathsf{CE}, T] &= \mathbb{E}[N_{j,t} + 1|\mathsf{CE}, T] \\ &\leq 8\ln B \cdot (2+\alpha_j)^2/(\alpha_j c_j)^2 + 1.\end{aligned}$$

Thus, we have

$$\begin{aligned}\mathbb{E}[R^b(\mathrm{ULV})|\mathsf{CE}, T] &= \sum_{j:d_j>0} \mathbb{E}[N_{j,T} \cdot c_j \cdot d_j|\mathsf{CE}, T] \\ &\leq \sum_{j:d_j>0} \Big(8\ln B \cdot (2+\alpha_j)^2/(\alpha_j c_j^2) + 1\Big).\end{aligned}$$

□

*Proof of the Main Theorem 2.* We analyze the regret of ULV for BMAB-VC. Let $R$ be the total expected regret of ULV. Following the budget-based framework, we have $R = R^a + R^b$. Observe that ULV will end up with at most 1 unit budget unused, which leads to $R^a \leq 2\rho_*$ (in which case the upper bound of OPT is replaced by $\rho^* + B\rho^*$). Now we focus on the second part $R^b$.

$$R^b \leq \mathbb{E}_T[R^b|\mathsf{CE}, T] + \mathbb{E}_T\big[(R^b|\neg\mathsf{CE}, T)\Pr[\neg\mathsf{CE}|T]\big] \tag{18}$$

$$\leq \mathbb{E}_T[R^b|\mathsf{CE}, T] + \mathbb{E}_T[B \cdot \rho_* \cdot (4B^{-3}T)] \tag{19}$$

$$\leq \sum_{j:d_j>0} \Big(8\ln B \cdot (2+\alpha_j)^2/(\alpha_j c_j^2) + 1\Big) + 4\rho_*/(B\lambda) \tag{20}$$

Inequality (19) is due to Lemma 6. Summarizing all above analysis yields our result. □

## 7. Experiments

In this section, we describe our experimental results on both BMAB-FC setting and BMAB-VC setting. In BMAB-FC setting, we test UF against KUBE and fractional KUBE (Tran-Thanh et al., 2012a); in BMAB-VC setting, we test ULV against UCB-BV1 (Ding et al., 2013).[6] The main questions we target to answer in this section are as follows:

- **Q1**: Does the proposed budget-based analysis framework yield a tighter theoretical upper bound?
- **Q2**: Do our algorithms demonstrate superior computational efficiency compared to the baseline methods?

### 7.1 Experiments for BMAB-FC

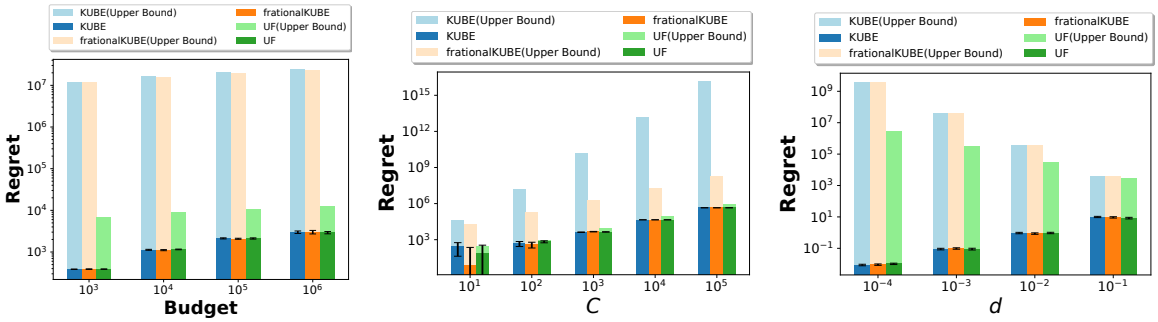
For BMAB-FC setting, we run three kinds of experiments by varying the total budget $B$, the pulling cost for each arm $c_j$, and the minimum density gap over all non-optimal arms $d$. The details of the experimental setup are as follows.

**Experimental Setup**. In the first experiment, we study the gap between the actual regrets and regret bounds by following the parameter settings in (Tran-Thanh et al., 2012a). We set the number of arms $K = 100$. We generate the fixed cost $c_j$ for each arm $j$ by selecting a uniform value from $[1, 10]$. The reward distribution of each arm $j$ is set to be truncated Gaussian, with mean reward $v_j$ takes a uniform value from $[0, 1]$ and variance $\sigma^2 = v_j/2$. In addition, we vary the total budgets $B$ from $10^3$ to $10^6$. In the second experiment, we investigate how the regret bounds respond to the gap of cost means, $\sum_j (c_j - c_*)_+$, while fixing the minimum density gap $d$ and the gap of reward means, $\sum_j (v_* - v_j)_+$. We set $K = 10$ and the total budget $B = 10^6$. We introduce a parameter $C$ and choose $C \in \{10^1, 10^2, 10^3, 10^4, 10^5\}$. To fix $d = 0.5$, we first set the mean reward $v_j$ for each arm $j$ as 1, then set the cost for the first arm $c_1 = 1/(d + 1/C)$; set the cost for the rest arms to $c_{j,j\neq 1} = C$. In the third experiment, we examine how the regret bounds respond to the

[6] All experiments are conducted on a PC with Quad-Core Intel Core i7 (2GHz) processor and 8GB memory.

Table 2: Parameter settings for BMAB-FC, where $K$ denotes the number of arms, $B$ denotes the total budget, $C$ is the parameter to adjust the gap of cost means, $d$ denotes the minimum density gap, and $\kappa$ is set to adjust the density gap.

| No. of Experiment | $K$ | Cost Setting (Fixed) | Reward Distribution | $B$ | $C$ | $d$ | $\kappa$ |
|---|---|---|---|---|---|---|---|
| 1 | 100 | Uniform value from $[0, 1]$ | Truncated Gaussian | $\{10^3, 10^4, 10^5, 10^6\}$ | - | - | - |
| 2 | 10 | $c_1 = 1/(d + 1/C)$ $c_{j,j\neq 1} = C$ | | $10^6$ | $\{10^1, 10^2, 10^3, 10^4, 10^5\}$ | 0.5 | - |
| 3 | 10 | $c_1 = 1$ $c_{j,j\neq 1} = c_1/(1-\kappa)$ | | 100 | - | $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$ | 0.1 |




(a) The actual regrets incurred and upper bounds when varying the total budget $B$.

(b) Results of varying the gap of cost means, fixing the minimum density gap and the gap of reward means.

(c) Results of varying the minimum density gap $d$, fixing the gap of cost means and reward means.

Figure 1: **Our method UF enhances the precision of the upper bound without compromising regret in BMAB-FC setting.** Comparison of the actually regrets incurred, and regret bounds of UF against KUBE and fractional KUBE.

parameter of the minimum density gap $d$, while fixing the other parameters. We set $K = 10$ and $B = 100$, and vary $d \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$ by first fixing the cost means for all arms as follows: for arm $j = 1$, we set $c_1 = 1$; for all arms $j, j \neq 1$, we set $c_{j,j\neq 1} = c_1/(1-\kappa)$, where $\kappa$ is set to 0.1 in our experiments. We set a uniform reward means for all arms as $v_j = (c_1/\kappa) \cdot d$. For each instance, we run all algorithms for 100 times and take the average as the final performance. The detailed setting is summarized in Table 2.

**Results and Discussions**. Figures 1 show the effectiveness and computation efficiency of UF and validate our theoretical analyses. From Figure 1a, we see that the upper bound of UF always dominates the actual performance of UF. This confirms our theoretical predictions in Theorem 1. What's more, our upper bound for UF significantly improves those for KUBE algorithms from around $10^7$ to $10^4$. This is consistent with our analyses in Section 3.1. Figure 1a also suggests that for the actually regrets incurred, the three algo-

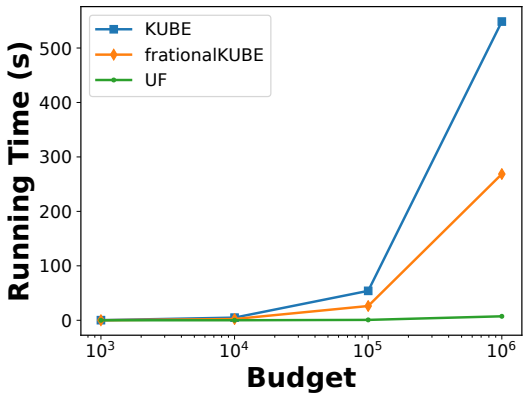



Figure 2: **Our algorithm** UF **achieves superior computational efficiency compared to baseline methods in** BMAB-FC **setting.** Running time achieved when varying the total budget $B$.

rithms perform comparably. When it comes to the running time, however, UF performs substantially better than fractional KUBE and KUBE, as shown in Figure 2. Note that the error bars in Figure 1a represent the 95% confidence intervals for the corresponding actual regrets.

Figure 1b shows the tightness of our upper bound with respect to the parameter of gaps in cost means. When the gap of cost means ($C$) increases, the actually regrets incurred by the three algorithms and our upper bound all increases in a minor way, compared with upper bounds of KUBE and fractional KUBE. In particular, the upper bound for KUBE increases most dramatically. The tiny gap between the lower bound (actually regrets) and upper bound of UF suggests the tightness in our regret analysis with respect to the parameter of gaps in cost means. There seems little space for improvement on that part.

From Figure 1c, we see that when the minimum density gap $d$ decreases from $10^{-1}$ to $10^{-4}$, the actually regrets incurred by the three algorithms all get reduced instead of increased, as predicted by the corresponding upper bounds. In spite of that, our upper bound for UF increases in the most insignificant way compared with the rest two. This is due to that our upper bound reduces the dependence on $d$ from $d^{-2}$ to $d^{-1}$. Current results of Figure 1c suggests the possibility of further improvement over the dependence of $d$.

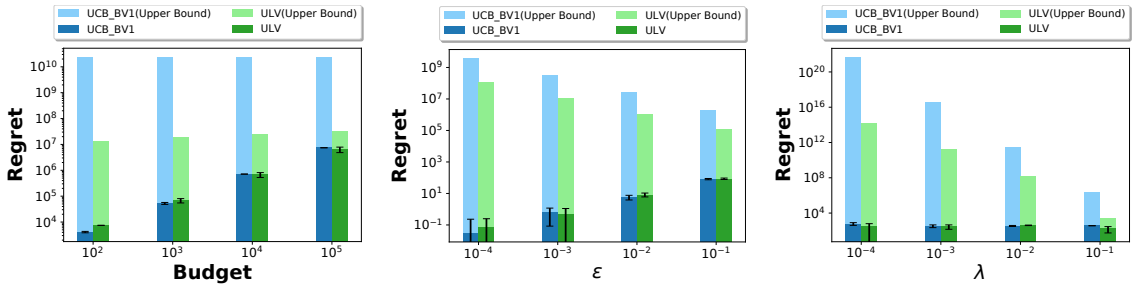
### 7.2 Experiments for BMAB-VC

For BMAB-VC setting, we run three kinds of experiments by varying the total budget $B$, the minimum density gap over all non-optimal arms $\epsilon$, and the minimum cost mean over all arms $\lambda$. The details of the experimental setup are as follows.

**Experimental Setup**. In the first experiment, we study the gap between the actual regrets and regret bounds by following the parameter settings in (Ding et al., 2013). We set the number of arms $K = 100$. We generate the mean cost $c_j$ for each arm $j$ by selecting a uniform value from $\{0.01, 0.02, ..., 0.99, 1\}$, and set the cost distribution type as truncated Gaussian, with mean of $c_j$ and variance $\sigma^2 = c_j/2$. The reward distribution of each arm $j$ is set to be Bernoulli, with mean reward $v_j$ takes a uniform value from $[0, 1]$. We vary the

Table 3: Parameter settings for BMAB-VC, where $K$ denotes the number of arms, $B$ denotes the total budget, $\epsilon$ denotes the minimum density gap, $\lambda$ denotes the minimum cost mean, and $\kappa$ is the parameter to adjust the density gap.

| No. of Experiment | $K$ | Cost Distribution | Reward Distribution | $B$ | $\epsilon$ | $\lambda$ | $\kappa$ |
|---|---|---|---|---|---|---|---|
| 1 | 100 | Truncated Gaussian | Bernoulli | $\{10^2, 10^3, 10^4, 10^5\}$ | - | - | - |
| 2 | 10 | | | 1000 | $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$ | - | - |
| 3 | 10 | | | 100 | 4 | $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$ | 0.1 |




(a) The actual regrets incurred when varying the total budget $B$.

(b) Results of varying the minimum density gap $\epsilon$, fixing the gap of cost means and reward means.

(c) Results of varying the minimum cost means $\lambda$, fixing the minimum density gap and the gap of reward means.

Figure 3: **Our method** ULV **enhances the precision of the upper bound without compromising regret in** BMAB-VC **setting.** Comparison of the actually regrets incurred and regret bounds of ULV against UCB-BV1.

total budgets $B$ from $10^2$ to $10^5$. In the second experiment, we investigate how our regret bound respond to different minimum density gap $\epsilon$, while fixing the gap of reward means, $\sum_j (v_* - v_j)_+$ and the gap of cost means, $\sum_j (c_j - c_*)_+$. We set the number of arms $K = 10$ and the total budget $B = 1000$. We vary $\epsilon \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$ by first fixing the cost means for all arms as follows: for arm $j = 1$, we set $c_1 = 0.5$; for all arms $j, j \neq 1$, we set $c_{j,j\neq 1} = c_1/(1-\kappa)$, where $\kappa = 0.1$. Then we set a uniform reward means for all arms as $v_j = (c_1/\kappa) \cdot \epsilon$. In the third experiment, we examine how our regret bound respond to the parameter of the minimum cost means over all arms $\lambda$, while fixing the rest. We set $K = 10$ and $B = 100$. We vary $\lambda \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$ by first fixing the cost means for all arms with $v_j = 0.5$. To fix $\epsilon = 4$, we set the mean cost of first arm as $c_1 = \lambda$; set the cost of rest arms as $c_{j,j\neq 1} = v_j/(v_j/\lambda - \epsilon)$. The detailed setting is summarized in Table 3.

**Results and Discussions**. Figures 3 show our algorithm ULV has a tighter theoretical upper bound and a lower computation complexity when compared with UCB-BV1. From

Figure 3a, we observe that although the actual regrets incurred by ULV and UCB-BV1 are comparable, our upper bound significantly improves that for UCB-BV1 from around $10^{10}$ to $10^7$. This is consistent with our theoretical analyses in Section 3.2. Note that UCB-BV1's actual regrets can outperform ULV when $B$ is smaller than $10^3$. This occurs because our algorithm is designed based on the assumption of a large budget $B \gg 1$. For the running time, our algorithm beats UCB-BV1 for all instances, specially, our algorithm is about 3 times faster than UCB-BV1 when $B$ increases to $10^5$, as shown in Figure 4. Recall that our regret bound improves the dominant term with $\epsilon$ and $\lambda$ from $\epsilon^{-2}\lambda^{-4}$ to $\epsilon^{-1}\lambda^{-3}$. These improvements can be seen in Figure 3b and Figure 3c, which confirm our theoretical predictions in Lemma 2 and highlight the superiority of the new budget-based analysis framework. Additionally, similar to the observation in Figure 1c in BMAB-FC, the regret bounds continuously increase as $\epsilon$ increases while the actual regrets decrease, as shown in Figure 3b. This implies the possibility of removing the dependence of $\epsilon$ to improve the upper bound.

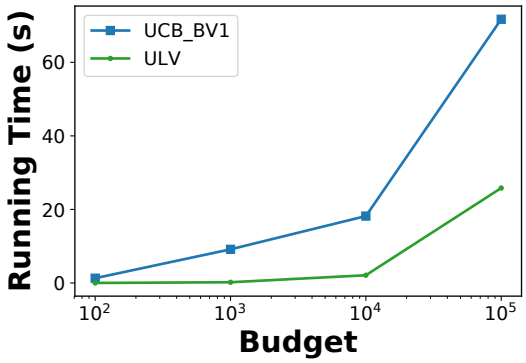



Figure 4: **Our algorithm** ULV **achieves superior computational efficiency compared to baseline methods in** BMAB-VC **setting.** Running time achieved when varying the total budget $B$.

## 8. Other Related Work

Our models belong to the *stochastic* setting where the random rewards (and costs, if they are random) associated with each arm are *i.i.d.* samples from an unknown fixed distribution. Budgeted MAB under the stochastic setting have been studied extensively during the last decade, see (Badanidiyuru et al., 2018b) for a detailed survey. Here we list only a few most relevant to us. The work of (Tran-Thanh et al., 2010) was the first to consider BMAB-FC and they presented a simple $\epsilon$-first approach that achieves a regret bound of $O(B^{2/3})$. Recently, authors in (Rangi & Franceschetti, 2018) generalized BMAB-FC in the way that each arm has an individual capacity (an upper bound on the number of pulls on it) in addition to the global budget. The work of (Badanidiyuru et al., 2018b) proposed a general framework of budgeted MAB: it can be viewed as a generalized version of BMAB-VC where there are multiple resources and pulling each arm will incur a random vector-valued cost. The work of (Xia et al., 2015) designed a Thompson-sampling-based algorithm for budgeted

MAB with variable costs and got distribution-dependent regret bounds. Compared with that work, both of our algorithm and analysis are much easier to implement and follow up.

There is another independent research line that studies the *adversarial* setting where the reward and cost on each arm are pre-arranged by an oblivious adversary during each round. Here are a few examples (Immorlica et al., 2019; Zhou & Tomlin, 2018). Authors in (Kesselheim & Singla, 2020) introduced a general framework of online learning with vector costs and considered both the adversarial and stochastic settings. It can be viewed as a minimization version of budgeted MAB but without rewards. The work of (Rangi et al., 2019b) proposed a general algorithm for budgeted MAB that is proved optimal under both the adversarial and stochastic settings.

## 9. Conclusion and Future Directions

In this paper, we formally stated a unified budget-based analysis framework for two versions of budgeted MAB, namely budgeted MAB with fixed and variable costs, respectively. We proposed two simple UCB-based algorithms to illustrate the power of the framework. Extensive experimental results show the effectiveness and computation efficiency of our algorithms. Moreover, our upper bounds for the proposed algorithms significantly improve those for the benchmark algorithms. We observe that when varying the minimum density gap $d$, the current regret bounds are much higher than the actual regrets incurred. This implies the possibility of further improvement over the dependence of $d$. Another interesting question is, *e.g.,* how to extend our framework to the case when $\lambda$ is unknown for BMAB-VC, and/or when each arm has an individual budget in addition to a global budget.

## Acknowledgement

Evan Yifan Xu's was partially supported by Jiangsu Province Engineering Research Center of Security for Ubiquitous Network. Pan Xu was partially supported by the NSF CRII Award IIS-1948157 and the BGU-NJIT seed grant.

## References

Auer, P. (2002). Finite-time analysis of the multiarmed bandit problem..

Babaioff, M., Dughmi, S., Kleinberg, R., & Slivkins, A. (2015). Dynamic pricing with limited supply. *ACM Transactions on Economics and Computation (TEAC)*, *3*(1), 1–26.

Badanidiyuru, A., Kleinberg, R., & Slivkins, A. (2018a). Bandits with knapsacks. *Journal of the ACM (JACM)*, *65*(3), 1–55.

Badanidiyuru, A., Kleinberg, R., & Slivkins, A. (2018b). Bandits with knapsacks. *J. ACM*, *65*(3), 13:1–13:55.

Chewi, S. (2020). Wald's identity. `https://inst.eecs.berkeley.edu/~ee126/fa17/wald.pdf`. Accessed: 2020-07-12.

Ding, W., Qin, T., Zhang, X.-D., & Liu, T.-Y. (2013). Multi-armed bandit with budget constraint and variable costs. In *Twenty-Seventh AAAI Conference on Artificial Intelligence*.

Gao, G., Wu, J., Xiao, M., & Chen, G. (2020). Combinatorial multi-armed bandit based unknown worker recruitment in heterogeneous crowdsensing. In *IEEE INFOCOM 2020-IEEE Conference on Computer Communications*, pp. 179–188. IEEE.

Huang, Z., Xu, Y., Hu, B., Wang, Q., & Pan, J. (2020). Thompson sampling for combinatorial semi-bandits with sleeping arms and long-term fairness constraints. *CoRR*, *abs/2005.06725*.

Immorlica, N., Sankararaman, K. A., Schapire, R., & Slivkins, A. (2019). Adversarial bandits with knapsacks. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 202–219. IEEE.

Jaillet, P., et al. (2017). Real-time bidding with side information. In *Advances in Neural Information Processing Systems*, pp. 5162–5172.

Kesselheim, T., & Singla, S. (2020). Online learning with vector costs and bandits with knapsacks. *Proceedings of Machine Learning Research vol*, *125*, 1–20.

Kleinberg, R., Niculescu-Mizil, A., & Sharma, Y. (2010). Regret bounds for sleeping experts and bandits. *Machine learning*, *80*(2-3), 245–272.

Lai, T. L., & Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, *6*(1), 4–22.

Li, F., Liu, J., & Ji, B. (2019). Combinatorial sleeping bandits with fairness constraints. In *2019 IEEE Conference on Computer Communications, INFOCOM 2019, Paris, France, April 29 - May 2, 2019*, pp. 1702–1710. IEEE.

Patil, V., Ghalme, G., Nair, V., & Narahari, Y. (2020). Achieving fairness in the stochastic multi-armed bandit problem.. In *AAAI*, pp. 5379–5386.

Rangi, A., & Franceschetti, M. (2018). Multi-armed bandit algorithms for crowdsourcing systems with online estimation of workers' ability. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS 2018, Stockholm, Sweden, July 10-15, 2018*, pp. 1345–1352.

Rangi, A., Franceschetti, M., & Tran-Thanh, L. (2019a). Unifying the stochastic and the adversarial bandits with knapsack. In *IJCAI*, pp. 3311–3317.

Rangi, A., Franceschetti, M., & Tran-Thanh, L. (2019b). Unifying the stochastic and the adversarial bandits with knapsack. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pp. 3311–3317. International Joint Conferences on Artificial Intelligence Organization.

Sankararaman, A., Basu, S., & Sankararaman, K. A. (2020). Dominate or delete: Decentralized competing bandits with uniform valuation. *CoRR*, *abs/2006.15166*.

Schumann, C., Counts, S. N., Foster, J. S., & Dickerson, J. P. (2019a). The diverse cohort selection problem. In Elkind, E., Veloso, M., Agmon, N., & Taylor, M. E. (Eds.), *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS '19, Montreal, QC, Canada, May 13-17, 2019*, pp. 601–609.

Schumann, C., Lang, Z., Foster, J., & Dickerson, J. (2019b). Making the cut: A bandit-based approach to tiered interviewing. In *Advances in Neural Information Processing Systems*, pp. 4639–4649.

Singla, A., Horvitz, E., Kohli, P., & Krause, A. (2015). Learning to hire teams. In *Third AAAI Conference on Human Computation and Crowdsourcing*.

Singla, A., & Krause, A. (2013). Truthful incentives in crowdsourcing tasks using regret minimization mechanisms. In *Proceedings of the 22nd international conference on World Wide Web*, pp. 1167–1178.

Singla, A., Santoni, M., Bartók, G., Mukerji, P., Meenen, M., & Krause, A. (2015). Incentivizing users for balancing bike sharing systems. In *Twenty-Ninth AAAI conference on artificial intelligence*.

Steiger, J., Li, B., & Lu, N. (2022). Learning from delayed semi-bandit feedback under strong fairness guarantees. In *IEEE INFOCOM 2022-IEEE Conference on Computer Communications*, pp. 1379–1388. IEEE.

Tran-Thanh, L., Chapman, A. C., de Cote, E. M., Rogers, A., & Jennings, N. R. (2010). Epsilon-first policies for budget-limited multi-armed bandits. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2010, Atlanta, Georgia, USA, July 11-15, 2010*.

Tran-Thanh, L., Chapman, A. C., Rogers, A., & Jennings, N. R. (2012a). Knapsack based optimal policies for budget-limited multi-armed bandits. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence, July 22-26, 2012, Toronto, Ontario, Canada.*

Tran-Thanh, L., Chapman, A. C., Rogers, A., & Jennings, N. R. (2012b). Knapsack based optimal policies for budget-limited multi-armed bandits. *CoRR*, *abs/1204.1909*.

Tran-Thanh, L., Stavrogiannis, L. C., Naroditskiy, V., Robu, V., Jennings, N. R., & Key, P. B. (2014). Efficient regret bounds for online bid optimisation in budget-limited sponsored search auctions. In Zhang, N. L., & Tian, J. (Eds.), *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence, UAI 2014, Quebec City, Quebec, Canada, July 23-27, 2014*, pp. 809–818. AUAI Press.

Wang, X., Ye, J., & Lui, J. C. (2022). Decentralized task offloading in edge computing: a multi-user multi-armed bandit approach. In *IEEE INFOCOM 2022-IEEE Conference on Computer Communications*, pp. 1199–1208. IEEE.

Xia, Y., Li, H., Qin, T., Yu, N., & Liu, T. (2015). Thompson sampling for budgeted multi-armed bandits. In *IJCAI*, pp. 3960–3966.

Zhou, D. P., & Tomlin, C. J. (2018). Budget-constrained multi-armed bandits with multiple plays. In *Thirty-Second AAAI Conference on Artificial Intelligence*.