Multi-Armed Bandit with Budget Constraint and Variable Costs

Wenkui Ding [*], **Tao Qin**[†], **Xu-Dong Zhang**[‡] and **Tie-Yan Liu**[†]
[*]Department of Electronic Engineering, Tsinghua University, Beijing, 100084, P.R.China
[†]Microsoft Research Asia, Building 2, No. 5 Danling Street, Haidian District, Beijing, 100080, P.R.China
[*]dingwenkui@gmail.com, [†]{taoqin,tyliu}@microsoft.com, [‡]zhangxd@tsinghua.edu.cn

Abstract

We study the multi-armed bandit problems with budget constraint and variable costs (MAB-BV). In this setting, pulling an arm will receive a random reward together with a random cost, and the objective of an algorithm is to pull a sequence of arms in order to maximize the expected total reward with the costs of pulling those arms complying with a budget constraint. This new setting models many Internet applications (e.g., ad exchange, sponsored search, and cloud computing) in a more accurate manner than previous settings where the pulling of arms is either costless or with a fixed cost. We propose two UCB based algorithms for the new setting. The first algorithm needs prior knowledge about the lower bound of the expected costs when computing the exploration term. The second algorithm eliminates this need by estimating the minimal expected costs from empirical observations, and therefore can be applied to more real-world applications where prior knowledge is not available. We prove that both algorithms have nice learning abilities, with regret bounds of $O(\ln B)$. Furthermore, we show that when applying our proposed algorithms to a previous setting with fixed costs (which can be regarded as our special case), one can improve the previously obtained regret bound. Our simulation results on real-time bidding in ad exchange verify the effectiveness of the algorithms and are consistent with our theoretical analysis.

Introduction

The multi-armed bandit (MAB) problems have been studied extensively in the literature (Agrawal, Hedge, and Teneketzis 1988; Auer, Cesa-Bianchi, and Fischer 2002; Auer 2003; Kleinberg 2004; Chu et al. 2011) because they provide a principled way to model sequential decision making in an uncertain environment. MAB algorithms have been widely adopted in real applications including adaptive routing (Awerbuch and Kleinberg 2004), online advertising (Chakrabarti et al. 2008; Babaioff, Sharma, and Slivkins 2009), and job finding in labor markets (Berry et al. 1997).

Most of the aforementioned research assumes that pulling an arm is costless. However, in reality, pulling an arm (taking an action) is usually costly and the total cost is constrained by a budget. Examples include the real-time bidding problem in ad exchange (Chakraborty et al. 2010), the bid optimization problem in sponsored search (Borgs et al. 2007), the on-spot instance bidding problem in Amazon EC2 (Zaman and Grosu 2010; Ben-Yehuda et al. 2011), and the cloud service provider selection problem in IaaS (Ardagna, Panicucci, and Passacantando 2011). For instance, in the real-time bidding problem in ad exchange, the bid of an ad is the arm and the received impression (or click) is the reward. The objective of an advertiser is to maximize the total impressions (or clicks) of his/her ad, by appropriately setting the bid. In this example, there is a cost associated with each pulling of an arm, which is the per-impression (or per-click) payment to the ad exchange by the advertiser. Furthermore, the total cost is constrained by a budget set by the advertiser for each of his/her ad campaign in advance.

Recently, two types of MAB problems with budget constraint have been studied. In the first type (Audibert, Bubeck, and others 2010; Bubeck, Munos, and Stoltz 2009; Guha and Munagala 2007), pulling each arm in the exploration phase has a unit cost and the budget is only imposed on the exploration arms. Instead, the exploitation arms are not associated with costs or constrained by budgets. The goal is to find the best arm given the budget constraint on the total number of exploration arms. This type of problems is also referred to as “best arm identification” or “pure exploration problem”. In the second type, pulling an arm is always associated with a cost and constrained by a budget, no matter in the exploration phase or the exploitation phase. Therefore, it better describes the aforementioned Internet applications. However, the attempt on this type of problems is quite limited. As far as we know, the only related work is (Tran-Thanh et al. 2010; 2012), and it studies the case where the cost of pulling an arm is fixed and becomes known after the arm is pulled once. For ease of reference, we call the setting *Multi-Armed Bandit problems with Budget constraint and Fixed costs (MAB-BF)*. It is noted that many real applications are more complex than this fixed-cost setting: the cost may vary from time to time even if the same arm is pulled. For example, in the real-time bidding problem in ad exchange, the payment (the cost of pulling an arm) is a function of many random factors (including the click behaviors of the users and the bidding behaviors of other advertisers) and should better be considered as a random variable rather than a fixed

[*]This work was done when the first author was visiting Microsoft Research Asia.

quantity (Feldman et al. 2010). In this variable-cost setting, we need to explore not only the reward of an arm but also its cost. To the best of our knowledge, this new setting has never been discussed in the literature, and our paper is the first work that formally investigates it. For ease of reference, we call this new setting *Multi-Armed Bandit problems with Budget constraint and Variable costs (MAB-BV)*.

In this paper, we design two upper confidence bound (UCB) (Agrawal, Hedge, and Teneketzis 1988) based algorithms to solve the MAB-BV problems: UCB-BV1 and UCB-BV2. Similar to previous UCB based algorithms, we consider an exploitation term and an exploration term while computing the index of an arm. The two proposed algorithms share the same exploitation term which is based on the observed rewards and costs of arms. They differ in the detailed way of computing the exploration terms. UCB-BV1 assumes that the minimum of the expected costs of all the arms are known as prior knowledge. Considering that the prior knowledge may be unavailable in some applications, UCB-BV2 eliminate this need by estimating the minimal average of their empirical observations.

We then analyze the regret bounds of the two proposed algorithms. This task turns out to be difficult due to the following reasons. First, to get the optimal (oracle) pulling policy, we need to solve a stochastic optimization problem. Second, the stopping time of the algorithms cannot be easily characterized due to the variable costs. To tackle the challenges, we develop a set of new proof techniques, and derive regret bounds of $O(\ln B)$ for the algorithms. Furthermore, we see a trade-off between regret bound and application scope of the algorithms: the UCB-BV2 algorithm has a looser bound but a broader application scope than UCB-BV1 because the latter requires additional prior knowledge.

It is noted that the MAB-BV problems under our investigation are generalizations of the MAB-BF problems (Tran-Thanh et al. 2010). Therefore, our proposed algorithms and theorems can also be applied to that setting and we obtain a better regret bound for the fixed-cost setting.

Problem Setup

In this section, we describe the setting of multi-armed bandit problems with budget constraint and variable costs (MAB-BV).

Similar to the classical K armed bandit problems, a bandit has K arms in the MAB-BV problems, and at time t, pulling arm i returns a reward $r_{i,t}$ with support $[0,1]$. Rewards $r_{i,1}, r_{i,2}, \cdots$ are random variables independently and identically sampled according to a probability distribution with expectation μ_i^r. Different from the classical MAB problems, at time t, pulling arm i is associated with a cost $c_{i,t}$, and costs $c_{i,1}, c_{i,2}, \cdots$ are random variables independently and identically distributed with expectation μ_i^c. We assume the cost takes discrete values from the set $\{\frac{k}{m}\}_{k=0}^m$ where $\frac{1}{m}$ is the unit cost and the largest cost is normalized to 1.[1]

We use B to denote the budget, which is an integral multiple of the unit cost $\frac{1}{m}$. The budget will constrain the total number of pulls (or the stopping time of the pulling procedure). That is, the stopping time, $t_a(B)$, of a pulling algorithm a is a random variable depending on B and can be characterized as follows:

$$\sum_{t=1}^{t_a(B)} c_{a_t,t} \leq B < \sum_{t=1}^{t_a(B)+1} c_{a_t,t}, \qquad (1)$$

where a_t is the index of the arm pulled by algorithm a at time t.

The total reward collected up to time $t_a(B)$ by the pulling algorithm a is defined as $R_a = \sum_{t=1}^{t_a(B)} r_{a_t,t}$. For mathematical convenience, we consider rewards collected up to time $t_a(B)$ but not $t_a(B)+1$. The expected total reward is $E[R_a] = E\left[\sum_{t=1}^{t_a(B)} r_{a_t,t}\right]$, where the expectation is taken over the randomness of rewards and costs, and possibly the algorithm.

The optimal total expected reward when knowing the distribution of all random variables is denoted as R^*. Then the expected regret of the algorithm a can be defined as below,

$$R^* - E[R_a] = R^* - E\left[\sum_{t=1}^{t_a(B)} r_{a_t,t}\right]. \qquad (2)$$

Note that the expected regret in the MAB-BV problems depends on B, but not on t as in the classical MAB problems. This is because in our setting B is the independent variable and $t_a(B)$ is a dependent random variable induced from B.

Our proposed MAB-BV problems can be used to describe many Internet applications. For one example, in the real-time bidding problem in ad exchange, an advertiser is given a budget and required to sequentially choose suitable bids for his/her ad to maximize his/her payoff. In this example, a bid corresponds to an arm; the received impression (or click) and the per-impression (or per-click) payment correspond to the reward and cost of an arm respectively. The payment is usually determined by many (random) factors, including the bids of other advertisers and the user behaviors and profiles. Therefore the cost should be regarded as a random variable.

For another example, in many sequential resource allocation problems like cloud computing (Wei et al. 2010), a user has a budget and performs sequential selections of virtual machines to maximize his/her overall profit. Here each virtual machine corresponds to an arm; the computational resource from the selected virtual machine and the charged price correspond to the reward and cost respectively. Again, the charged price depends on many factors such as the time of the day, the energy cost, and the combativeness in the virtual machine (Ardagna, Panicucci, and Passacantando 2011). Therefore, the cost is also a random variable in this example.

Algorithms

In this section, we design two upper confidence bound (UCB) based algorithms for the MAB-BV problems, as shown in the following table.

In the table, $n_{i,t}$ is the times that arm i has been pulled before step t, $\bar{r}_{i,t}$ is the average reward of arm i before step

[1] $\frac{1}{m}$ can be the minimal unit of the currency, such as 1 cent; then $\frac{1000}{m}$ is 10 dollar.

Algorithm 1 UCB-BV1/UCB-BV2

```
Initialization: Pull each arm $i$ once in the first $K$ steps, set $t = K$.
while $\sum_{s=1}^{t} c_{a_s,s} \le B$ do
    Set $t = t + 1$.
    Calculate the index $D_{i,t}$ of each arm $i$ as follows.
    UCB-BV1
        $D_{i,t} = \frac{\bar{r}_{i,t}}{\bar{c}_{i,t}} + \frac{(1+\frac{1}{\lambda})\sqrt{\frac{\ln(t-1)}{n_{i,t}}}}{\lambda - \sqrt{\frac{\ln(t-1)}{n_{i,t}}}}$
    UCB-BV2
        $D_{i,t} = \frac{\bar{r}_{i,t}}{\bar{c}_{i,t}} + \frac{1}{\lambda_t}\left(1 + \frac{1}{\lambda_t - \sqrt{\frac{\ln(t-1)}{n_{i,t}}}}\right)\sqrt{\frac{\ln(t-1)}{n_{i,t}}}$
    Pull the arm $a_t$ with the largest index: $a_t = \arg\max_i D_{i,t}$.
end while
Return: $t_B = t - 1$.
```

t, and $\bar{c}_{i,t}$ is the average cost. The definitions of $n_{i,t}$, $\bar{r}_{i,t}$, and $\bar{c}_{i,t}$ are given as follows:

$$n_{i,t} = \sum_{s=1}^{t-1} I(a_s = i),$$

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

where $I(\cdot)$ is the indicator function.

The major difference between UCB-BV1/UCB-BV2 and conventional UCB algorithms lies in that our proposed algorithms will stop when running out of budget (i.e., the budget determines the stopping time of the algorithms) but there is no explicit stopping time in the conventional UCB algorithms. This difference comes from the nature of the MAB-BV problems, but not because of our specific proposal of the algorithms.

As can be seen, the two algorithms share the same exploitation term, which is the average reward-to-cost ratio. This term forces the algorithms to choose those arms with higher marginal rewards. The differences between the two algorithms lie in the exploration term.

- There is a parameter λ in UCB-BV1, which is the prior knowledge characterizing the lower bound of the expected costs:

  $$\lambda \le \min_i \mu_i^c.$$

  In some applications like on-spot bidding in cloud computing, this prior knowledge can be easily obtained from the service provider (Ardagna, Panicucci, and Passacantando 2011). But in some other applications, it might be hard to obtain such knowledge. Therefore, the application scope of UCB-BV1 is limited.
- UCB-BV2 estimates the minimum of expectations of costs by using their empirical observations:

  $$\lambda_t = \min_i \bar{c}_{i,t},$$

  and then uses the estimate to compute the exploration term. Therefore, this algorithm does not require the prior knowledge as UCB-BV1 does and can be applied to more applications.

One may have noticed that UCB-BV2 is not obtained by simply replacing λ in UCB-BV1 with λ_t. It is because this simple replacement does not lead to a reasonable regret bound. Instead, the formula used in our proposed UCB-BV2 algorithm has much better theoretical properties (as shown in the next section).

Regret Analysis

As mentioned in the previous section, our proposed algorithms can be regarded as natural extensions of classical UCB algorithms, and therefore one may think that the analysis on their regret bounds will also be similar. However, we would like to point out that the regret analysis for our algorithms is much more difficult due to the following reasons.

First, because of the budget constraint, the optimal (oracle) policy for MAB-BV is no longer to repeatedly pull the arm with the largest expected reward (or reward-to-cost ratio) as in the UCB algorithms. Furthermore, because the costs are random variables, we cannot obtain the optimal policy by solving a simple knapsack problem as in the setting of MAB-BF (Tran-Thanh et al. 2010). In our case, we will have to solve a constrained stochastic optimization problem in order to obtain the optimal policy. We discuss this issue and derive an upper bound for the objective function of this stochastic optimization problem in Lemma 1.

Second, due to the budget constraint, there will be a stopping time $t_a(B)$ for arm pulling, which will affect the regret bound. When the cost is fixed (i.e., in the setting of MAB-BF), $t_a(B)$ can be obtained by taking the expectation on Eqn. (1) and decomposing the overall cost into the costs for distinct arms (this is because in this setting the cost for an arm is the same whenever the arm is pulled). However, in our case, $t_a(B)$ and $c_{a_t,t}$ are dependent random variables, so we cannot obtain the stopping time by directly taking the expectation on Eqn. (1). To tackle the challenge, we construct a single-armed bandit process (with budget constraint and unknown costs) and obtain its stopping time by means of induction (see Lemma 2), and then bound the stopping time of the original MAB-BV problems on this basis.

In the following subsections, we first describe how we tackle the aforementioned difficulties, and then derive the regret bounds for the proposed algorithms.

Reward of the Optimal Policy

The following lemma upper bounds the reward of the optimal policy. In the lemma, i^* denotes the arm with the largest expected reward-to-cost ratio: $i^* = \arg\max_i \frac{\mu_i^r}{\mu_i^c}$.

Lemma 1. *Consider the following stochastic optimization problem*

$$\max_a E\left[\sum_{t=1}^{t_a(B)} r_{a_t,t}\right], \qquad \textit{s.t.} \sum_{t=1}^{t_a(B)} c_{a_t,t} \leq B$$

where $r_{i,1}, r_{i,2}, \cdots$ are non-negative, i.i.d random variables with expectation μ_i^r and $c_{i,1}, c_{i,2}, \cdots$ are non-negative, i.i.d random variables with expectation μ_i^c. Its optimum is upper bounded by $\frac{\mu_{i^*}^r}{\mu_{i^*}^c}(B+1)$.

Proof sketch. Denote $R(B)$ as the optimum of the optimization problem. When $-1 \leq B \leq 0$, the result holds trivially. When $B = \frac{k}{m} > 0$, given $\forall k' < k, R(\frac{k'}{m}) \leq \frac{\mu_{i^*}^r}{\mu_{i^*}^c}(\frac{k'}{m} + 1)$, we need to prove $R(\frac{k}{m}) \leq \frac{\mu_{i^*}^r}{\mu_{i^*}^c}(\frac{k}{m} + 1)$. Consider the first step of the optimal policy. No matter how the optimal policy operates at the first step, we can assume that it pulls arm i with a probability q_i and $\sum_i q_i = 1$ and let $p_{i,j}$ denote the probability of the cost of the arm i to be $\frac{j}{m}$. [2] Then we get

$$R\left(\frac{k}{m}\right) \leq \sum_{i=1}^{K} q_i \sum_{j=1}^{m} p_{i,j} \left(E\left[r_i \middle| c_i = \frac{j}{m}\right] + R\left(\frac{k-j}{m}\right)\right)$$

$$\leq \sum_{i=1}^{K} q_i \left(\mu_i^r + \frac{\mu_{i^*}^r}{\mu_{i^*}^c} \sum_{j=1}^{m} p_{i,j} \left(\frac{k-j}{m} + 1\right)\right) \leq \frac{\mu_{i^*}^r}{\mu_{i^*}^c}\left(\frac{k}{m} + 1\right).$$

□

Stopping Time of UCB-BV Algorithms

The following lemma characterizes the expected stopping time of the proposed UCB-BV algorithms.

Lemma 2. *Consider an MAB-BV problem and denote $n_{i,t(B)}$ as the times that arm i has been pulled. If for $\forall i \neq i^*, \exists \delta_i > 0, \rho_i > 0$, s.t. $E[n_{i,t(B)}|t(B)] \leq \delta_i \ln t(B) + \rho_i$, then we have*

$$E[t(B)] \leq \frac{B+1}{\mu_{i^*}^c} + \delta \ln 2\left(\frac{B+1}{\mu_{i^*}^c} + \delta\ln(2\delta) + \rho\right) + \rho, \tag{3}$$

$$E[t(B)] > \frac{B-\rho}{\mu_{i^*}^c} - \frac{\delta}{\mu_{i^*}^c}\ln 2\left(\frac{B+1}{\mu_{i^*}^c} + \delta\ln(2\delta) + \rho\right) - 1, \tag{4}$$

where $\delta = \sum_{i \neq i^*} \delta_i, \rho = \sum_{i \neq i^*} \rho_i$.

Proof sketch. We first prove that for a single-armed bandit whose arm has the same reward and cost distributions as the *i-th* arm of the original multi-armed bandit, the stopping time $s_i(B)$ of this single-armed bandit process satisfies the following inequality.

$$\frac{B}{\mu_i^c} - 1 < E[s_i(B)] \leq \frac{B+1}{\mu_i^c} \tag{5}$$

Eqn. (5) holds trivially for $-1 \leq B \leq 0$. When $B = \frac{k}{m} > 0$, given Eqn. (5) holds for $\forall k' < k$, then it also holds for B, we then prove the correctness Eqn. (5) for $B = \frac{k}{m}$. Consider the first step of the process. Let p_j denote the probability to induce a cost equal to $\frac{j}{m}$, then the remaining budget is $\frac{k-j}{m}$; the expected stopping time obtained by the remaining budget equals $s_i(\frac{k-j}{m})$. So $s_i(B)$ can be expressed in the following recursive formula.

$$E[s_i(B)] = \sum_{j=1}^{\min\{k,m\}} p_j(1 + E[s_i(\frac{k-j}{m})])$$

(1) If $k \leq m$, we get $E[s_i(B)] > \sum_{j=1}^{k_B} \frac{p_j(B-\frac{j}{m})}{\mu_i^c} = \frac{B}{\mu_i^c} - \sum_{j=k_B+1}^{m} \frac{p_j}{\mu_i^c} B - \sum_{j=1}^{k_B} \frac{p_j}{\mu_i^c}\frac{j}{m} \geq \frac{B}{\mu_i^c} - \sum_{j=1}^{m} \frac{p_j}{\mu_i^c}\frac{j}{m} = \frac{B}{\mu_i^c} - 1$;
(2) if $k > m$, we get $E[s_i(B)] > \sum\limits_{j=1}^{m} \frac{p_j}{\mu_i^c}(B - \frac{j}{m}) = \frac{B}{\mu_i^c} - 1$;
(3) Further, we have $E[s_i(B)] \leq \sum\limits_{j=1}^{m} p_j(1 + \frac{B-\frac{j}{m}+1}{\mu_i^c}) = \frac{B+1}{\mu_i^c}$.

Combining the above results and the conditions given in the theorem for the multi-armed bandit process, we can get the following two inequalities.

$$\begin{aligned} E[t(B)] &\leq E[s_{i^*}(B)] + \sum_{i \neq i^*} \delta_i E[\ln t(B)] + \sum_{i \neq i^*} \rho_i \\ &\leq \frac{B+1}{\mu_{i^*}^c} + \delta E[\ln t(B)] + \rho \\ E[t(B)] &\geq E[s_{i^*}(B - \sum_{i \neq i^*} \delta_i \ln t(B) - \sum_{i \neq i^*} \rho_i)] \\ &> \frac{B - \delta E[\ln t(B)] - \rho}{\mu_{i^*}^c} - 1 \end{aligned}$$

By substituting the below inequality into the above two inequalities, we prove the two inequalities in the theorem.

$$E[\ln t(B)] \leq \frac{E[t(B)]}{2\delta} + \ln(2\delta) - 1$$

□

Regret Bounds for UCB-BV Algorithms

Based on the results obtained in the previous subsections, we can obtain the following theorem which gives regret bounds to our proposed UCB-BV algorithms. We use $\Delta_i = \frac{\mu_{i^*}^r}{\mu_{i^*}^c} - \frac{\mu_i^r}{\mu_i^c}$ in the following theorem.

Theorem 1. *The expected regret of the UCB-BV algorithms ($\lambda \leq \min_i \mu_i^c$ for UCB-BV1) is at most*

$$\mathbb{R}(\delta,\rho) = \alpha \ln\big(\frac{B+1}{\mu_{i^*}^c} + \delta\ln 2\big(\frac{B+1}{\mu_{i^*}^c} + \delta\ln(2\delta) + \rho\big) + \rho\big)$$
$$+ \frac{\mu_{i^*}^r}{\mu_{i^*}^c}\big(\delta \ln 2\big(\frac{B+1}{\mu_{i^*}^c} + \delta\ln(2\delta) + \rho\big) + \rho + \mu_{i^*}^c + 1\big) + \beta, \tag{6}$$

where $\alpha = \sum_{i:\mu_i^r < \mu_{i^*}^r} \delta_i(\mu_{i^*}^r - \mu_i^r)$, $\beta = \sum_{i:i \neq i^*} \rho_i(\mu_i^r - \mu_{i^*}^r)$, $\delta = \sum_{i:i \neq i^*} \delta_i$, $\rho = \sum_{i:i \neq i^*} \rho_i$, for UCB-BV1 $\delta_i = (\frac{2+\frac{2}{\lambda}+\Delta_i}{\Delta_i \lambda})^2$ and $\rho_i = 2(1 + \frac{\pi^2}{3})$, for UCB-BV2 $\delta_i = (\frac{2+\frac{2}{\lambda}+3\Delta_i}{\Delta_i \lambda})^2$ and $\rho_i = 3(1 + \frac{\pi^2}{3})$.

[2] If the optimal policy is deterministic and pulls arm k at the first step, then $q_k = 1$ and $q_i = 0, \forall i \neq k$.

Proof. Here we give the proof for UCB-BV1, the proof for UCB-BV2 can be found in the supplementary document.

First, we want to prove the pulling number of any suboptimal arm i can be bounded as follows

$$E[n_{i,t_a(B)}|t_a(B)] \leq \left(\frac{2+\frac{2}{\lambda}+\Delta_i}{\Delta_i\lambda}\right)^2 \ln t_a(B)+2(1+\frac{\pi^2}{3}).$$

Define $T = \left(\frac{2+\frac{2}{\lambda}+\Delta_i}{\Delta_i\lambda}\right)^2 \ln t_a(B)$. Given $t_a(B)$, there are two cases for $n_{i,t_a(B)}$: (a) $n_{i,t_a(B)} < T$; (b) $n_{i,t_a(B)} \geq T$. For the first case, it can be trivially obtained that $E[n_{i,t_a(B)}|t_a(B), n_{i,t_a(B)} < T] \leq T$. Next we consider the second case.

$$\begin{aligned}
& E[n_{i,t_a(B)}|t_a(B), n_{i,t_a(B)} \geq T] \\
= \ & E\left[\sum_{t=K+1}^{t_a(B)} I(a_t = i) \, \middle| t_a(B), n_{i,t_a(B)} \geq T\right] + 1 \\
\leq \ & \sum_{t=K+1}^{t_a(B)} P\left(\frac{\bar{r}_{i,t}}{\bar{c}_{i,t}} + E_{i,t} \geq \frac{\bar{r}_{i^*,t}}{\bar{c}_{i^*,t}} + E_{i^*,t}, n_{i,t} \geq T\right) + T \\
\leq \ & P\left(\frac{\bar{r}_{i,t}}{\bar{c}_{i,t}} \geq \frac{\mu_i^r}{\mu_i^c} + E_{i,t}\right) + P\left(\frac{\bar{r}_{i^*,t}}{\bar{c}_{i^*,t}} \leq \frac{\mu_{i^*}^r}{\mu_{i^*}^c} - E_{i^*,t}\right) \\
& + P\left(\frac{\mu_{i^*}^r}{\mu_{i^*}^c} < \frac{\mu_i^r}{\mu_i^c} + 2E_{i,t}, n_{i,t} \geq N\right) + T \qquad (7)
\end{aligned}$$

where $E_{i,t}$ is the exploration term in UCB-BV1.

Note that $\frac{\bar{r}_{i,t}}{\bar{c}_{i,t}} \geq \frac{\mu_i^r}{\mu_i^c} + E_{i,t}$ implies at least one of the following two events happens

$$\bar{r}_{i,t} \geq \mu_i^r + \epsilon_{i,t}, \qquad \bar{c}_{i,t} \leq \mu_i^c - \epsilon_{i,t},$$

where $\epsilon_{i,t} = \sqrt{\frac{\ln t}{n_{i,t}}}$. Otherwise,

$$\begin{aligned}
& \frac{\bar{r}_{i,t}}{\bar{c}_{i,t}} - \frac{\mu_i^r}{\mu_i^c} = \frac{(\bar{r}_{i,t} - \mu_i^r)\mu_i^c + (\mu_i^c - \bar{c}_{i,t})\mu_i^r}{\bar{c}_{i,t}\mu_i^c} \\
< \ & \frac{\epsilon_{i,t}}{\bar{c}_{i,t}} + \frac{\epsilon_{i,t}\mu_i^r}{\bar{c}_{i,t}\mu_i^c} \leq \frac{\epsilon_{i,t}}{\lambda - \epsilon_{i,t}} + \frac{\epsilon_{i,t}}{(\lambda - \epsilon_{i,t})\lambda} = E_{i,t}.
\end{aligned}$$

Using Chernoff-Hoeffding inequality, we have

$$P(\bar{r}_{i,t} \geq \mu_i^r + \epsilon_{i,t}) \leq \exp(-2\epsilon_{i,t}^2 n_{i,t}) = t^{-2},$$
$$P(\bar{c}_{i,t} \leq \mu_i^c - \epsilon_{i,t}) \leq \exp(-2\epsilon_{i,t}^2 n_{i,t}) = t^{-2}.$$

So $$P\left(\frac{\bar{r}_{i,t}}{\bar{c}_{i,t}} \geq \frac{\mu_i^r}{\mu_i^c} + E_{i,t}\right) \leq \sum_{t=1}^{\infty} P(\bar{r}_{i,t} \geq \mu_i^r + \epsilon_{i,t})$$
$$+\sum_{t=1}^{\infty} P(\bar{c}_{i,t} \leq \mu_i^c - \epsilon_{i,t}) \leq 2\sum_{t=1}^{\infty} t^{-2} = 1 + \frac{\pi^2}{3}. \qquad (8)$$

Similarly, $$P\left(\frac{\bar{r}_{i,t}}{\bar{c}_{i,t}} \leq \frac{\mu_i^r}{\mu_i^c} - E_{i^*,t}\right) \leq 1 + \frac{\pi^2}{3}. \qquad (9)$$

It can be verified that $\forall n_{i,t} \geq T$, we have

$$\lambda > \sqrt{\frac{\ln t}{N}} \quad \text{and} \quad E_{i,t} \leq \frac{(1+\frac{1}{\lambda})\sqrt{\frac{\ln t}{N}}}{\lambda - \sqrt{\frac{\ln t}{N}}} \leq \frac{\Delta_i}{2}.$$

So we have

$$P\left(\frac{\mu_{i^*}^r}{\mu_{i^*}^c} < \frac{\mu_i^r}{\mu_i^c} + 2E_{i,t}, n_{i,t} \geq N\right) = 0. \qquad (10)$$

Combining Eqn (7) $\sim$ Eqn (10), we obtain

$$E[n_{i,t_a(B)}|t_a(B), n_{i,t_a(B)} \geq T] \leq T + 2(1 + \frac{\pi^2}{3})$$

for the second case.

Applying Lemma 1 we obtain the regret bound as follows.

$$\begin{aligned}
R^* - E[R_a] \ \leq \ & \left(\frac{\mu_{i^*}^r}{\mu_{i^*}^c}(B+1) - \mu_{i^*}^r E[t_a(B)]\right) \\
+ \ & \left(\mu_{i^*}^r E[t_a(B)] - E\Big[\sum_{t=1}^{t_a(B)} r_{a_t,t}\Big]\right) \qquad (11)
\end{aligned}$$

The first term on the r.h.s of Eqn. (11) can be bounded by applying Lemma 2, where $\delta = \sum_{i \neq i^*}(\frac{2+\frac{2}{\lambda}+\Delta_i}{\Delta_i\lambda})^2$ and $\rho = 2(K-1)(1+\frac{\pi^2}{3})$. The second term on the r.h.s of Eqn. (11) can be bounded similar to traditional MAB algorithms: $\mu_{i^*}^r E[t_a(B)] - E\big[\sum_{t=1}^{t_a(B)} r_{a_t,t}\big] = E\left[\sum_{i\neq i^*} n_{i,t_a(B)}(\mu_{i^*}^r - \mu_i^r)\right]$.

The regret bound for UCB-BV2 can be obtained in a similar manner. The main difference lies in the use of a double sided Chernoff-Hoeffding inequality for the cost and a delicately designed N to ensure $\lambda > 2\sqrt{\frac{\ln t}{N}}$ and $2E_{i,t} \leq \Delta_i$. □

From the above theorem we can see that both the two algorithm can achieve a regret bound of $O(\ln B)$. Further, UCB-BV1 has a tighter bound than UCB-BV2. This is easy to understand because it leverages additional prior knowledge about the expected costs. However, this also limits the application scope of UCB-BV1 as discussed in the introduction.

Discussions

It is clear that the MAB-BV problems are more general than the MAB-BF problems studied in (Tran-Thanh et al. 2010) where the cost is regarded as a fixed quantity. Actually we can directly apply our proposed algorithms to solve the MAB-BF problems. Due to space limitations, we only make discussions on UCB-BV1 for the MAB-BF problems as below. The discussions on UCB-BV2 are very similar.

In the MAB-BF problems, after pulling each arm once, we can know the value of μ_i^c.[3] With this information, it is not difficult to compute the index $D_{i,t}$ of UCB-BV1 as

$$D_{i,t} = \frac{\bar{r}_{i,t}}{\mu_i^c} + \frac{(1+\frac{1}{\min_j \mu_j^c})\sqrt{\frac{\ln(t-1)}{n_{i,t}}}}{\min_j \mu_j^c - \sqrt{\frac{\ln(t-1)}{n_{i,t}}}}.$$

The following corollary shows that UCB-BV1 can achieve a regret bound of $O(\ln B)$ when applied to solve the MAB-BF problems.

[3] Please note that here we reuse the symbol μ_i^c as the fixed cost of pulling arm i.

Corollary 1. *The expected regret of UCB-BV1 for the MAB-BF problems is no larger than* $\mathbb{R}(\delta, \rho)$ *defined in Eqn. (6) with* $\delta_i = (\frac{2+\frac{2}{\min_j \mu_j^c}+\Delta_i}{\Delta_i \min_j \mu_j^c})^2$ *and* $\rho_i = 2(1+\frac{\pi^2}{3})$.

One can see that the bound of $O(\ln B)$ given in the above corollary is better than the bound of $O(B^{\frac{2}{3}})$ obtained in (Tran-Thanh et al. 2010) when B is large.

Numerical Simulations

In the previous section, we have obtained theoretical regret bounds of $O(\ln(B))$ for both UCB-BV1 and UCB-BV2. The bounds guarantee that the two algorithms perform well when the budget B is sufficiently large. However, it remains unclear whether the algorithms will perform well when it is not the case. In this section, we conducted some experimental investigations on this issue, taking real-time bidding in ad exchange as the target application.

The setting of our simulation is as follows. (1) A bandit with ten arms is created, which means an advertiser can have ten choices for his bidding price of his ad. (2) To simulate real-time bidding in ad exchange, the reward of each pulling of an arm is sampled from a Bernoulli distribution: 0 means the advertiser does not receive an impression (or click) by using a specific bidding price (the pulled arm); we use the expected reward of the arm as the mean of the Bernoulli distribution. The cost of pulling an arm is randomly sampled from $\{0, 1/100, 2/100, 3/100, \cdots, 100/100\}$ according to a multinomial distribution. (3) We set the budget as 100, 200, ..., and up to 10000. For each value of budget, we run our proposed two MAB-BV algorithms for 100 times and check their average performances. (4) For comparison purpose, we implemented UCB1 (Auer, Cesa-Bianchi, and Fischer 2002) (designed for classical MAB problems) and the ϵ-first algorithm with $\epsilon = 0.1$ (Tran-Thanh et al. 2010) (designed for the MAB-BF problems) as baselines.

The regrets of the four algorithms are shown in Figure 1(a). We can see that UCB1 performs the worst, and our proposed algorithms outperform both baselines. The results are easy to understand because the baselines are not delicately designed for the MAB-BV setting and cannot well leverage the structure of the problem to achieve good performances. To see the differences between our two algorithms in a clearer manner, we re-draw their regrets in Figure 1(b). From it, we can see UCB-BV1 is actually better UCB-BV2. This is consistent with our theoretical analysis: UCB-BV1 has a tighter regret bound because it uses additional prior knowledge about the bandit.

In the previous section, we mentioned that our proposed UCB-BV algorithms can be applied to the MAB-BF problems and can also have nice regret bounds in that setting. To verify this, we conduct a second simulation. Specifically, we set the variance of the cost in the first simulation to be zero. The regrets of our UCB-BV algorithms and the two baselines are shown in Figure 1(c) and 1(d). Again, according to the experimental results, our proposed algorithms perform much better than the baselines, and UCV-BV1 performs better than UCB-BV2.

To sum up, the experimental results are consistent with our theoretical analysis and verify the effectiveness of our proposed algorithms.

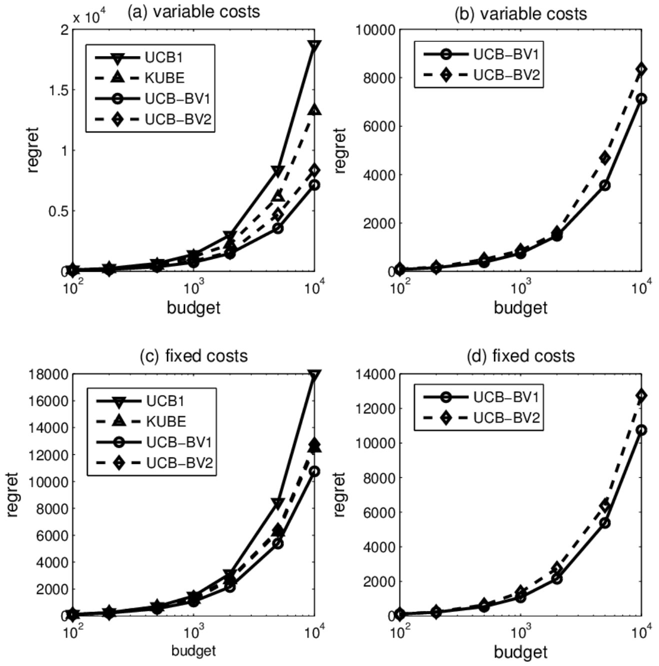

Figure 1: Experimental results

Conclusions and Future Work

In this paper, we have studied the multi-armed bandit problems with budget constraint and variable costs (MAB-BV). We have proposed two learning algorithms for the problem and proven their regret bounds of $O(\ln(B))$. We have also applied the proposed algorithms to the multi-armed bandit problems with budget constraint and fixed costs (MAB-BF), and shown that they can achieve better regret bounds than existing methods for that setting. Numerical simulations demonstrate the effectiveness of our proposed algorithms.

As for future work, we plan to study the following aspects. First, we will investigate generalized versions of the MAB-BV problems, including (1) the setting with many arms and even continuum arms, and (2) the setting with unbounded rewards (and costs). Second, we will investigate the case of correlated arms where the rewards and costs of different arms are interdependent. Third, we plan to study multi-armed bandit problems with multiple budget constraints and multiple costs (e.g., taking an action involves both a time cost and a money cost).

References

Agrawal, R.; Hedge, M.; and Teneketzis, D. 1988. Asymptotically efficient adaptive allocation rules for the multi-armed bandit problem with switching cost. *Automatic Control, IEEE Transactions on* 33(10):899–906.

Ardagna, D.; Panicucci, B.; and Passacantando, M. 2011.

A game theoretic formulation of the service provisioning problem in cloud systems. In *Proceedings of the 20th international conference on World wide web*, 177–186. ACM.

Audibert, J.; Bubeck, S.; et al. 2010. Best arm identification in multi-armed bandits. In *Proceedings of the 23rd Annual Conference on Learning Theory*.

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

Auer, P. 2003. Using confidence bounds for exploitation-exploration trade-offs. *The Journal of Machine Learning Research* 3:397–422.

Awerbuch, B., and Kleinberg, R. 2004. Near-optimal adaptive routing: Shortest paths and geometric generalizations. In *Proceeding of the 36th Annual ACM Symposium on Theory of Computing*, 45–53.

Babaioff, M.; Sharma, Y.; and Slivkins, A. 2009. Characterizing truthful multi-armed bandit mechanisms. In *Proceedings of the 10th ACM conference on Electronic commerce*, 79–88. ACM.

Ben-Yehuda, O.; Ben-Yehuda, M.; Schuster, A.; and Tsafrir, D. 2011. Deconstructing amazon ec2 spot instance pricing. In *Cloud Computing Technology and Science (CloudCom), 2011 IEEE Third International Conference on*, 304–311. IEEE.

Berry, D.; Chen, R.; Zame, A.; Heath, D.; and Shepp, L. 1997. Bandit problems with infinitely many arms. *The Annals of Statistics* 25(5):2103–2116.

Borgs, C.; Chayes, J.; Immorlica, N.; Jain, K.; Etesami, O.; and Mahdian, M. 2007. Dynamics of bid optimization in online advertisement auctions. In *Proceedings of the 16th international conference on World Wide Web*, 531–540. ACM.

Bubeck, S.; Munos, R.; and Stoltz, G. 2009. Pure exploration in multi-armed bandits problems. In *Algorithmic Learning Theory*, 23–37. Springer.

Chakrabarti, D.; Kumar, R.; Radlinski, F.; and Upfal, E. 2008. Mortal multi-armed bandits. *Advances in Neural Information Processing Systems* 30:255–262.

Chakraborty, T.; Even-Dar, E.; Guha, S.; Mansour, Y.; and Muthukrishnan, S. 2010. Selective call out and real time bidding. *Internet and Network Economics* 145–157.

Chu, W.; Li, L.; Reyzin, L.; and Schapire, R. 2011. Contextual bandits with linear payoff functions. In *Conference on Artificial Intelligence and Statistics (AISTATS)*.

Feldman, J.; Mirrokni, V.; Muthukrishnan, S.; and Pai, M. 2010. Auctions with intermediaries. In *Proceedings of the 11th ACM conference on Electronic commerce*, 23–32. ACM.

Guha, S., and Munagala, K. 2007. Approximation algorithms for budgeted learning problems. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, 104–113. ACM.

Kleinberg, R. 2004. Nearly tight bounds for the continuum-armed bandit problem. *Advances in Neural Information Processing Systems* 17:697–704.

Tran-Thanh, L.; Chapman, A.; Munoz De Cote Flores Luna, J.; Rogers, A.; and Jennings, N. 2010. Epsilon–first policies for budget–limited multi-armed bandits. In *Twenty-Fourth AAAI Conference on Artificial Intelligence*, 1211–1216.

Tran-Thanh, L.; Chapman, A.; Rogers, A.; and Jennings, N. 2012. Knapsack based optimal policies for budget–limited multi–armed bandits. 1134–1140.

Wei, G.; Vasilakos, A.; Zheng, Y.; and Xiong, N. 2010. A game-theoretic method of fair resource allocation for cloud computing services. *The Journal of Supercomputing* 54(2):252–269.

Zaman, S., and Grosu, D. 2010. Combinatorial auction-based allocation of virtual machine instances in clouds. In *Cloud Computing Technology and Science (CloudCom), 2010 IEEE Second International Conference on*, 127–134. IEEE.