Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)

Infinitely Many-Armed Bandits with Budget Constraints

Haifang Li
Institute of Automation,
Chinese Academy of Sciences
lihaifang@amss.ac.cn

Yingce Xia
University of Science
and Technology of China
yingce.xia@gmail.com

Abstract

We study the infinitely many-armed bandit problem with budget constraints, where the number of arms can be infinite and much larger than the number of possible experiments. The player aims at maximizing his/her total expected reward under a budget constraint B for the cost of pulling arms. We introduce a weak stochastic assumption on the ratio of expected-reward to expected-cost of a newly pulled arm which characterizes its probability of being a near-optimal arm. We propose an algorithm named RCB-I to this new problem, in which the player first randomly picks K arms, whose order is sub-linear in terms of B, and then runs the algorithm for the finite-arm setting on the selected arms. Theoretical analysis shows that this simple algorithm enjoys a sub-linear regret in term of the budget B. We also provide a lower bound of any algorithm under Bernoulli setting. The regret bound of RCB-I matches the lower bound up to a logarithmic factor. We further extend this algorithm to the any-budget setting (i.e., the budget is unknown in advance) and conduct corresponding theoretical analysis.

1 Introduction

The multi-armed bandit (MAB) problem is a classical sequential decision problem, where a player receives a random reward by pulling one of the K arms of a slot machine at each round and wants to maximize the expected cumulative reward. Each arm has an unknown reward distribution. The player can only observe the reward of the pulled arm at each round. The core problem of MAB is the tradeoff between exploration (i.e., pulling the less pulled arms) and exploitation (i.e., sticking to the empirical best arms). Numerous real-world applications can be described by MAB models, such as personalized news recommendation (Li et al. 2010), auction mechanism design (Mohri and Munoz 2014), online advertising (Tran-Thanh et al. 2014), crowdsourcing (Zhou, Chen, and Li 2014) and so on. MAB problems have been extensively studied in machine learning and statistics community (Lai and Robbins 1985; Beygelzimer et al. 2011; Bubeck and Cesa-Bianchi 2012). Many algorithms have been proposed, like UCB1, ϵ_n-Greedy (Auer, Cesa-Bianchi, and Fischer 2002), UCB-V (Audibert, Munos, and Szepesvári 2007; 2009), KL-UCB (Garivier and Cappé 2011), and Bayes-UCB (Kaufmann, Cappé, and Garivier 2012).

Recently, the MAB problem with budget constraints (i.e., budgeted MAB), has been utilized to model some new industrial situations, for example, online bidding optimization in sponsored search (Amin et al. 2012; Tran-Thanh et al. 2014) and on-spot instance bidding in cloud computing (Agmon Ben-Yehuda et al. 2013; Ardagna, Panicucci, and Passacantando 2011). In budgeted MAB, besides the random reward, each arm is also associated with an unknown random cost. At each round, the player pulls an arm, receives a random reward and pays a random cost, until his budget runs out. Many algorithms have been developed under different settings for the budgeted MAB. Tran-Thanh et al. (2010), Tran-Thanh et al. (2012), and Vanchinathan et al. (2015) designed ϵ-first, KUBE and GP-SELECT algorithms respectively to study the budgeted MAB with deterministic cost of each arm. Ding et al. (2013) studied the problem with unknown discrete cost distributions and proposed the UCB-BV algorithm. Xia et al. (2015a) designed the budget-UCB algorithm to investigate the random continuous costs. Another line of research, the best arm identification problem for budgeted MAB, was proposed in (Xia et al. 2016b).

Berry et al. (1997) first addressed the MAB problem with infinitely many arms, where each arm is associated with a Bernoulli reward distribution. They tried to characterize the case in which the number of arms can be infinite and much larger than the possible number of experiments. The general reward distribution problems are studied by Wang, Audibert, and Munos (2009) and Carpentier and Valko (2015). Wang, Audibert, and Munos (2009) proposed UCB-V(∞) algorithm to deal with this case. Carpentier and Valko (2015) addressed the optimal arm identification problem for infinite-arm setting and provided SiRI algorithm. However, the costs per round in the aforementioned works are all 1's. In this paper, we introduce the random cost into the infinitely many-armed setting, and propose the *infinitely many-armed bandits problem with budget constraints* (denoted as ∞-BMAB, where the first "B"refers to "Budgeted".).

All prior works on budgeted MAB focus on the finite number of arms, in which each arm will be sufficiently explored. These algorithms cannot be directly applied to an infinitely many-armed bandit problem, since it is impossible to make exploration on each arm. We design efficient algo-

rithms to this new problem and perform formal theoretical analysis. Our contributions can be summarized in the following two aspects:

Algorithm: We propose a *Ratio Confidence Bound for ∞-BMAB* (briefly denoted as RCB-I) algorithm in this work. The algorithm consists of two steps: first, we randomly pick K arms, where the choice of K should balance the trade-off between exploring enough arms (in order to include the arm whose expected reward to expected-cost ratio[1] is close enough to the supremum ratio of all candidates), and avoiding selecting too many arms (in case that we waste much budget on exploring them). Then we adopt a UCB-style algorithm to tackle the finite arms problem (See Section 3). The choice of K usually depends on budget B. On the other hand, we will often come across the case that B is unknown in advance. Therefore, we propose an any-budget algorithm with a variant of the RCB-I algorithm, named RCB-AIR to deal with this case, in which we will check whether we should explore a new arm at the beginning of each round (See Section 5).

Theoretical analysis: We make extensive theoretical analysis on ∞-BMAB. We prove that both RCB-I and RCB-AIR enjoy sub-linear regrets in terms of the budget. Compared with the prior works, there are two challenges for ∞-BMAB to be tackled: (i) we cannot make full exploration of all the arms since the number of arms is infinite; (ii) we cannot decompose the expected pulling number of each suboptimal arm as that of the finite-arm setting, otherwise we cannot get a finite regret bound. For the first challenge, we make investigation on finitely chosen arms, and bridge the regret between the finitely selected arms and all the infinite arms through the stochastic regularity assumption on the ratio distribution (see (1)). As to the second one, we decompose the expected pulling number of each suboptimal arm till round τ_B as in Eqn. (10), which includes a product term. It is the product form that enables us utilizing Eqn. (1) and obtaining a finite regret bound. Then we provide a lower bound for any algorithm for Bernoulli bandits (whose rewards and costs are either 0 or 1) with infinitely many arms. We show that our proposed RCB-I can match the lower bound up to a logarithmic factor.

2 Problem Formulation

We consider a stochastic infinitely many-armed bandit problem with budget constraints. At each round $t \in \mathbb{N}_+$, the player pulls an arm $i \in \mathbb{N}_+$, receives a random reward $r_i(t)$ and pays a random cost $c_i(t)$, until he/she runs out of the budget B. B is a positive real number, which might be known/not known in advance. Both the reward $r_i(t)$ and cost $c_i(t)$ are supported in $[0, 1]$. We follow the setting in (Xia et al. 2015a) and make the assumptions about independence of rewards and costs: (i) the rewards (costs) of the same arm at different rounds are independent and identically distributed; (ii) the rewards and costs of an arm are independent of the other arms. Note that we do not assume that the rewards of an arm are independent of its costs. For any $i \in \mathbb{N}_+$, denote the expected reward and expected cost of arm i as u_i^r and u_i^c respectively. Without loss of generality, we assume $u_i^r \in (0, 1)$ and $u_i^c \in [\lambda, 1)$, where λ belongs to $(0, 1)$. This assumption is very reasonable and natural, since in practice whatever non-trivial action a player takes, he/she needs to afford a certain non-zero cost. Denote the ratio of expected reward to expected cost of arm i as ρ_i, i.e., $\rho_i = \frac{u_i^r}{u_i^c}$.

Since there is infinite number of arms in our setting, it is impossible to pull each arm once. Similar to (Berry et al. 1997; Wang, Audibert, and Munos 2009; Carpentier and Valko 2015), we make the following stochastic regularity assumption on the ratio of expected reward to expected cost: given ρ^*, which is the supremum (or, maximum) expected-reward to expected-cost ratio, the probability that a newly pulled arm is ϵ-*optimal* is of order ϵ^β for small $\epsilon(> 0)$, where $\beta(\geq 0)$ is a parameter known in advance. Mathematically,

$$\mathbb{P}\{\rho > \rho^* - \epsilon\} = \Theta(\epsilon^\beta), \text{ for } \epsilon \to 0, \tag{1}$$

where ρ is the ratio of the expected reward to the expected cost of the newly pulled arm, and $\Theta(\epsilon^\beta)$ means that there exist two positive constants c_1 and c_2 such that for any $\epsilon \in [0, \rho^*]$,

$$c_1\epsilon^\beta \leq \mathbb{P}\{\rho > \rho^* - \epsilon\} \leq \mathbb{P}\{\rho \geq \rho^* - \epsilon\} \leq c_2\epsilon^\beta. \tag{2}$$

One can verify that the uniform distribution on $(0, \rho^*)$ satisfies the Eqn. (1) with $\beta = 1$, any $c_1 \in (0, 1/\rho^*]$ and any $c_2 \in [1/\rho^*, \infty)$.

Our goal is to design algorithms for ∞-BMAB in order to maximize its expected cumulative rewards, or equivalently to minimize the pseudo-regret, before the budget runs out. We define the pseudo-regret of any algorithm as follows:

$$\mathcal{R}^{\text{alg}} = R^* - \mathbb{E}[\sum_{t=1}^{\infty} r_{I_t}(t)\mathbf{1}\{B_t \geq 0\}], \tag{3}$$

where R^* is the supremum of expected reward of all the possible pulling algorithms[2], I_t denote the arm chosen at round t, $\mathbf{1}\{\cdot\}$ is the indicator function, B_t is the remaining budget at round t, i.e., $B_t = B - \sum_{s=1}^{t} c_{I_s}(s)$, $\mathcal{R}^{\text{alg}}$ is the regret of algorithm "alg", and the expectation $\mathbb{E}$ is taken w.r.t. the randomness of the rewards, costs and the pulling algorithm[3]. Note that in our setting, there are not such hard constraints that the pulling procedures cannot exceed specific rounds, which make our work differ from (György et al. 2007; Badanidiyuru, Kleinberg, and Slivkins 2013; Badanidiyuru, Langford, and Slivkins 2014; Wu et al. 2015).

3 Algorithms

As pointed in (Xia et al. 2016a), even when the number of arms is finite, and the reward and cost of each arm are deterministic, the budgeted MAB problem is an *unbounded knapsack problem* (Martello and Toth 1990), which is NP-hard (Lueker 1975). Consequently, it is much more challenging

[1] According to the previous literature and the analysis of our work, the ratio of expected-reward to expected-cost of each arm is an important factor.

[2] According to the analysis in Appendix B, we have that R^* is smaller than $(B + 1)\rho^*$, which shows that R^* exists. Due to space limitations, all the appendices are left in the online full version of this work (Li and Xia 2016).

[3] All the notations are summarized at the end of this paper.

to obtain the optimal algorithms for the infinitely many arms setting.

According to (Xia et al. 2016a), we have that for budgeted MAB with finite arms: (a) When the reward and cost distributions of each arm are known, always pulling the arm with the maximum ratio (denote the ratio as $\tilde{\rho}$) of expected reward to expected cost can bring almost the same expected reward as the optimal policy (the policy which can bring the maximum expected reward given the reward and cost distributions of all the arms) , the gap of which is at most $2\tilde{\rho}$. (b) When the reward and cost distributions of each arm are not known, at each round, we should try to pull the arm with the maximum ratio of empirical reward to empirical cost, while maintaining enough exploration on the less pulled arms.

Given the assumption in Eqn. (1), we know that if we randomly pick K arms where K is large enough, the maximum expected-reward to expected-cost ratio of the selected arms is close enough to that of the infinitely many arms. Then, we could run the algorithms for finite-arm budgeted MAB on the randomly picked arms. When choosing K, we should consider the tradeoff between exploring more arms in order to search a candidate arm with the expected-reward to expected-cost ratio close enough to the best one and exploiting the empirical best arm in order not to waste much budget on exploration.

Algorithm for Finite-arm Case For ease the reference, let $T_i(t-1), \overline{r}_{i,t}, \overline{c}_{i,t}$ and $\mathcal{E}_{i,t}$ denote the number of pulling rounds, the empirical average reward, the empirical average cost and a confidence term of arm i before (excluding) round t respectively. Mathematically,

$$T_i(t-1) = \sum_{s=1}^{t-1} \mathbf{1}\{I_s = i\},\ \mathcal{E}_{i,t} = \sqrt{\frac{\mathcal{E}_{t-1}}{2T_i(t-1)}},$$
$$\overline{r}_{i,t} = \frac{\sum_{s=1}^{t-1} r_i(s)\mathbf{1}\{I_s = i\}}{T_i(t-1)},\ \overline{c}_{i,t} = \frac{\sum_{s=1}^{t-1} c_i(s)\mathbf{1}\{I_s = i\}}{T_i(t-1)},$$

where $\{\mathcal{E}_t\}_{t\geq 0}$ is a nondecreasing sequence of nonnegative numbers. We call $\mathcal{E}_t$ as exploration sequence.

We adapt a ratio confidence bound style algorithm introduced in (Xia et al. 2016a) as a subroutine of our algorithm, which can deal with the finite-arm budgeted MAB. For ease of reference, denote the subroutine as RCB. At each round, one should pull the arm with the maximum index defined as follows:

$$D_{i,T_i(t-1),t} = \frac{\min\{\overline{r}_{i,t} + \mathcal{E}_{i,t}, 1\}}{\max\{\overline{c}_{i,t} - \mathcal{E}_{i,t}, 0\}}. \tag{4}$$

Note that arms with larger ratio of empirical average reward to empirical average cost, or fewer pulling rounds, would have larger indices. RCB is formally described[4] in Algorithm 1.

Algorithm for Infinite-arm Case Given the algorithm for finite-arm case, we only need to design the number of randomly picked arms. After careful derivations, we find that if we randomly select K arms, where

$$K = \begin{cases} \Theta(B^{\beta/2}) & \text{if } \beta < 1, \\ \Theta(B^{\frac{\beta}{1+\beta}}) & \text{if } \beta \geq 1. \end{cases} \tag{5}$$

[4]In step 4 of Algorithm 1, if there is more than one arm with maximum index $D_{i,T_i(t-1),t}$, randomly pick one.

Algorithm 1: RCB subroutine

1 *Input*: The randomly picked K arms;
2 Pull each arm once at the first K rounds and set $t \leftarrow K+1$;
3 **while** *the budget has not run out* **do**
4 Pull arm I_t with the largest index of Eqn. (4), i.e., $I_t = \arg\max_{i\in[K]} D_{i,T_i(t-1),t}$;
5 Update $T_{I_t}(t)$, $\overline{r}_{I_t,t}$, $\overline{c}_{I_t,t}$ and the left budget; set $t \leftarrow t+1$.

we could achieve sub-linear regret w.r.t. the budget. Our proposed algorithm for budgeted MAB with infinite arms, ***R**atio **C**onfidence **B**ound for **I**nfinitely many-armed bandits with budget constraints* (briefly denoted as RCB-I), is shown in Algorithm 2.

Algorithm 2: RCB-I Algorithm

1 *Input*: The ratio distribution regularity parameter β, the budget B;
2 Randomly choose K arms, which is defined in Eqn. (5);
3 Run RCB subroutine on the selected K arms.

4 Theoretical Analysis

In this section, we conduct theoretical analysis for our proposed algorithm. We first give an upper bound of the regret of RCB-I. Then, we derive a lower bound for any algorithm under budgeted Bernoulli setting (whose rewards and costs are either 0 or 1). At last, make some discussions on upper bound and lower bound of the regret.

For ease of reference, we introduce the following two notations, which would be used quite frequently.

(i) $\Delta_k := \rho^* - \rho_k$, which describes the difference of the expected-reward to expected-cost ratio between the possible optimal one and that of a suboptimal arm k.

(ii) $\tau_B := \lfloor \frac{2B}{\lambda} \rfloor$, which can be seen as the *pseudo stopping time* of the budgeted MAB problem, since when B is large enough, the probability that the pulling rounds can exceed τ_B, bounded by $\mathcal{X}(B)$, is very small, where $\mathcal{X}(B)$ denotes the order $O\big(B\exp(-\frac{B\lambda}{2})\big)$.

4.1 Upper Bound of $\mathcal{R}^{\text{RCB-I}}$

In this subsection, we derive an upper bound of the regret for the RCB-I algorithm, as shown in Theorem 1.

Theorem 1. *For the ∞-BMAB satisfying Eqn.* (1)*, when the exploration sequence $\mathcal{E}_t$ satisfies: $2\log(4(\log_2 t+1)) \leq \mathcal{E}_t \leq \log t$, the upper bound of the regret of RCB-I is shown as below.*

$$\mathcal{R}^{\text{RCB-I}} \leq \begin{cases} CB^{1/2}\log B & \text{if } \beta < 1, \\ CB^{1/2}(\log B)^2 & \text{if } \beta = 1, \\ CB^{\frac{\beta}{1+\beta}}\log B & \text{if } \beta > 1, \end{cases} \tag{6}$$

where C is a constant depending only on c_1, c_2, β, λ.

Our proof consists of three main steps: First we analyze the regret on the randomly chosen K arms. Then we make a bridge of the regret between the randomly chosen K arms and infinitely many arms through the stochastic regularity assumption on the ratio distribution (see (1)). Finally, we summarize all the derivations and eventually get Theorem 1.

To increase readability, we leave some detailed derivations in the Appendix C and provide the proof sketch only.

(S1): *Regret analysis on the selected K arms.*

Define the regret of RCB on the given K arms[5], compared to the optimal policy obtained from the infinity many arms, as follows:

$$\mathcal{R}_K^{\text{RCB-I}} = \mathcal{R}^* - \mathbb{E}\big[\sum_{t=1}^{\infty}\sum_{k=1}^{K} r_k(t)\mathbf{1}\{I_t = k, B_t \geq 0\}\big|\mathcal{S}_K\big], \quad (7)$$

where $\mathcal{S}_K$ represents the event "randomly select K arms", and R^* is defined in (3).

Conditioned on $\mathcal{S}_K$, by similar derivations to the Eqn. (10) in (Xia et al. 2016a), we can obtain that[6]

$$\mathcal{R}_K^{\text{RCB-I}} \leq \frac{2}{\lambda} + \sum_{k=1}^{K}\Delta_k\mathbb{E}[T_k(\tau_B)] + O\big(B\exp(-\frac{1}{2}B\lambda)\big), \quad (8)$$

where $T_k(\tau_B)$ denotes the pulling number of arm k from round 1 to round τ_B. The order $O(\cdot)$ in (8) comes from the randomness of the stopping time. The reasons why we bridge $\mathcal{R}_K^{\text{RCB-I}}$ and $T_k(\tau_B)$ are: (i) The randomness of the stopping time is removed since we introduce the pseudo stopping time τ_B , due to which which we can use concentration inequalities safely; (ii) We can adapt the techniques about bounding the pulling rounds of suboptimal arms from finite-armed MAB's.

Next we only need to focus on upper bounding $\mathbb{E}[T_k(\tau_B)]$.

(S1-1): *Decompose* $\mathbb{E}[T_k(\tau_B)]$.

Given $\mathcal{S}_K$, for any positive integer L_k, $\mathbb{E}[T_k(\tau_B)]$ can be decomposed into three components: a constant invariant to t and the two probability terms. Specifically, we get that

$$\mathbb{E}[T_k(\tau_B)] \leq L_k + \sum_{t=1}^{\tau_B}\sum_{s=L_k}^{t-1}\mathbb{P}\{E_{k,s,t}\} \quad (9)$$

$$+\sum_{t=1}^{\tau_B}\prod_{k'\neq k}\mathbb{P}\{\exists s' \in [1, t-1], E'_{k',s',t}\}, \quad (10)$$

where $L_k = \lceil \frac{2\log\tau_B}{\eta(\lambda)^2\Delta_k^2}\rceil, \eta(\lambda) = \frac{\lambda^2}{3+2\lambda}$, $\varphi_k = \rho_k + \frac{1}{2}\Delta_k$, and $E_{k,s,t}, E'_{k',s',t}$ denote these two events $\mathbf{1}\{D_{k,s,t} > \varphi_k\}$ and $\mathbf{1}\{D_{k',s',t} \leq \varphi_k\}$ respectively.

In the following two sub-steps, we get down to bounding the two probability terms in (9) and (10).

Remark: The $\prod_{k'\neq k}\mathbb{P}\{\cdot\}$ in term (10) does not exist in the analysis of finite-armed budgeted bandits. If we directly use the proof technique for the finite-armed settings, term (10) would become linear w.r.t. K, and consequently, make the regret bound not a finite number.

[5]For ease of reference, throughout step (**S1**), denote the indices of the selected K arms as $1, 2, \cdots, K$.

[6]Throughout (**S1**), the expectation $\mathbb{E}(\cdot)$ and the probability $\mathbb{P}\{\cdot\}$ are $\mathbb{E}(\cdot|\mathcal{S}_K)$ and probability $\mathbb{P}\{\cdot|\mathcal{S}_K\}$. We omit the $\mathcal{S}_K$ for simplicity.

(S1-2): *Bound term* (9).

It is easy to verify the following inequality holds for any $k \geq 1$:

$$\varphi_k = \rho_k + \frac{1}{2}\Delta_k \geq \frac{u_k^r + \frac{\lambda^2}{3+2\lambda}\Delta_k}{u_k^c - \frac{\lambda^2}{3+2\lambda}\Delta_k}. \quad (11)$$

For any $k \in [K]$, $s \in [L_k, t-1]$ and $t \in [1, \tau_B]$, we define the following two events:

(i) $E_{k,s,t}^r : \overline{r}_{k,t} - u_k^r > \eta(\lambda)\Delta_k - \sqrt{\frac{\mathcal{E}_{t-1}}{2s}}$;

(ii) $E_{k,s,t}^c : \overline{c}_{k,t} - u_k^c < -\eta(\lambda)\Delta_k + \sqrt{\frac{\mathcal{E}_{t-1}}{2s}}$.

If $E_{k,s,t}$ holds, at least one of the two events $E_{k,s,t}^r$ and $E_{k,s,t}^c$ would hold. As a result, we have

$$\mathbb{P}\{E_{k,s,t}\} \leq \mathbb{P}\{E_{k,s,t}^r\} + \mathbb{P}\{E_{k,s,t}^c\}. \quad (12)$$

By leveraging Hoeffding's inequality on the two terms in the right-hand side of (12), we obtain that

$$\begin{aligned}\mathbb{P}\{E_{k,s,t}^r\} &\leq \exp(-\frac{s\eta(\lambda)^2\Delta_k^2}{2});\\ \mathbb{P}\{E_{k,s,t}^c\} &\leq \exp(-\frac{s\eta(\lambda)^2\Delta_k^2}{2}).\end{aligned} \quad (13)$$

Consequently, by conducting some derivations, we have

$$\sum_{t=1}^{\tau_B}\sum_{s=L_k}^{t-1}\mathbb{P}\{E_{k,s,t}\} \leq \frac{5}{\eta(\lambda)^2\Delta_k^2}. \quad (14)$$

(S1-3): *Bound term (10).*

For ease of the reference, we define another three events as follows, for any $k' \neq k \in [K]$, $s \in [1, t-1]$ and $t \in [1, \tau_B]$:

(iii) $\tilde{E}_{k',s',t} : D_{k',s',t} \leq \rho_{k'}$;

(iv) $\tilde{E}_{k',s',t}^r : \overline{r}_{k',t} - u_{k'}^r \leq -\sqrt{\frac{\mathcal{E}_{t-1}}{2s}}$;

(v) $\tilde{E}_{k',s',t}^c : \overline{c}_{k',t} - u_{k'}^c \geq \sqrt{\frac{\mathcal{E}_{t-1}}{2s}}$.

It is easy to obtain that

$$\begin{aligned}&\prod_{k'\neq k}\mathbb{P}\{\exists s' \in [1, t-1], E'_{k',s',t}\}\\ \leq &\prod_{k':\rho_{k'} > \rho^* - \frac{1}{2}\Delta_k}\mathbb{P}\{\exists s' \in [1, t-1], \tilde{E}_{k',s',t}\},\end{aligned} \quad (15)$$

If $\tilde{E}_{k',s',t}$ holds, at least one event of $\tilde{E}_{k',s',t}^r$ and $\tilde{E}_{k',s',t}^c$ holds. Thus, we have

$$\begin{aligned}\mathbb{P}\{\exists s' \in [1, t-1], \tilde{E}_{k',s',t}\} \leq &\mathbb{P}\{\exists s' \in [1, t-1], \tilde{E}_{k',s',t}^r\}\\ &+\mathbb{P}\{\exists s' \in [1, t-1], \tilde{E}_{k',s',t}^c\}.\end{aligned} \quad (16)$$

The two terms in the right-hand side of (16) could be upper bounded as below. By applying the peeling argument with a geometric grid over the time interval $[1, t-1]$ and hoeffding maximal inequality (Bubeck 2010), we have

$$\mathbb{P}\{\exists s' \in [1, t-1], \tilde{E}_{k',s',t}^r\} \leq (\log_2(t-1)+1)\exp(-\frac{\mathcal{E}_{t-1}}{2}).$$

Similarly, we have

$$\mathbb{P}\{\exists s' \in [1, t-1], \tilde{E}^c_{k',s',t}\} \leq (\log_2(t-1)+1)\exp(-\frac{\mathcal{E}_{t-1}}{2}).$$

Since $\mathcal{E}_t \geq 2\log(4(\log_2 t + 1))$, the following inequality holds:

$$\mathbb{P}\{\exists s' \in [1, t-1], \tilde{E}_{k',s',t}\} \leq \frac{1}{2}. \tag{17}$$

Therefore,

$$\sum_{t=1}^{\tau_B}\prod_{k'\neq k}\mathbb{P}\{\exists s' \in [1, t-1], E'_{k',s',t}\} \leq \tau_B 2^{-N_{\Delta_k}}, \tag{18}$$

where N_{Δ_k} is the cardinal of $\{k' \in [K] : \rho_{k'} > \varphi_k\}$.

In conclusion, by combining inequalities (9), and (14) in (**S1-2**) and (18) in (**S1-3**), we get that

$$\mathbb{E}[T_k(\tau_B)] \leq \lceil \frac{2\log\tau_B}{\eta(\lambda)^2\Delta_k^2}\rceil + \frac{5}{\eta(\lambda)^2\Delta_k^2} + \tau_B 2^{-N_{\Delta_k}}. \tag{19}$$

(S2): *Bridge $\mathcal{R}_K^{RCB\text{-}I}$ and $\mathcal{R}^{RCB\text{-}I}$.*

The first step **(S1)** makes progress based on the condition that the randomly chosen K arms are given. In this step, we try to utilize the stochastic regularity assumption on the expected-reward to expected-cost ratio distribution (See (1) in Section 2) in order to bridge $\mathcal{R}_K^{\text{RCB-I}}$ and $\mathcal{R}^{\text{RCB-I}}$.

According to Eqn. (1), the quantities $\Delta_1, ..., \Delta_K$ are i.i.d. random variables satisfying $0 \leq \Delta_k \leq \rho^*$ and $\mathbb{P}\{\Delta_k \leq \epsilon\} = \Theta(\epsilon^\beta)$. Combine (8) and (19), and take expectations w.r.t. all sources of randomness. Therefore, we have

$$\begin{aligned}\mathcal{R}^{\text{RCB-I}} =&\mathbb{E}[\mathcal{R}_K^{\text{RCB-I}}] \leq \frac{2}{\lambda} + K\mathbb{E}\big[\big(\frac{10}{\eta(\lambda)^2\Delta_1}\log\tau_B\big)\wedge(\tau_B\Delta_1)\big]\\ &+\tau_B K\mathbb{E}\big[\Delta_1\cdot 2^{-N_{\Delta_1}}\big] + O\big(B\exp(-\frac{1}{2}B\lambda)\big),\end{aligned} \tag{20}$$

where $a \wedge b := \min\{a, b\}$.

Then we only need to bound the two expectation terms in the right-hand side of (20) in the next two sub-steps.

(S2-1): *Bound the first expectation term in (20).*

Since $\mathbb{P}\{\Delta_1 \leq \epsilon\} = \Theta(\epsilon^\beta)$ and according to the expectation definition, we can obtain that

$$\mathbb{E}[(\frac{10\log\tau_B}{\eta(\lambda)^2\Delta_1})\wedge(\tau_B\Delta_1)] \leq \begin{cases} O\big(\tau_B^{\frac{1-\beta}{2}}\log\tau_B\big) & \text{if } \beta<1,\\ O\big((\log\tau_B)^2\big) & \text{if } \beta=1,\\ O\big(\log\tau_B\big) & \text{if } \beta>1.\end{cases}$$

(S2-2): *Bound the second expectation term in (20).*

Conditioning on Δ_1, N_{Δ_1} follows a binomial distribution with parameters $K-1$ and $\mathbb{P}\{\rho_1 > \rho^* - \frac{\Delta_1}{2}|\Delta_1\}$. Then according to the total expectation formula and the expectation definition, we can obtain that

$$\mathbb{E}[\Delta_1 2^{-N_{\Delta_1}}] \leq O\big(K^{-1-1/\beta}\log(K)\big). \tag{21}$$

(S3): *Bound the regret of the RCB-I algorithm $\mathcal{R}^{RCB\text{-}I}$.*

Combine the above steps, we get that

$$\mathcal{R}^{\text{RCB-I}} \leq \begin{cases} C\big[KB^{\frac{1-\beta}{2}}\log B + BK^{-1/\beta}\log K\big] & \text{if } \beta<1,\\ C\big[K(\log B)^2 + BK^{-1/\beta}\log K\big] & \text{if } \beta=1,\\ C\big[K\log B + BK^{-1/\beta}\log K\big] & \text{if } \beta>1,\end{cases}$$

where C is a constant depending only on c_1, c_2, β (See Eqn. (2)) and λ. Substitute K by Eqn. (5), and then we can get the desired result.

4.2 Lower Bound of Any Algorithm

In this subsection, we derive a lower bound for the regret of any algorithm for Bernoulli ∞-BMAB. A Bernoulli ∞-BMAB is a special bandit, where the rewards and costs of each arm follow two Bernoulli distributions with unknown parameters. The lower bound of regret of any algorithm for Bernoulli ∞-BMAB is shown as follows:

Theorem 2. *For any Bernoulli ∞-BMAB, if the parameters of any newly pulled arm satisfy Eqn.* (1)*, for any $\beta > 0$, any algorithm suffers a regret larger than $cB^{\frac{\beta}{1+\beta}}$ for some small enough constant c depending on c_2, β and λ.*

Due to space restrictions, here we just only show the proof sketch of Theorem 2, and leave the completed proof into the Appendix D.

(S1): *Lower bound analysis on $\mathcal{S}_K$.*

Given K randomly selected arms (denoted as $\mathcal{S}_K$), under the Bernoulli setting, according to (Xia et al. 2015b), we have

$$\mathcal{R}_K^{\text{RCB-I}} = \sum_{k=1}^K u_k^c\Delta_k\mathbb{E}[T_k(B)] \geq \lambda\sum_{k=1}^K\Delta_k\mathbb{E}[T_k(B)], \tag{22}$$

where $T_k(B)$ denotes the pulling number of arm k until the budget B runs out.

Since the cost of each arm is no larger than 1, the algorithm runs at least B rounds, i.e., $\sum_{k=1}^K\mathbb{E}[T_k(B)] \geq B$.

Now let $0 < \delta < \delta' < \rho^*$. Similar to the derivations of Theorem 3 in (Wang, Audibert, and Munos 2009), we have

$$\mathcal{R}_K^{\text{RCB-I}} \geq \lambda B\delta\mathbf{1}\{\hat{\rho} \leq \rho^* - \delta\} + \lambda\kappa\delta'\mathbf{1}\{\hat{\rho} > \rho^* - \delta; \overline{K} \geq \kappa\}. \tag{23}$$

where $\kappa > 0$ is a parameter to be determined, $\overline{K}$ denotes the cardinality of $\{k \in \{\tilde{I}_1, ..., \tilde{I}_{K^*-1} : \rho_k \leq \rho^* - \delta'\}$, $K^* := \min\{l \in \mathbb{N}^+, \rho_{\tilde{I}_l} > \rho^* - \delta\}$, $\tilde{I}_l$ is the l-th arm drawn, $\hat{\rho}$ denotes the expected reward to expected cost ratio of the best arm in $\{\tilde{I}_1, ..., \tilde{I}_K\}$.

(S2): *Bridge $\mathcal{R}_K^{RCB\text{-}I}$ and $\mathcal{R}^{RCB\text{-}I}$.*

First, let we take $\kappa = \frac{B\delta}{\delta'}$ and take expectations on both sides of (23). Therefore, we have

$$\mathcal{R}^{\text{RCB-I}} = \mathbb{E}[\mathcal{R}_K^{\text{RCB-I}}] \geq \lambda B\delta\mathbb{P}\{\overline{K} \geq \kappa\}. \tag{24}$$

Next, we need to obtain the distribution of $\overline{K}$. Since K^* follows a geometric distribution with parameter $\mathbb{P}\{\rho > \rho^* - \delta\}$, and given K^*, $\overline{K}$ follows a binomial distribution with parameters $K^* - 1$ and $\mathbb{P}\{\rho \leq \rho^* - \delta'\}$, by technical derivations, we can obtain that the random variable $\overline{K}$ follows a distribution as follows.

$$\mathbb{P}\{\overline{K} = \iota\} = \frac{\big[\mathbb{P}\{\rho \leq \rho^* - \delta'\}\big]^\iota}{\big[\mathbb{P}\{\rho \notin (\rho^* - \delta', \rho^* - \delta]\}\big]^{\iota+1}}. \tag{25}$$

Therefore, we have

$$\mathcal{R}^{\text{RCBI}} \geq \lambda B\delta\frac{\mathbb{P}\{\rho \leq \rho^* - \delta'\}^\kappa}{\mathbb{P}\{\rho \notin (\rho^* - \delta', \rho^* - \delta]\}^\kappa\mathbb{P}\{\rho > \rho^* - \delta\}}. \tag{26}$$

Taking $\delta = \delta' B^{-\frac{1}{1+\beta}}$, where δ' could be any constant in $(0, \rho^*)$, we have $\kappa = B^{\frac{\beta}{1+\beta}}$, and we obtain the desired result.

4.3 Discussions

In this subsection, we make some discussions on Theorem 1 and Theorem 2.

(1) Theorem 1 shows that RCB-I achieves a sub-linear regret bound with respect to the budget B, and we have $\lim_{B\to\infty}(\mathcal{R}^{\text{RCB-I}}/B) = 0$.

(2) Comparing Theorem 1 with Theorem 2, we obtain that the upper bound of the regret of RCB-I matches the lower bound up to a logarithmic factor[7] $\log B$.

(3) Compared with the budgeted MAB with finite arms, the regret bound of proposed algorithm, as well as the lower bound for any algorithm under Bernoulli reward/cost distributions setting, cannot achieve $O(\ln B)$. This is because we have to explore enough arms so as to obtain an arm with the expected-reward to expected-cost ratio close enough to the ρ^* with high probability.

5 Extension to Any Budgets

In practice, we often come across the case that B is not known in advance, or changed through time. Inspired by (Wang, Audibert, and Munos 2009), in this section, we present an any-budget algorithm with a variant of the RCB-I algorithm, named RCB-AIR (short for Arm Increasing Rule) to deal with the case. The main idea is that, at the beginning of each round, we will check whether we should explore a new arm (which is determined by the number of the explored arms and the round number), then run the procedure for the finite arm case.

Denote $\mathcal{K}_t$ as the arms pulled up to round t. Define $K_t = |\mathcal{K}_t|$. We set $\mathcal{K}_t = \emptyset$ and $K_0 = 0$. The RCB-AIR algorithm is shown in Algorithm 3.

Algorithm 3: RCB-AIR Algorithm

Input: The ratio distribution regularity parameter $\beta > 0$;
while *the budget has not run out* **do**
  At round t, if $K_{t-1} < \begin{cases} t^{\beta/2} & \text{if } \beta < 1 \\ t^{\frac{\beta}{\beta+1}} & \text{if } \beta \geq 1 \end{cases}$,
    randomly pick a new arm a_t and set $\mathcal{K}_t \leftarrow \mathcal{K}_{t-1} \cup \{a_t\}$; otherwise, set $\mathcal{K}_t \leftarrow \mathcal{K}_{t-1}$;
  Run Step 4 and 5 of Algorithm 1 on the selected $\mathcal{K}_t$.

The regret bound of RCB-AIR algorithm is shown as follows:

Theorem 3. *For the ∞-BMAB satisfying Eqn.* (1)*, when the exploration sequence $\{\mathcal{E}_t\}$ satisfies $2\log(4(\log_2 t + 1)) \leq \mathcal{E}_t \leq \log t$, for any budget B, the upper bound of the regret for RCB-AIR is shown as follows.*

$$\mathcal{R}^{\text{RCB-AIR}} \leq \begin{cases} C(\log B)^2 B^{\frac{1}{2}} & \text{if } \beta > 1, \\ C(\log B)^2 B^{\frac{\beta}{1+\beta}} & \text{if } \beta \leq 1, \end{cases} \tag{27}$$

where C is a constant depending only on c_1, c_2, β, λ.

[7] In fact, the upper bound matches (up to a logarithmic factor) the lower bound in the case $\beta \geq 1$. We will consider the case $\beta < 1$ in the future.

Similar to the proof of Theorem 1, the proof of Theorem 3 also consists of three steps: First, analyze the regret on the randomly chosen $\{K_t\}_{t=1}^{\tau_B}$ arms. Note to mention that the arms chosen until round τ_B progressively enter in competition, which is different from the RCB-I setting; second, relate $\mathcal{R}^{\text{RCB-AIR}}_{K_{\tau_B}}$ to $\mathcal{R}^{\text{RCB-AIR}}$ by leveraging the stochastic assumption on the expected-reward to expected-cost ratio distribution; third, combine the results of the above two steps. We leave the detailed proofs into Appendix E.

6 Conclusion and Future Work

In this paper, we design an arm pulling algorithm, RCB-I, for the ∞-BMAB. We have proved that when the budget is known in advance, RCB-I achieves a sub-linear regret bound with respect to the budget, and matches (up to a logarithmic factor) the lower bound. We further make an extension to any budget setting, propose the RCB-AIR algorithm, and conduct corresponding theoretical analysis.

For the future work, there are many important and interesting directions: (1) the distribution-free regret bounds remain empty for our proposed algorithm, which need to be solved; (2) whether the lower bound for $\beta < 1$ can be improvable is also an intriguing problem in need of discussion; (3) the case that the rewards and costs of different arms are correlated requires further investigation; (4) the best arm identification problem for ∞-BMAB is another important problem.

Notations

Notation	Meaning
B	budget
B_t	budget at round t
$r_i(t), c_i(t)$	reward (cost) of arm i at round t
u_i^r, u_i^c	expected reward (cost) of arm i
λ	uniformly lower bound of expected cost
ρ_i, ρ^*	expected-reward to expected-cost ratio of arm i and maximum ratio
β	parameter of ratio distribution
Δ_k	difference between ρ^* and ρ_k
$T_i(t-1)$	arm i's pulling number before round t
I_t	the arm pulled at t-th round
K, K_t	randomly chosen arms (at time t)
$\mathcal{E}_t$	exploration sequence at round t
$\mathcal{D}_{i,T_i(t-1),t}$	arm i's estimated ratio index at round t
R^*	supremum of expected reward of all possible pulling algorithms
$\mathcal{R}^{\text{alg}}, \mathcal{R}^{\text{RCB-I}}, \mathcal{R}^{\text{RCB-AIR}}$	regret of any algorithm, RCB-I and RCB-AIR algorithm
τ_B	speudo stopping time

Table 1: Notations: *We summarize the notations used in our paper in this table.*

Acknowledgments

The authors wish to thank Tie-Yan Liu, Wei Chen, Tao Qin and Wensheng Zhang for helpful discussions related to this work.

References

Agmon Ben-Yehuda, O.; Ben-Yehuda, M.; Schuster, A.; and Tsafrir, D. 2013. Deconstructing amazon ec2 spot instance pricing. *ACM Transactions on Economics and Computation* 1(3):16.

Amin, K.; Kearns, M.; Key, P.; and Schwaighofer, A. 2012. Budget optimization for sponsored search: Censored learning in mdps. *Eprint Arxiv*.

Ardagna, D.; Panicucci, B.; and Passacantando, M. 2011. A game theoretic formulation of the service provisioning problem in cloud systems. In *WWW*, 177–186. ACM.

Audibert, J.-Y.; Munos, R.; and Szepesvári, C. 2007. Tuning bandit algorithms in stochastic environments. In *International Conference on Algorithmic Learning Theory*. Springer.

Audibert, J.-Y.; Munos, R.; and Szepesvári, C. 2009. Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science* 410(19):1876–1902.

Auer, P.; Cesa-Bianchi, N.; and Fischer, P. 2002. Finite-time analysis of the multiarmed bandit problem. *Machine learning* 47(2-3):235–256.

Badanidiyuru, A.; Kleinberg, R.; and Slivkins, A. 2013. Bandits with knapsacks. In *FOCS*, 207–216. IEEE.

Badanidiyuru, A.; Langford, J.; and Slivkins, A. 2014. Resourceful contextual bandits. In *COLT*, 1109–1134.

Berry, D. A.; Chen, R. W.; Zame, A.; Heath, D. C.; and Shepp, L. A. 1997. Bandit problems with infinitely many arms. *The Annals of Statistics* 2103–2116.

Beygelzimer, A.; Langford, J.; Li, L.; Reyzin, L.; and Schapire, R. E. 2011. Contextual bandit algorithms with supervised learning guarantees. In *AISTATS*, 19–26.

Bubeck, S., and Cesa-Bianchi, N. 2012. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *arXiv preprint arXiv:1204.5721*.

Bubeck, S. 2010. *Bandits games and clustering foundations*. Ph.D. Dissertation, Université des Sciences et Technologie de Lille-Lille I.

Carpentier, A., and Valko, M. 2015. Simple regret for infinitely many armed bandits. *arXiv preprint arXiv:1505.04627*.

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

Garivier, A., and Cappé, O. 2011. The kl-ucb algorithm for bounded stochastic bandits and beyond. In *COLT*, 359–376.

György, A.; Kocsis, L.; Szabó, I.; and Szepesvári, C. 2007. Continuous time associative bandit problems. In *IJCAI*, 830–835.

Kaufmann, E.; Cappé, O.; and Garivier, A. 2012. On bayesian upper confidence bounds for bandit problems. In *AISTATS*, 592–600.

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

Li, H., and Xia, Y. 2016. Infinitely many-armed bandits with budget constraints. https://goo.gl/bCuuS5.

Li, L.; Chu, W.; Langford, J.; and Schapire, R. E. 2010. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, 661–670. ACM.

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

Martello, S., and Toth, P. 1990. *Knapsack problems: algorithms and computer implementations*. John Wiley & Sons, Inc.

Mohri, M., and Munoz, A. 2014. Optimal regret minimization in posted-price auctions with strategic buyers. In *NIPS*, 1871–1879.

Tran-Thanh, L.; Chapman, A.; de Cote, E. M.; Rogers, A.; and Jennings, N. R. 2010. Epsilon–first policies for budget–limited multi-armed bandits. In *Twenty-Fourth AAAI Conference on Artificial Intelligence*.

Tran-Thanh, L.; Chapman, A.; Rogers, A.; and Jennings, N. R. 2012. Knapsack based optimal policies for budget–limited multi–armed bandits. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*.

Tran-Thanh, L.; Stavrogiannis, L. C.; Naroditskiy, V.; Robu, V.; Jennings, N. R.; and Key, P. 2014. Efficient regret bounds for online bid optimisation in budget-limited sponsored search auctions.

Vanchinathan, H. P.; Marfurt, A.; Robelin, C.-A.; Kossmann, D.; and Krause, A. 2015. Discovering valuable items from massive data. In *KDD*. ACM.

Wang, Y.; Audibert, J.-Y.; and Munos, R. 2009. Algorithms for infinitely many-armed bandits. In *NIPS*, 1729–1736.

Wu, H.; Srikant, R.; Liu, X.; and Jiang, C. 2015. Algorithms with logarithmic or sublinear regret for constrained contextual bandits. In *NIPS*.

Xia, Y.; Ding, W.; Zhang, X.-D.; Yu, N.; and Qin, T. 2015a. Budgeted bandit problems with continuous random costs. In *ACML*.

Xia, Y.; Li, H.; Qin, T.; Yu, N.; and Liu, T. Y. 2015b. Thompson sampling for budgeted multi-armed bandits. In *International Conference on Artificial Intelligence*.

Xia, Y.; Qin, T.; Ma, W.; Yu, N.; and Liu, T.-Y. 2016a. Budgeted multi-armed bandits with multiple plays. In *IJCAI*.

Xia, Y.; Qin, T.; Yu, N.; and Liu, T.-Y. 2016b. Best action selection in a stochastic environment. In *AAMAS*, 758–766.

Zhou, Y.; Chen, X.; and Li, J. 2014. Optimal pac multiple arm identification with applications to crowdsourcing. In *ICML*, 217–225.