Best Action Selection in a Stochastic Environment*

Yingce Xia[1], Tao Qin[2], Nenghai Yu[1], Tie-Yan Liu[2]
[1]University of Science and Technology of China [2]Microsoft Research
yingce.xia@gmail.com, {taoqin, tie-yan.liu}@microsoft.com, ynh@ustc.edu.cn

ABSTRACT

We study the problem of selecting the best action from multiple candidates in a stochastic environment. In such a stochastic setting, when taking an action, a player receives a random reward and affords a random cost, which are drawn from two unknown distributions. We target at selecting the best action, the one with the maximum ratio of the expected reward to the expected cost, after exploring the actions for n rounds. In particular, we study three mechanisms: (i) the uniform exploration mechanism $\mathcal{M}_{\mathrm{U}}$; (ii) the successive elimination mechanism $\mathcal{M}_{\mathrm{SE}}$; and (iii) the ratio confidence bound exploration mechanism $\mathcal{M}_{\mathrm{RCB}}$. We prove that for all the three mechanisms, the probabilities that the best action is not selected (i.e., the error probabilities) can be upper bounded by $O(\exp\{-cn\})$, where c is a constant related to the mechanisms and coefficients about the actions. We then give an asymptotic lower bound of the error probabilities of the consistent mechanisms for Bernoulli setting, and discuss its relationship with the upper bounds in different aspects. Our proposed mechanisms can be degenerated to cover the cases where only the reward/costs are random. We also test the proposed mechanisms through numerical experiments.

Keywords

Design, Economics, Bandit Algorithm, Stochastic

1. INTRODUCTION

Sponsored search is a very effective means for and widely used by many businesses to advertise and promote their products. In sponsored search, an advertiser needs to bid a keyword for her ad to participate in ad auctions. After bidding a keyword, whether she can win the ad auction depends on how many other advertisers bid exactly the same keyword and other related keywords. If she wins the auction, whether her ad is clicked is determined by the behaviors of the search users, which depends on many random factors. If her ad is clicked, the payment to the search engine is determined by the click-through rate (CTR) and the bid of the ad next to hers according to the generalized second pricing (GSP) mechanism.[1] Thus, the reward (click or not) and cost (the payment) of choosing a keyword are both random in sponsored search. To maximize her utility, the advertiser needs to identify the keyword (as well as setting a proper bid) with the maximal ratio of the expected clicked number to expected cost as soon as possible. Such a problem can be abstracted as the best action selection in a stochastic environment, in which an action is the operation of bidding a keyword (including selecting the keyword and setting a bid).

Cloud computing provides an effective way for firms to reduce their computation and IT costs. There are more and more firms moving their computation tasks to cloud. Amazon EC2, the largest provider in cloud computing, provides a specific kind of virtual machines, spot instances[2], for price sensitive firms. Spot instances let one bid on spare Amazon EC2 instances to name one's own price for computation capacity. The spot price fluctuates based on the supply and demand of available EC2 capacity and thus is random. Different types of spot instances have different computation powers (due to the different resource configurations, such as CPU, memory, IO, bandwidth, and geographical locations) and different prices. Since the real performance of a spot instance is impacted by the resource consumption of other instances that are virtualized in the same physical server, there is no guarantee that two spot instances of the same type have the same performance, e.g., how many web visitors can be supported concurrently if hosting a website on the spot instance. Actually, it is more natural to model the performances of spot instances as random variables. Therefore, to be better off, a firm would like to select the type of instance with the maximum ratio of the expected performance to the expected cost as soon as possible after trying different types of instances for several rounds. This instance selection problem can also be modeled as the best action selection in a stochastic environment.

The above two problems are related to the *best arm identification* for multi-armed bandits [6, 1, 15, 20]. A multi-armed bandit (MAB) is a slot machine with K arms. After pulling an arm, the player will receive a reward drawn from an unknown distribution associated with the arm. After pulling the arms for n rounds, the player needs to recommend the arm with the largest expected reward she thinks. The algorithms proposed for best arm identification cannot directly fit our concerned problems, because in our problem, the player will receive two observations, a random reward

*This work was conducted at Microsoft Research Asia.

Appears in: *Proceedings of the 15th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2016), J. Thangarajah, K. Tuyls, C. Jonker, S. Marsella (eds.), May 9–13, 2016, Singapore.*

[1]For more details, please refer to a recent survey about sponsored search auctions [18].
[2]`https://aws.amazon.com/ec2/spot/`

and a random cost, after each selection, and is interested in the ratio of the observations. In this work, we design new mechanisms to solve the best action selection problem in a stochastic environment.

Model Setup There are $K(\geq 2)$ candidate actions to be selected. For any $i \in [K]$ (denote the set $\{1, 2, \cdots, K\}$ as $[K]$ for simplicity), action i is associated with a reward distribution and a cost distribution, both of which are unknown and with bounded supports. Without loss of generality, we assume the two distributions are supported in $[0, 1]$. At round t, taking action i results in a random reward $X^{\lambda}_{i,t}$ and a random cost $X^{\mu}_{i,t}$, which are independently sampled[3] from the two distributions and are independent of the past actions and observations. Denote the expected reward and cost of action i as λ_i and μ_i respectively, i.e., $\mathbb{E}\{X^{\lambda}_{i,t}\} = \lambda_i$ and $\mathbb{E}\{X^{\mu}_{i,t}\} = \mu_i$. We assume that $0 < \lambda_i, \mu_i < 1$ for any $i \in [K]$. The best action is defined as the one with the maximum ratio of the expected reward to the expected cost. For simplicity, we assume that there is a single best action. W.l.o.g, we set that $\frac{\lambda_1}{\mu_1} > \frac{\lambda_2}{\mu_2} \geq \frac{\lambda_3}{\mu_3} \geq \cdots \geq \frac{\lambda_K}{\mu_K}$ throughout this work. Therefore, action 1 is the best one, and the other action are suboptimal ones. Please note that this order is just for ease of theoretical analysis, and the player (e.g., the advertiser to explore keywords in sponsored search and the firm to try different instances in cloud computing) has no such order information.

The player can explore the actions for n rounds, after which she needs to find out the best action she thinks. Here n is a positive integer known and fixed in advance. Denote the action that the player selects after n rounds as J_n. The player would like to maximize the probability of finding the best action, i.e., minimize the error probability defined in (1), where $\mathcal{M}$ is the exploration mechanism the player uses.

$$\mathscr{E}_n(\mathcal{M}) = \mathbb{P}\{J_n \text{ is not the best action}\}. \tag{1}$$

When the context is clear, we will omit $\mathcal{M}$ from $\mathscr{E}_n(\mathcal{M})$.

Proposed Mechanisms As far as we know, there is no literature about how to select the action with the maximum ratio of the expected reward to the expected cost in a stochastic environment. In this work, we design three mechanisms for the problem. The first is the naive uniform exploration mechanism $\mathcal{M}_{\mathrm{U}}$, in which we take turns to try each action, observe the rewards and costs. The second is the successive elimination mechanism $\mathcal{M}_{\mathrm{SE}}$, in which we divide the n selection rounds into $K-1$ phases (recall K is the number of candidate actions); and at the end of each phase, we eliminate the empirical worst action. The third is the ratio confidence bound mechanism $\mathcal{M}_{\mathrm{RCB}}$, which is an adaptive version of the UCB mechanism [1]. For $\mathcal{M}_{\mathrm{U}}$ and $\mathcal{M}_{\mathrm{RCB}}$, we recommend the action with the maximum ratio of the accumulative rewards to the accumulative costs. For $\mathcal{M}_{\mathrm{SE}}$, there is only one action that survives after n rounds of selections and therefore we recommend this action.

Theoretical Results We prove that for all the above three mechanisms, $\mathscr{E}_n$ can be upper bounded by $O(\exp\{-cn\})$, where c is a constant related to the mechanisms and coefficients about the actions. Thus, all the three mechanisms are consistent in the sense that their error probabilities converge to zero exponentially as n tends to infinity. We also give an asymptotic lower bound of the error probability of the consistent mechanisms for the setting that the reward and cost distributions of all the actions are Bernoulli, which is $\Omega(\exp\{-\mathcal{D}_* n\})$ and $\mathcal{D}_*$ is an action-related coefficient. We show that for the aforementioned Bernoulli setting, the upper bounds of the proposed three mechanisms match the lower bound, in terms of the order of n. That is, the three mechanisms are optimal in certain conditions. In addition, we also make discussions on the impact of the action-related parameters on the constants in the above bounds.

[3]Here we assume that the reward of an action is independent of its cost. Note that an action with a higher cost does not always have a higher reward. For example, in the keyword bidding problem, a keyword $k1$ can have a higher cost than another keyword $k2$ because more advertisers bid for $k1$, but $k1$ does not necessarily lead to more clicks. We leave the setting with dependent rewards and costs to future work.

2. RELATED WORK

Bandit algorithms play important roles for mechanism design in stochastic environments [14, 4, 3, 21, 9, 19]. The most important related work about our concerned problem is the literature about best arm identification. There are two settings for the best arm identification problem in MAB:

Fixed budget setting: The player needs to minimize the error probability within n rounds of pulling and n is fixed and known in advance. A UCB style exploration mechanism was proposed in [1]. The idea of elimination policy was proposed in [1, 2, 24]. Different criteria to recommend the best arm were discussed in [7]: the maximum empirical average reward, the maximum pulling time, and the probabilities proportional to the pulling time of each arm.

Fixed confidence setting: The player needs to minimize the pulling rounds while guaranteeing that the error probability is smaller than a given threshold. For this setting, a lower bound of the sample complexity was given in [17], and an action elimination mechanism was proposed in [11]. Furthermore, the sample complexities under both settings were compared in [16], and a unified approach for both settings was presented in [12].

Another line of related work is the budgeted MAB [23, 10, 26, 19, 25], in which the goal of the player is to maximize the accumulate reward before the budget runs out. The mechanism design in our paper can be seen as a dual problem of the above five works. Compared with the budgeted MAB, we focus on searching for the best action rather than minimizing the cumulative regret.

3. MECHANISMS

Before describing the mechanisms, we first introduce some notations that are frequently used in the following sections. For any $t \in [n]$ and $i \in [K]$, (1) I_t denotes the action selected at round t; (2) $T_i(t)$, $\bar{X}^{\lambda}_{i,T_i(t)}$ and $\bar{X}^{\mu}_{i,T_i(t)}$ denote the number of selected rounds, the average reward and average cost of action i until round t respectively. Mathematically,

$$\begin{aligned} &T_i(t) = \sum_{s=1}^{t} \mathbf{1}\{I_s = i\};\ \bar{X}^{\lambda}_{i,T_i(t)} = \frac{1}{T_i(t)} \sum_{s=1}^{t} X^{\lambda}_{i,s} \mathbf{1}\{I_s = i\}; \\ &\bar{X}^{\mu}_{i,T_i(t)} = \frac{1}{T_i(t)} \sum_{s=1}^{t} X^{\mu}_{i,s} \mathbf{1}\{I_s = i\}. \end{aligned} \tag{2}$$

We propose three mechanisms to the applications modeled in Section 1. Our mechanisms are inspired from the best arm identification problem in conventional bandits.

First, we adapt the *uniform exploration mechanism* [6, 22] to the best action selection, and call the corresponding mechanism $\mathcal{M}_{\mathrm{U}}$, which is shown in Algorithm 1. The idea

behind $\mathcal{M}_{\mathrm{U}}$ is very simple: one just selects actions one after one and then recommends the action with the largest ratio of the empirical average reward to the average cost.

Algorithm 1: Uniform Exploration Mechanism ($\mathcal{M}_{\mathrm{U}}$)

for $t \leftarrow 1$ *to* n **do**
  Select action (t mod K)+1;
Return $J_n = \arg\max_i\{\bar{X}^{\lambda}_{i,T_i(n)}/\bar{X}^{\mu}_{i,T_i(n)}\}$.

The uniform exploration mechanism is simple. However, it has a clear drawback: after exploring the actions for a number of rounds, even if we know with high probabilities that some actions cannot be the best one, we still need to explore them as frequently as others.

Algorithm 2: Successive Elimination Mechanism($\mathcal{M}_{\mathrm{SE}}$)

Initialization: $A_1 = [K]$; $n_0 = 0$;$\{n_k\}_{k\in[K-1]}$ in (3);
for *phase* $k \leftarrow 1$ *to* $K-1$ **do**
  For any $i \in A_k$, select action i for $n_k - n_{k-1}$ rounds;
  $A_{k+1} \leftarrow A_k \backslash \{\arg\min_{i\in A_k}(\bar{X}^{\lambda}_{i,n_k}/\bar{X}^{\mu}_{i,n_k})\}$;
Return J_n = the unique action in A_K.

The *elimination policy* [11, 24] is a strategy that can overcome this drawback. It sequentially eliminates suboptimal actions. We adapt it to our mechanisms and propose the *Successive Elimination mechanism* $\mathcal{M}_{\mathrm{SE}}$ (See Algorithm 2). The n_k for any $k \in [K-1]$ that is used in $\mathcal{M}_{\mathrm{SE}}$ is defined as below:

$$H(K) = \frac{1}{2} + \sum_{i=2}^{K} \frac{1}{i}; \quad n_k = \left\lceil \frac{1}{H(K)} \frac{n-K}{K+1-k} \right\rceil. \tag{3}$$

Mechanism $\mathcal{M}_{\mathrm{SE}}$ works as follows. It divides the rounds into $K-1$ phases. At the end of each phase, it removes the action with the smallest ratio of the average rewards to the average costs. Intuitively, actions with closer ratios to the best one are more likely to survive after more phases.

Algorithm 3: RCB Exploration Mechanism ($\mathcal{M}_{\mathrm{RCB}}$)

Select each action once at the first K rounds;
for $t \leftarrow K+1$ *to* n **do**
  $\forall i \in [K]$, calculate the $\hat{\lambda}_i(t)$ and $\hat{\mu}_i(t)$ defined in (4);
  if $\{i|\hat{\mu}_i(t) \le 0\}$ *is not empty* **then**
    Select action $I_t = \arg\min_i\{T_i(t-1)|\hat{\mu}_i(t) \le 0\}$;
  else
    Select action $I_t = \arg\max_i\{\hat{\lambda}_i(t)/\hat{\mu}_i(t)\}$;
Return $J_n = \arg\max_i\{\bar{X}^{\lambda}_{i,T_i(n)}/\bar{X}^{\mu}_{i,T_i(n)}\}$.

The UCB-style mechanisms (see Chapter 2 of [5] for an introduction) play an important role of best arm identification. We also adapt a UCB-style mechanism for our best action selection problem, which we call the *Ratio Confidence Bound mechanism* $\mathcal{M}_{\mathrm{RCB}}$ (see Algorithm 3). In $\mathcal{M}_{\mathrm{RCB}}$, we introduce an upper confidence bound for the reward and a lower confidence bound for the cost as defined in (4):

$$\hat{\lambda}_i(t) = \bar{X}^{\lambda}_{i,T_i(t-1)} + \mathcal{E}^{\alpha}_i(t),\ \hat{\mu}_i(t) = \bar{X}^{\mu}_{i,T_i(t-1)} - \mathcal{E}^{\alpha}_i(t), \tag{4}$$

where $i \in [K], t \in [n], \mathcal{E}^{\alpha}_i(t) = \sqrt{\alpha/T_i(t-1)}$ and α is a positive hyper-parameter. Note that both $\hat{\lambda}_i(t)$ and $\hat{\mu}_i(t)$ depend on α, and we omit the α for simplicity.

The "ratio" in $\mathcal{M}_{\mathrm{RCB}}$ is between the upper confidence bound of the average reward and the lower confidence bound of the average cost.[4] To run $\mathcal{M}_{\mathrm{RCB}}$, one needs a positive hyper parameter α as the input.

From the above descriptions, one can see that the ratios of the empirical average rewards to costs are important to all the three mechanisms. Note that it is practically possible that the denominator of $\bar{X}^{\lambda}_{i,T_i(t)}/\bar{X}^{\mu}_{i,T_i(t)}$ is zero for some i and t. In this case, we calculate the ratio as follows:

$$\begin{aligned} &\bar{X}^{\lambda}_{i,T_i(t)}/\bar{X}^{\mu}_{i,T_i(t)} = 0 \quad \text{if } \bar{X}^{\lambda}_{i,T_i(t)} = 0; \\ &\bar{X}^{\lambda}_{i,T_i(t)}/\bar{X}^{\mu}_{i,T_i(t)} = \infty \quad \text{if } \bar{X}^{\lambda}_{i,T_i(t)} > 0. \end{aligned} \tag{5}$$

Furthermore, we break ties randomly when more than one action is selected in the arg min or arg max operators in the three mechanisms.

4. MECHANISM ANALYSIS

In this section, we prove the upper bounds of the error probabilities of the proposed three mechanisms.

Before the formal theoretical analysis, we introduce two new concepts defined as follows: For any $i \ge 2$, we define Δ_i and ϱ_i as follows:

$$\Delta_i = \frac{\lambda_1}{\mu_1} - \frac{\lambda_i}{\mu_i}; \qquad \varrho_i = \frac{\mu_1\mu_i\Delta_i}{\lambda_1+\mu_1+\lambda_i+\mu_i}. \tag{6}$$

Δ_i measures the gap between the best action and a suboptimal action i, in terms of the ratio of the expected reward to the expected cost. ϱ_i measures the difference between the best action and a suboptimal action i in another way: in order to make their ratios equal, we can increase λ_i and μ_1 by ϱ_i, while decreasing λ_1 and μ_i by ϱ_i, i.e.,

$$\frac{\lambda_1 - \varrho_i}{\mu_1 + \varrho_i} = \frac{\lambda_i + \varrho_i}{\mu_i - \varrho_i}. \tag{7}$$

It is easy to verify $\lambda_1 > \varrho_i$ and $\mu_i > \varrho_i$ for any $i \ge 2$.
We further define the following notations: (1) $\varrho_1 = \min_{i\ge2}\{\varrho_i\}$; (2) $H_1 = \max\{\mu_1^{-2}, \varrho_1^{-2}\} + \sum_{i=2}^{K} \varrho_i^{-2}$; (3) $H_2 = \max_{i\ge2} i\varrho_i^{-2}$.

The upper bounds of error probabilities of $\mathcal{M}_{\mathrm{U}}, \mathcal{M}_{\mathrm{SE}}$ and $\mathcal{M}_{\mathrm{RCB}}$ depends on the aforementioned notations. We first present the result for $\mathcal{M}_{\mathrm{U}}$:

THEOREM 1. *The error probability of $\mathcal{M}_U$ can be upper bounded as below:*

$$\mathscr{E}_n(\mathcal{M}_U) \le 2\sum_{i=1}^{K} e^{2\varrho_i^2} \exp\{-2\varrho_i^2 \frac{n}{K}\}. \tag{8}$$

PROOF. If $\mathcal{M}_{\mathrm{U}}$ recommends a suboptimal action, i.e., $J_n \ne 1$, we know that the following event ξ_{unif} is true:

$$\xi_{\mathrm{unif}} = \bigcup_{i=2}^{K} \left\{ \frac{\bar{X}^{\lambda}_{i,T_i(n)}}{\bar{X}^{\mu}_{i,T_i(n)}} \ge \frac{\bar{X}^{\lambda}_{1,T_1(n)}}{\bar{X}^{\mu}_{1,T_1(n)}} \right\}. \tag{9}$$

According to (7), we know at least one of the following two events holds:

$$(1)\ \frac{\bar{X}^{\lambda}_{1,T_1(n)}}{\bar{X}^{\mu}_{1,T_1(n)}} \le \frac{\lambda_1 - \varrho_i}{\mu_1 + \varrho_i}; \quad (2)\ \frac{\bar{X}^{\lambda}_{1,T_1(n)}}{\bar{X}^{\mu}_{1,T_1(n)}} \ge \frac{\lambda_i + \varrho_i}{\mu_i - \varrho_i}. \tag{10}$$

[4] In the 5th step of Alogorithm 3, if $\hat{\mu}_i(t)$ is smaller than zero, it is highly probable that 0 is within the lower confidence bound for the cost. Therefore, we have to continue exploring these actions to get more accurate estimations of the expected costs.

By intersecting the ξ_{unif} defined in (9) and the union of the two events in (10), we have

$$\xi_{\text{unif}} \subseteq \bigcup_{i=2}^{K} \Big\{ \Big\{ \frac{\bar{X}^{\lambda}_{1,T_1(n)}}{\bar{X}^{\mu}_{1,T_1(n)}} \le \frac{\lambda_1 - \varrho_i}{\mu_1 + \varrho_i} \Big\} \bigcup \Big\{ \frac{\bar{X}^{\lambda}_{i,T_i(n)}}{\bar{X}^{\mu}_{i,T_i(n)}} \ge \frac{\lambda_i + \varrho_i}{\mu_i - \varrho_i} \Big\} \Big\},$$

which can be further decomposed as

$$\begin{aligned}
\xi_{\text{unif}} \subseteq & \bigcup_{i=2}^{K} \Big\{ \{\bar{X}^{\lambda}_{1,T_1(n)} \le \lambda_1 - \varrho_i\} \cup \{\bar{X}^{\mu}_{1,T_1(n)} \ge \mu_1 + \varrho_i\} \\
& \cup \{\bar{X}^{\lambda}_{i,T_i(n)} \ge \lambda_i + \varrho_i\} \cup \{\bar{X}^{\mu}_{i,T_i(n)} \le \mu_i - \varrho_i\} \Big\} \\
= & \bigcup_{i=2}^{K} \Big\{ \{\bar{X}^{\lambda}_{i,T_i(n)} \ge \lambda_i + \varrho_i\} \cup \{\bar{X}^{\mu}_{i,T_i(n)} \le \mu_i - \varrho_i\} \Big\} \quad (11) \\
& \cup \{\bar{X}^{\lambda}_{1,T_1(n)} \le \lambda_1 - \varrho_1\} \cup \{\bar{X}^{\mu}_{1,T_1(n)} \ge \mu_1 + \varrho_1\}. \quad (12)
\end{aligned}$$

Please note that: (1) even if $\bar{X}^{\mu}_{i,T_i(n)} = 0$ for some i, this case also belongs to (11) or (12); (2) when using $\mathcal{M}_{\text{U}}$, $T_i(n)$ for any i is deterministic, rather than random. Therefore, according to Hoeffding's inequality[5], we can obtain

$$\begin{aligned}
\mathbb{P}\{\xi_{\text{unif}}\} \le & \sum_{i=2}^{K} \Big\{ \mathbb{P}\{\bar{X}^{\lambda}_{i,T_i(n)} \ge \lambda_i + \varrho_i\} + \mathbb{P}\{\bar{X}^{\mu}_{i,T_i(n)} \le \mu_i - \varrho_i\} \Big\} \\
& + \mathbb{P}\{\bar{X}^{\lambda}_{1,T_1(n)} \le \lambda_1 - \varrho_1\} + \mathbb{P}\{\bar{X}^{\mu}_{1,T_1(n)} \ge \mu_1 + \varrho_1\} \\
\le & \, 2 \sum_{i=1}^{K} \exp\{-2T_i(n)\varrho_i^2\}.
\end{aligned}$$

It is obvious that $T_i(n) \ge \frac{n}{K} - 1$ for any $i \in [K]$ when using $\mathcal{M}_{\text{U}}$, according to which we know

$$\mathscr{E}_n(\mathcal{M}_{\text{U}}) \le \mathbb{P}\{\xi_{\text{unif}}\} \le 2 \sum_{i=1}^{K} e^{2\varrho_i^2} \exp\{-2\varrho_i^2 \frac{n}{K}\}. \quad (13)$$

Therefore, we can conclude Theorem 1. □

The proof of Theorem 1 illustrates the importance of the concept ϱ_i defined in (6). The ϱ_i concept is also crucial to the proofs of other mechanisms. $\mathcal{M}_{\text{U}}$ is a simple and naive mechanism. Next we show the theoretical results for $\mathcal{M}_{\text{SE}}$:

THEOREM 2. *The error probability of $\mathcal{M}_{SE}$ can be upper bounded as below:*

$$\mathscr{E}_n(\mathcal{M}_{SE}) \le 2K(K-1) \exp\Big\{ - \frac{2(n-K)}{H_2 \ln 2K} \Big\}. \quad (14)$$

PROOF. First of all, we can verify that Algorithm 2 takes at most n steps. This is because the action eliminated at phase $k(\le K-1)$ is selected for n_k rounds. Thus, the total action selection rounds of $\mathcal{M}_{\text{SE}}$ is:

$$\Big(\sum_{k=1}^{K-1} n_k \Big) + n_{K-1} \le K + \frac{n-K}{\frac{1}{2} + \sum_{i=2}^{K} \frac{1}{i}} \Big(\frac{1}{2} + \sum_{i=2}^{K} \frac{1}{i} \Big) \le n.$$

During phase $k(\le K-1)$, at least one of the k worst actions survives. Therefore, if action 1 (i.e., the best action) has been eliminated at phase k, it means that

$$\frac{\bar{X}^{\lambda}_{1,n_k}}{\bar{X}^{\mu}_{1,n_k}} \le \max_{i \in \{K, K-1, K+1-k\}} \frac{\bar{X}^{\lambda}_{i,n_k}}{\bar{X}^{\mu}_{i,n_k}}. \quad (15)$$

[5]A quick introduction can be found at `https://en.wikipedia.org/wiki/Hoeffding's_inequality`

According to (15), we have that

$$\begin{aligned}
\mathscr{E}_n(\mathcal{M}_{\text{SE}}) \le & \sum_{k=1}^{K-1} \sum_{i=K+1-k}^{K} \mathbb{P}\Big\{ \frac{\bar{X}^{\lambda}_{1,n_k}}{\bar{X}^{\mu}_{1,n_k}} \le \frac{\bar{X}^{\lambda}_{i,n_k}}{\bar{X}^{\mu}_{i,n_k}} \Big\} \\
\le & \sum_{k=1}^{K-1} \sum_{i=K+1-k}^{K} \Big(\mathbb{P}\Big\{ \frac{\bar{X}^{\lambda}_{1,n_k}}{\bar{X}^{\mu}_{1,n_k}} \le \frac{\lambda_1 - \varrho_i}{\mu_1 + \varrho_i} \Big\} + \mathbb{P}\Big\{ \frac{\bar{X}^{\lambda}_{i,n_k}}{\bar{X}^{\mu}_{i,n_k}} \ge \frac{\lambda_i + \varrho_i}{\mu_i - \varrho_i} \Big\} \Big) \\
\le & \sum_{k=1}^{K-1} \sum_{i=K+1-k}^{K} \Big(\mathbb{P}\Big\{ \bar{X}^{\lambda}_{1,n_k} \le \lambda_1 - \varrho_i \Big\} + \mathbb{P}\Big\{ \bar{X}^{\mu}_{1,n_k} \ge \mu_1 + \varrho_i \Big\} \\
& + \mathbb{P}\Big\{ \bar{X}^{\lambda}_{i,n_k} \ge \lambda_i + \varrho_i \Big\} + \mathbb{P}\Big\{ \bar{X}^{\mu}_{i,n_k} \le \mu_i - \varrho_i \Big\} \Big) \\
\le & \sum_{k=1}^{K-1} \sum_{i=K+1-k}^{K} 4 \exp\{-2n_k \varrho_i^2\} \\
\le & \sum_{k=1}^{K-1} \sum_{i=K+1-k}^{K} 4 \exp\{-2(\frac{1}{H(K)} \frac{n-K}{K+1-k}) \varrho_i^2\} \\
\le & \sum_{k=1}^{K-1} \sum_{i=K+1-k}^{K} 4 \exp\{-2(\frac{1}{H(K)} \frac{n-K}{i \varrho_i^{-2}})\} \\
\le & 2K(K-1) \exp\Big\{ - \frac{2(n-K)}{H(K) H_2} \Big\} \le 2K(K-1) \exp\Big\{ - \frac{2(n-K)}{H_2 \ln 2K} \Big\}.
\end{aligned}$$

Therefore, we have proved Theorem 2. □

From Theorem 1 and Theorem 2, we can see that the error probabilities of both $\mathcal{M}_{\text{U}}$ and $\mathcal{M}_{\text{SE}}$ decrease exponentially in terms of n. We can verify that when $\varrho_1^2 \le K/[H_2 \ln(2K)]$, (intuitively, ϱ_1 is very small and K is large) for a sufficiently large n, the upper bound of $\mathscr{E}_n(\mathcal{M}_{\text{SE}})$ will be smaller than that of $\mathscr{E}_n(\mathcal{M}_{\text{U}})$.

Finally, we have the following theorem for $\mathcal{M}_{\text{RCB}}$. (Keep in mind that for any i and t, $\hat{\lambda}_i(t)$ and $\hat{\mu}_i(t)$ in $\mathcal{M}_{\text{RCB}}$ are related to the hyper-parameter α.)

THEOREM 3. *If $\mathcal{M}_{RCB}$ runs with hyper parameter $0 < \alpha \le \frac{9}{16} \frac{n-K}{H_1}$, we have that*

$$\mathscr{E}_n(\mathcal{M}_{RCB}) \le 4nK \exp\big\{ - \frac{\alpha}{50} (\lambda_1 + \mu_1)^2 \big\}. \quad (16)$$

In particular, for $\alpha = \frac{9}{16} \frac{n-K}{H_1}$, we have

$$\mathscr{E}_n(\mathcal{M}_{RCB}) \le 4nK \exp\Big\{ - \frac{9(\lambda_1 + \mu_1)^2}{800 H_1} (n-K) \Big\}. \quad (17)$$

We will prove Theorem 3 by three steps from ($\mathcal{S}1$) to ($\mathcal{S}3$):

($\mathcal{S}1$) *Preparation*: Denote the event $\bar{\xi}_{\text{RCB}}$ as below:

$$\Big\{ \forall i \in [K], s \in [n], |\bar{X}^{\lambda}_{i,s} - \lambda_i| < x_0 \sqrt{\frac{\alpha}{s}}, |\bar{X}^{\mu}_{i,s} - \mu_i| < x_0 \sqrt{\frac{\alpha}{s}} \Big\},$$

in which $x_0 = (\lambda_1 + \mu_1)/[3(\lambda_1 + \mu_1) + 2\max_{i \ge 2}(\lambda_i + \mu_i)]$. According to Hoeffding's inequality, we can get

$$\mathbb{P}\{\bar{\xi}_{\text{RCB}}\} \le 4nK \exp\{-2x_0^2 \alpha\}. \quad (18)$$

It is easy to verify that

$$(\lambda_1 + \mu_1)/10 \le x_0 < 1/3. \quad (19)$$

Combining (18) and (19), $\mathbb{P}\{\bar{\xi}_{\text{RCB}}\}$ can be upper bounded as below.

$$\mathbb{P}\{\bar{\xi}_{\text{RCB}}\} \le 4nK \exp\big\{ - \frac{\alpha}{50} (\lambda_1 + \mu_1)^2 \big\}. \quad (20)$$

We present two lemmas that will be used later. Please note that the two lemmas can also be seen as side products of the number of selection rounds of the actions:

LEMMA 4. *Conditioned on ξ_{RCB}, for any $i \geq 2$ and $t \leq n$, we have that*

$$T_i(t) \leq \frac{(1+x_0)^2\alpha}{\varrho_i^2} + 1. \tag{21}$$

LEMMA 5. *Conditioned on ξ_{RCB}, for any $i \geq 2$ and $t \leq n$, at least one of the following two events is true:*

(a) $T_i(t) \geq \frac{\alpha}{4}(1-x_0)^2(\lambda_1+\mu_1)^2 F_i$, where

$$F_i = \min\Big\{\frac{1}{(\mu_i\mu_1\Delta_i)^2}, \frac{T_1(t)-1}{\alpha(1+x_0)^2(\lambda_i+\mu_i)^2}\Big\};$$

(b) $T_1(t) \leq (1+x_0)^2\alpha\mu_1^{-2} + 1$.

We prove Lemma 4 by induction. Lemma 5 can be similarly proved and we omit its proof due to space limitations.

Proof of Lemma 4: Lemma 4 is proved by induction. It is obvious that (21) holds for $t \leq K$. Assume (21) holds at round $t(\geq K)$. At round $t+1$, if $I_{t+1} \neq i$, we have

$$T_i(t+1) = T_i(t) \leq \frac{(1+x_0)^2\alpha}{\varrho_i^2} + 1.$$

If $I_{t+1} = i$, at least one of the following two cases is true:

$$\text{Case (i)} \quad \bar{X}^{\mu}_{i,T_i(t)} - \sqrt{\tfrac{\alpha}{T_i(t)}} \leq 0;$$

$$\text{Case (ii)} \quad \frac{\bar{X}^{\lambda}_{i,T_i(t)} + \sqrt{\frac{\alpha}{T_i(t)}}}{\bar{X}^{\mu}_{i,T_i(t)} - \sqrt{\frac{\alpha}{T_i(t)}}} \geq \frac{\bar{X}^{\lambda}_{1,T_1(t)} + \sqrt{\frac{\alpha}{T_1(t)}}}{\bar{X}^{\mu}_{1,T_1(t)} - \sqrt{\frac{\alpha}{T_1(t)}}}, \tag{22}$$

$$\bar{X}^{\mu}_{i,T_i(t)} - \sqrt{\tfrac{\alpha}{T_i(t)}} > 0,\ \bar{X}^{\mu}_{1,T_1(t)} - \sqrt{\tfrac{\alpha}{T_1(t)}} > 0.$$

Next we will analyze the above two cases:
For Case (i): Since ξ_{RCB} holds, we have $|\bar{X}^{\mu}_{i,T_i(t)} - \mu_i| < x_0\sqrt{\frac{\alpha}{T_i(t)}}$, which implies that $\bar{X}^{\mu}_{i,T_i(t)} > \mu_i - x_0\sqrt{\frac{\alpha}{T_i(t)}}$. Therefore, we obtain that $T_i(t) \leq \frac{(1+x_0)^2\alpha}{\mu_i^2}$. As a result, we have

$$T_i(t+1) \leq \frac{(1+x_0)^2\alpha}{\mu_i^2} + 1 \leq \frac{(1+x_0)^2\alpha}{\varrho_i^2} + 1. \tag{23}$$

For Case (ii): In this case, if $T_i(t) \leq (1+x_0)^2\alpha/\mu_i^2$, we can obtain Lemma 4 by (23). Therefore, in the left content for the proof based on Case (ii), we only need to consider $T_i(t) > \frac{(1+x_0)^2\alpha}{\mu_i^2}$, which implies that $\mu_i - (1+x_0)\sqrt{\frac{\alpha}{T_i(t)}} > 0$. Conditioned on ξ_{RCB}, we have $|\bar{X}^{\mu}_{j,T_j(t)} - \mu_j| < x_0\sqrt{\frac{\alpha}{T_j(t)}}$ for any $j \in [K]$. We also have that $\bar{X}^{\mu}_{j,T_j(t)} - \sqrt{\frac{\alpha}{T_j(t)}} > 0$ for any $j \in \{1, i\}$. They jointly imply that

$$0 < \bar{X}^{\mu}_{1,T_1(t)} - \sqrt{\frac{\alpha}{T_1(t)}} \leq \mu_1 - (1-x_0)\sqrt{\frac{\alpha}{T_1(t)}}.$$

Thus, conditioned on ξ_{RCB}, we have

$$\frac{\lambda_i + (1+x_0)\sqrt{\frac{\alpha}{T_i(t)}}}{\mu_i - (1+x_0)\sqrt{\frac{\alpha}{T_i(t)}}} \geq \frac{\bar{X}^{\lambda}_{i,T_i(t)} + \sqrt{\frac{\alpha}{T_i(t)}}}{\bar{X}^{\mu}_{i,T_i(t)} - \sqrt{\frac{\alpha}{T_i(t)}}}$$

$$\geq \frac{\bar{X}^{\lambda}_{1,T_1(t)} + \sqrt{\frac{\alpha}{T_1(t)}}}{\bar{X}^{\mu}_{1,T_1(t)} - \sqrt{\frac{\alpha}{T_1(t)}}} \geq \frac{\lambda_1 + (1-x_0)\sqrt{\frac{\alpha}{T_1(t)}}}{\mu_1 - (1-x_0)\sqrt{\frac{\alpha}{T_1(t)}}} > \frac{\lambda_1}{\mu_1} > \frac{\lambda_i + \varrho_i}{\mu_i - \varrho_i},$$

which shows that $T_i(t) \leq (1+x_0)^2\alpha/\varrho_i^2$. Accordingly, $T_i(t+1) \leq [(1+x_0)^2\alpha/\varrho_i^2] + 1$. Lemma 4 is proved. □

($\mathcal{S}2$) *Problem Transformation*: If we can prove the following proposition, then we can conclude Theorem 3.

PROPOSITION 6. *Given ξ_{RCB} is true, for any $i \in [K]$, we have $x_0\sqrt{\frac{\alpha}{T_i(n)}} \leq \varrho_i$.*

This is because if Proposition 6 holds, given ξ_{RCB} is true, for any $i \geq 2$, we have that $\bar{X}^{\mu}_{i,T_i(n)} > \mu_i - x_0\sqrt{\frac{\alpha}{T_i(n)}} > \varrho_i - x_0\sqrt{\frac{\alpha}{T_i(n)}} \geq 0$. And further, we can obtain

$$\begin{aligned}\frac{\bar{X}^{\lambda}_{1,T_1(n)}}{\bar{X}^{\mu}_{1,T_1(n)}} &> \frac{\lambda_1 - x_0\sqrt{\frac{\alpha}{T_1(n)}}}{\mu_1 + x_0\sqrt{\frac{\alpha}{T_1(n)}}} \geq \frac{\lambda_1 - \varrho_i}{\mu_1 + \varrho_i} \\ &= \frac{\lambda_i + \varrho_i}{\mu_i - \varrho_i} \geq \frac{\lambda_i + x_0\sqrt{\frac{\alpha}{T_i(n)}}}{\mu_i - x_0\sqrt{\frac{\alpha}{T_i(n)}}} > \frac{\bar{X}^{\lambda}_{i,T_i(n)}}{\bar{X}^{\mu}_{i,T_i(n)}},\end{aligned} \tag{24}$$

which means that action 1 can be eventually recommended. That is, by Proposition 6, ξ_{RCB} holds means that the best action can be recommended. Therefore, $\mathscr{E}_n(\mathcal{M}_{\text{RCB}})$ is smaller than $\mathbb{P}\{\bar{\xi}_{\text{RCB}}\}$, which has been shown in (20).

($\mathcal{S}3$) *Proof of Proposition* 6: By Lemma 4 and (19),

$$\begin{aligned}&T_1(n) - 1 = n - 1 - \sum_{i=2}^{K} T_i(n) \geq n - K - \sum_{i=2}^{K} \frac{\alpha(1+x_0)^2}{\varrho_i^2} \\ &\geq \frac{16\alpha}{9}\max\{\frac{1}{\mu_1^2}, \frac{1}{\varrho_1^2}\} + \frac{16\alpha}{9}\sum_{i=2}^{K}\frac{1}{\varrho_i^2} - \sum_{i=2}^{K}\frac{\alpha(1+x_0)^2}{\varrho_i^2} \\ &> \alpha(1+x_0)^2\max\{\mu_1^{-2}, \varrho_1^{-2}\}.\end{aligned} \tag{25}$$

Therefore, when $t = n$, the (b) in Lemma 5 does not hold and thus, the (a) in Lemma 5 must hold.
We can verify that for any $i \geq 2$,

$$\begin{cases} \frac{\alpha}{4}(1-x_0)^2(\lambda_1+\mu_1)^2\frac{1}{(\mu_i\mu_1\Delta_i)^2} \geq \frac{\alpha x_0^2}{\varrho_i^2}; \\ \frac{\alpha}{4}(1-x_0)^2(\lambda_1+\mu_1)^2\frac{\alpha(1+x_0)^2}{\varrho_1^2\alpha(1+x_0)^2(\lambda_i+\mu_i)^2} \geq \frac{\alpha x_0^2}{\varrho_i^2}. \end{cases} \tag{26}$$

According to (25), (26) and Lemma 5, we can conclude that for any $i \in [K]$, $T_i(n) \geq \frac{\alpha x_0^2}{\varrho_i^2}$, i.e., Proposition 6 holds. □

(17) suggests that $\mathscr{E}_n(\mathcal{M}_{\text{RCB}})$ is bounded by $O(n\exp\{-cn\})$ where c is a constant independent of n. $O(n\exp\{-cn\})$ is bounded by $O(\exp\{-\tilde{c}n\})$ for any $\tilde{c} \in (0, c)$. Therefore, like $\mathscr{E}_n(\mathcal{M}_{\text{U}})$ and $\mathscr{E}_n(\mathcal{M}_{\text{SE}})$, $\mathscr{E}_n(\mathcal{M}_{\text{RCB}})$ decreases exponentially in terms of n.

From Theorem 3, one may notice that the upper bound in (16) seems to increase w.r.t n, which is counterintuitive. Actually, this impression is not correct. Parameter α can also depend on n, and in particular, when $\alpha = \frac{9}{16}\frac{n-K}{H_1}$, the upper bound in (17) will decrease exponentially w.r.t. n.

Based on Theorem 1, 2 and 3, we can derive certain sample complexities that make the error probabilities of the three mechanisms smaller than $\varepsilon \in (0,1)$:

COROLLARY 7. *For any $\varepsilon \in (0,1)$, to find the best action with probability at least $1-\varepsilon$, a total of $O(K\varrho_1^{-2}\ln(1/\varepsilon))$ rounds of selection suffices for $\mathcal{M}_U$, $O((H_2\ln(2K))\ln(1/\varepsilon))$ for $\mathcal{M}_{SE}$, and $O(R(\varepsilon)\ln R(\varepsilon))$ for $\mathcal{M}_{RCB}$ with $\alpha = (9(n-K))/(16H_1)$, where $R(\varepsilon) = (H_1/(\lambda_1+\mu_1)^2)\ln(1/\varepsilon)$.*

According to Corollary 7, if we are only interested in the complexity related to ε, all the three mechanisms are $O(\ln\frac{1}{\varepsilon})$; $\mathcal{M}_{\text{SE}}$ and $\mathcal{M}_{\text{RCB}}$ need less sample complexity than $\mathcal{M}_{\text{U}}$ when ϱ_1 is small and K is large.

5. LOWER BOUND FOR THE BERNOULLI SETTING

In this section, we provide an asymptotic lower bound for consistent mechanisms when both the rewards and costs of all the actions are sampled from Bernoulli distributions. For ease of references, we call such settings *Bernoulli settings*, which are a common choice for lower bound analysis for bandit problems. Similar to [16], we say a mechanism $\mathcal{M}$ is *consistent* if for every underlying reward/cost distribution, for any $\varepsilon \in (0,1)$, there exists an integer $n(\varepsilon) > 0$, such that for any $n > n(\varepsilon)$, $\mathscr{E}_n(\mathcal{M}) < \varepsilon$. That is, for a consistent mechanism $\mathcal{M}$, $\mathscr{E}_n(\mathcal{M})$ converges to zero as n tends to infinity.

Define the KL divergence of two Bernoulli distributions with parameter x and y as follows:

$$D(x\|y) = x\ln\frac{x}{y} + (1-x)\ln\frac{1-x}{1-y}, \text{ where } x, y \in (0,1). \quad (27)$$

Furthermore, for any $i \geq 2$ and $\gamma \geq 0$, define the δ-gap $\delta_i(\gamma)$ as follows:

$$\delta_i(\gamma) = \frac{\mu_1\mu_i\Delta_i}{\gamma\lambda_1 + \mu_1}. \quad (28)$$

$\delta_i(\gamma)$ can be seen as an asymmetric adaptive version of Δ_i: to make suboptimal action i the best action, one can increase λ_i by $\delta_i(\gamma)$ while decreasing the μ_i by $\gamma\delta_i(\gamma)$. Please note that when $\gamma = 1$, the asymmetric δ-gap becomes symmetric. Furthermore, it is easy to verify that $\frac{\lambda_1}{\mu_1} = \frac{\lambda_i+\delta_i(\gamma)}{\mu_i-\gamma\delta_i(\gamma)}$.

The following theorem shows that for Bernoulli settings, the asymptotic lower bound of error probability of any consistent mechanism is $\Omega(\exp\{-\mathcal{D}_* n\})$, which decreases exponentially in terms of n. ($\mathcal{D}_*$ is defined in Theorem 8.)

THEOREM 8. *Considering best action selection with Bernoulli settings, for any consistent mechanism $\mathcal{M}$, we have*

$$\overline{\lim}_{n\to\infty} -\frac{1}{n}\ln\mathscr{E}_n(\mathcal{M}) \leq \mathcal{D}_*,$$

in which $\mathcal{D}_* = \inf_{(j,\gamma)\in\Gamma}\{D(\lambda_j+\delta_j(\gamma)\|\lambda_j)+D(\mu_j-\gamma\delta_j(\gamma)\|\mu_j)\}$ *and* $\Gamma = \{(j,\gamma_j)|j \in [K]\backslash\{1\}, \lambda_j + \delta_j(\gamma_j) < 1, \gamma_j \geq 0\}$. *Note that $\mathcal{D}_*$ always exists and is positive.*

PROOF. Define $\hat{D}_{\nu,\upsilon}$ as follows:

$$\begin{aligned}\hat{D}_{\nu,\upsilon} = &\sum_{i=1}^{K}\sum_{t=1}^{n}\ln\frac{\nu_i^\lambda X_{i,t}^\lambda + (1-\nu_i^\lambda)(1-X_{i,t}^\lambda)}{\upsilon_i^\lambda X_{i,t}^\lambda + (1-\upsilon_i^\lambda)(1-X_{i,t}^\lambda)}\\ &+\sum_{i=1}^{K}\sum_{t=1}^{n}\ln\frac{\nu_i^\mu X_{i,t}^\mu + (1-\nu_i^\mu)(1-X_{i,t}^\mu)}{\upsilon_i^\mu X_{i,t}^\mu + (1-\upsilon_i^\mu)(1-X_{i,t}^\mu)},\end{aligned} \quad (29)$$

in which (1) $\nu_i^\lambda, \nu_i^\mu, \upsilon_i^\lambda, \upsilon_i^\mu \in (0,1)$ for any $i \in [K]$; (2) $X_{i,t}^\lambda, X_{i,t}^\mu \in \{0,1\}$ are the reward and the cost of action i at round t; (3) ν and υ denote the two product Bernoulli distributions. Mathematically,

$$\begin{aligned}\nu &= \otimes_{i=1}^K \text{Bern}(\nu_i^\lambda) \times \otimes_{i=1}^K \text{Bern}(\nu_i^\mu),\\ \upsilon &= \otimes_{i=1}^K \text{Bern}(\upsilon_i^\lambda) \times \otimes_{i=1}^K \text{Bern}(\upsilon_i^\mu),\end{aligned} \quad (30)$$

where[6] $\text{Bern}(p)$ represents the Bernoulli distribution with success probability p.

[6]For simplicity, $\otimes_{i=1}^K\text{Bern}(\nu_i^\lambda)$ represents a product of K independent distributions, where the i-th one is $\text{Bern}(\nu_i^\lambda)$. Both $\otimes$ and $\times$ represent the products of distributions.

Let $\mathcal{F}_n$ denote the σ-algebra generated by $\{X_{i,t}^\lambda\}_{i\in[K],t\in[n]}$, $\{X_{i,t}^\mu\}_{i\in[K],t\in[n]}$ and the selection history. According to the Lemma 15 of [16], we can obtain that for any $A \in \mathcal{F}_n$,

$$\begin{aligned}\mathbb{P}_\upsilon\{A\} &= \mathbb{E}_\nu\{\exp(-\hat{D}_{\nu,\upsilon})\mathbf{1}\{A\}\}\\ &= \mathbb{E}_\nu\{\exp(-\hat{D}_{\nu,\upsilon})\mathbf{1}\{A\}|\mathbf{1}\{A\}=1\}\mathbb{P}_\nu\{A\}\\ &= \mathbb{E}_\nu\{\exp(-\hat{D}_{\nu,\upsilon})|\mathbf{1}\{A\}=1\}\mathbb{P}_\nu\{A\}\\ &= \mathbb{E}_\nu\{\exp(-\hat{D}_{\nu,\upsilon})|A\}\mathbb{P}_\nu\{A\} \geq \exp(-\mathbb{E}_\nu[\hat{D}_{\nu,\upsilon}|A])\mathbb{P}_\nu\{A\}.\end{aligned}$$

Accordingly, we have

$$\mathbb{E}_\nu[\hat{D}_{\nu,\upsilon}|A] \geq \ln\frac{\mathbb{P}_\nu\{A\}}{\mathbb{P}_\upsilon\{A\}}. \quad (31)$$

Since $A \in \mathcal{F}_n$, $\bar{A} \in \mathcal{F}_n$. By substituting the A in (31) with $\bar{A}$, we have

$$\mathbb{E}_\nu[\hat{D}_{\nu,\upsilon}|\bar{A}] \geq \ln\frac{\mathbb{P}_\nu\{\bar{A}\}}{\mathbb{P}_\upsilon\{\bar{A}\}}.$$

Then we can obtain that

$$\begin{aligned}\mathbb{E}_\nu\{\hat{D}_{\nu,\upsilon}\} &= \mathbb{E}_\nu\{\hat{D}_{\nu,\upsilon}|A\}\mathbb{P}_\nu\{A\} + \mathbb{E}_\nu\{\hat{D}_{\nu,\upsilon}|\bar{A}\}\mathbb{P}_\nu\{\bar{A}\}\\ &\geq \mathbb{P}_\nu\{A\}\ln\frac{\mathbb{P}_\nu\{A\}}{\mathbb{P}_\upsilon\{A\}} + \mathbb{P}_\nu\{\bar{A}\}\ln\frac{\mathbb{P}_\nu\{\bar{A}\}}{\mathbb{P}_\upsilon\{\bar{A}\}}\\ &= D(\mathbb{P}_\nu\{A\}\|\mathbb{P}_\upsilon\{A\}).\end{aligned} \quad (32)$$

By some derivations, one can verify that

$$\mathbb{E}_\nu\{\hat{D}_{\nu,\upsilon}\} = n\sum_{i=1}^{K} D(\nu_i^\lambda\|\upsilon_i^\lambda) + n\sum_{i=1}^{K} D(\nu_i^\mu\|\upsilon_i^\mu). \quad (33)$$

Combining (32) and (33), we have

$$n\sum_{i=1}^{K} D(\nu_i^\lambda\|\upsilon_i^\lambda) + n\sum_{i=1}^{K} D(\nu_i^\mu\|\upsilon_i^\mu) \geq D(\mathbb{P}_\nu\{A\}\|\mathbb{P}_\upsilon\{A\}). \quad (34)$$

To get the lower bound of the error probability for consistent mechanisms, we specialize A, ν and υ as follows:

$A = \{J_n = 1\}$; (It is obvious that $A \in \mathcal{F}_n$;)

$\upsilon = \otimes_{i=1}^K\text{Bern}(\lambda_i) \times \otimes_{i=1}^K\text{Bern}(\mu_i)$; $\nu = \nu^\lambda \otimes \nu^\mu$;

$\nu^\lambda = \otimes_{i=1}^{j-1}\text{Bern}(\lambda_i) \otimes \text{Bern}(\lambda_j+\delta_j(\gamma)+\varepsilon) \otimes_{i=j+1}^K \text{Bern}(\lambda_i)$;

$\nu^\mu = \otimes_{i=1}^{j-1}\text{Bern}(\mu_i) \otimes \text{Bern}(\mu_j-\gamma\delta_j(\gamma)-\varepsilon) \otimes_{i=j+1}^K \text{Bern}(\mu_i)$,

in which the γ and ε are constraint by $\lambda_j+\delta_j(\gamma)+\varepsilon \in (0,1)$ and $\mu_j - \gamma\delta_j(\gamma) - \varepsilon \in (0,1)$, and $j \geq 2$.

For any consistent mechanism $\mathcal{M}$ and any $\varepsilon \in (0,1)$, we can always find a large enough number $n(\varepsilon)$ such that for any $n \geq n(\varepsilon)$, $\mathbb{P}_\nu\{A\} \leq \varepsilon \leq \mathbb{P}_\upsilon\{A\}$. Therefore, by setting the notations in (34) with the corresponding ones specified above, we have that for any $j \geq 2$,

$$\begin{aligned}&D(\lambda_j+\delta_j(\gamma)+\varepsilon\|\lambda_j) + D(\mu_j-\gamma\delta_j(\gamma)-\varepsilon\|\mu_j)\\ \geq&\frac{1}{n}D(\mathbb{P}_\nu\{A\}\|\mathbb{P}_\upsilon\{A\}) \geq \frac{1}{n}D(\varepsilon\|\mathbb{P}_\upsilon\{A\})\\ \geq&\frac{1}{n}(\varepsilon\ln\varepsilon + (1-\varepsilon)\ln\frac{1-\varepsilon}{\mathscr{E}_n}).\end{aligned} \quad (35)$$

Let $J_j(\gamma)$ denote $D(\lambda_j+\delta_j(\gamma)\|\lambda_j) + D(\mu_j-\gamma\delta_j(\gamma)\|\mu_j)$ for any $(j,\gamma) \in \Gamma$ (defined in Theorem 8). Taking the supremum limit with $\varepsilon \to 0$ on both sides of (35), we have

$$\overline{\lim}_{n\to\infty} -\frac{1}{n}\ln\mathscr{E}_n \leq J_j(\gamma). \quad (36)$$

Finally, we need to prove the existence of $\mathcal{D}_*$. For any $j \geq 2$, denote the domain of $J_j(\gamma)$ as Γ_j, which is $\{\gamma|(j,\gamma) \in \Gamma\}$. One can verify that for any $j \geq 2$, in Γ_j, $J_j(\gamma)$ either increases w.r.t γ, or first decreases then increases w.r.t γ. By careful derivations, we can get that $\inf_{\gamma\in\Gamma_j} J_j(\gamma)$ can be achieved at some $\gamma_j^* \in \Gamma_j$. Since K is limited, we can always find the $\mathcal{D}_*$, which is $\min_{j\geq 2} J_j(\gamma_j^*)$. □

5.1 Discussions

Generally speaking, the upper bounds of the error probabilities for the proposed three mechanisms match the lower bound given in Theorem 8, in terms of the order of n. To gain more understanding of the relationship between the upper bounds and the lower bound, we also need to check the constants in the bounds. For this purpose, we investigate how the action-related coefficients (like Δ_i) behave in the lower and upper bounds of the error probabilities. We take $\mathcal{M}_{\text{SE}}$ as an example to give detailed discussions. We can get similar conclusions for $\mathcal{M}_{\text{U}}$ and $\mathcal{M}_{\text{RCB}}$.

First, let us have a look at how the error probability $\mathscr{E}_n$ depends on Δ_i, the sub-optimality of action i compared to the best action. Consider the following example: There is a Bernoulli setting with the following parameters: for any $i \in [K]$, $\lambda_i, \mu_i \in [p, 1-p]$, in which $p \in (0, \frac{1}{2})$. We can bound $\mathcal{D}_*$ by setting $\gamma \to \infty$ in Theorem 8 as below.

$$\mathcal{D}_* \le \frac{(\mu_1\mu_i\Delta_i)^2}{\lambda_1^2\mu_i(1-\mu_i)} \le \frac{\Delta_i^2}{p^3}, \forall i \ge 2. \tag{37}$$

Define $\Delta_{\min} = \min_{i\ge 2}\Delta_i$. By (37), the lower bound of the error probability of any consistent mechanism is

$$\mathscr{E}_n \ge \Omega\Big(\exp\Big\{-\frac{1}{p^3}\Delta_{\min}^2 n\Big\}\Big). \tag{38}$$

The upper bound of $\mathscr{E}_n(\mathcal{M}_{\text{SE}})$ is:

$$\mathscr{E}_n(\mathcal{M}_{\text{SE}}) \le O\Big(\exp\Big\{-\frac{p^4}{8K\ln(2K)}\Delta_{\min}^2 n\Big\}\Big). \tag{39}$$

(38) and (39) share the common term $\Delta_{\min}^2 n$ within the exponential order. Therefore, as $\Delta_{\min}$ becomes smaller, both the lower bound of the error probability of any consistent mechanism and the upper bound of $\mathscr{E}_n(\mathcal{M}_{\text{SE}})$ increase.

Then we study how the error probabilities of the mechanisms change w.r.t. the magnitude of the parameters of the actions. For ease of reference, we denote the parameters of the Bernoulli K-action setting as $\{\lambda_i, \mu_i\}_{i=1}^K$. Let us check three variants of Bernoulli action setting by introducing a hyper parameter $\theta \in (0, 1]$. The three variants of the settings are $(\mathcal{V}_1)$ $\{\theta\lambda_i, \mu_i\}_{i=1}^K$; $(\mathcal{V}_2)$ $\{\lambda_i, \theta\mu_i\}_{i=1}^K$; $(\mathcal{V}_3)$ $\{\theta\lambda_i, \theta\mu_i\}_{i=1}^K$. The θ for the variant $(\mathcal{V}_1)$ should be smaller than $\min\{\mu_1/(\lambda_1\mu_i), 1\}$ for any $i \ge 2$. We get that the asymptotic lower bound (related to n and θ simultaneously) of any consistent mechanism for the three variants is

$$\mathscr{E}_n \ge \Omega(\exp\{-L(\mathcal{V}_i)\theta n\}, \qquad i \in \{1,2,3\}, \tag{40}$$

where
$$L(\mathcal{V}_1) = \min_{i\ge 2}\frac{(\lambda_1\mu_i - \lambda_i\mu_1)^2}{\mu_1^2\lambda_i(1-\lambda_i)}; \tag{41}$$

$$L(\mathcal{V}_2) = L(\mathcal{V}_3) = \min_{i\ge 2}\frac{(\lambda_1\mu_i - \lambda_i\mu_1)^2}{\lambda_1^2\mu_i(1-\mu_i)}. \tag{42}$$

(41) and (42) can be obtained by a similar derivation to that of (37). Specifically, we need to set γ in Theorem 8 as 0 to get (41) and as ∞ to get (42).

The upper bounds of the error probabilities $\mathcal{M}_{\text{SE}}$ under the three variants share the following common form:

$$O\Big(\exp\Big\{-\min_{i\ge 2}\Big(\frac{\lambda_1\mu_i - \lambda_i\mu_1}{\lambda_1+\lambda_i+\mu_1+\mu_i}\Big)^2\frac{2\theta^2 n}{K\ln(2K)}\Big\}\Big). \tag{43}$$

From (40) and (43), we come to two conclusions: (1) Decreasing the expected rewards (or, the expected costs) of all the actions lead to a larger lower bound and a larger upper bound of the error probabilities. (2) If the expected rewards and the expected costs decrease simultaneously, the lower bound of the error probability of any consistent mechanism becomes larger and the upper bound of the error probability of $\mathcal{M}_{\text{SE}}$ also becomes larger.

6. TWO DEGENERATED CASES

In this section, we consider the two degenerated cases as follows[7]. For any $i \in [K]$ and $t \in [n]$,

(Case 1): the rewards are random variables but the costs are deterministic, i.e., $X_{i,t}^{\mu} = \mu_i$.

(Case 2): the rewards are deterministic but the costs are random variables, i.e., $X_{i,t}^{\lambda} = \lambda_i$.

Our proposed three mechanisms can be applied to solve the above cases: (1) Both $\mathcal{M}_{\text{U}}$ and $\mathcal{M}_{\text{SE}}$ can be directly applied to the two degenerated cases. (2) $\mathcal{M}_{\text{RCB}}$ can be applied with minor revisions. Since we do not need to introduce confidence bounds for deterministic observations,

For Case 1, we set $\hat{\mu}_i(t)$ in (4) as μ_i;

For Case 2, we set $\hat{\lambda}_i(t)$ in (4) as λ_i.

We extend the definition of ϱ_i for any $i \ge 2$ by introducing two binary variables $\beta_\lambda, \beta_\mu \in \{0, 1\}$:

$$\varrho_i(\beta_\lambda, \beta_\mu) = \frac{\mu_1\mu_i\Delta_i}{\beta_\mu(\lambda_1+\lambda_i) + \beta_\lambda(\mu_1+\mu_i)}. \tag{44}$$

Similarly, define $\varrho_1(\beta_\lambda, \beta_\mu) = \min_{i\ge 2}\varrho_i(\beta_\lambda, \beta_\mu)$. We can regard β_λ and β_μ as binary variables, which represent whether the rewards and costs are random variables or not ("1" means "yes"). We redefine H_1 and H_2 as $H_1(\beta_\lambda, \beta_\mu)$ and $H_2(\beta_\lambda, \beta_\mu)$ with the corresponding $\varrho_i(\beta_\lambda, \beta_\mu)$'s.

The error probabilities for the two degenerated cases can be bounded in unified forms using β_λ and β_μ. For $\mathcal{M}_{\text{U}}$ and $\mathcal{M}_{\text{SE}}$, we have

$$\mathscr{E}_n(\mathcal{M}_{\text{U}}) \le (\beta_\lambda+\beta_\mu)\sum_{i=1}^K e^{2\varrho_i^2(\beta_\lambda,\beta_\mu)}\exp\{-2\varrho_i^2(\beta_\lambda,\beta_\mu)\frac{n}{K}\}.$$

$$\mathscr{E}_n(\mathcal{M}_{\text{SE}}) \le (\beta_\lambda+\beta_\mu)K(K-1)\exp\Big\{-\frac{2(n-K)}{\ln(2K)H_2(\beta_\lambda,\beta_\mu)}\Big\}.$$

By setting $0 < \alpha \le \frac{9}{16}\frac{n-K}{H_1(\beta_\lambda,\beta_\mu)}$ for $\mathcal{M}_{\text{RCB}}$, we get

$$\mathscr{E}_n(\mathcal{M}_{\text{RCB}}) \le 2(\beta_\lambda+\beta_\mu)nK\exp\Big\{-\frac{2\alpha(\beta_\mu\lambda_1+\beta_\lambda\mu_1)^2}{25(\beta_\lambda+\beta_\mu)^2}\Big\}.$$

In particular, for $\alpha = \frac{9}{16}\frac{n-K}{H_1(\beta_\lambda,\beta_\mu)}$, we have

$$\mathscr{E}_n(\mathcal{M}_{\text{RCB}}) \le 2(\beta_\lambda+\beta_\mu)nK\exp\Big\{-\frac{9(n-K)(\beta_\mu\lambda_1+\beta_\lambda\mu_1)^2}{200(\beta_\lambda+\beta_\mu)^2H_1(\beta_\lambda,\beta_\mu)}\Big\}.$$

We can verify that by substituting both the β_λ and the β_μ with 1 in the above formulas, (which means that the rewards and costs are random) the obtained error probabilities are the same as the corresponding ones in Theorem 1, 2 and 3.

The lower bounds for the degenerated cases are shown below: (45) for Case 1 (i.e., $\beta_\lambda = 1, \beta_\mu = 0$) and (46) for Case 2 (i.e., $\beta_\lambda = 0, \beta_\mu = 1$).

$$\mathscr{E}_n \ge \Omega\Big(\exp\{-D(\lambda_1 - \mu_1\Delta_{\min}\|\lambda_1)n\}\Big); \tag{45}$$

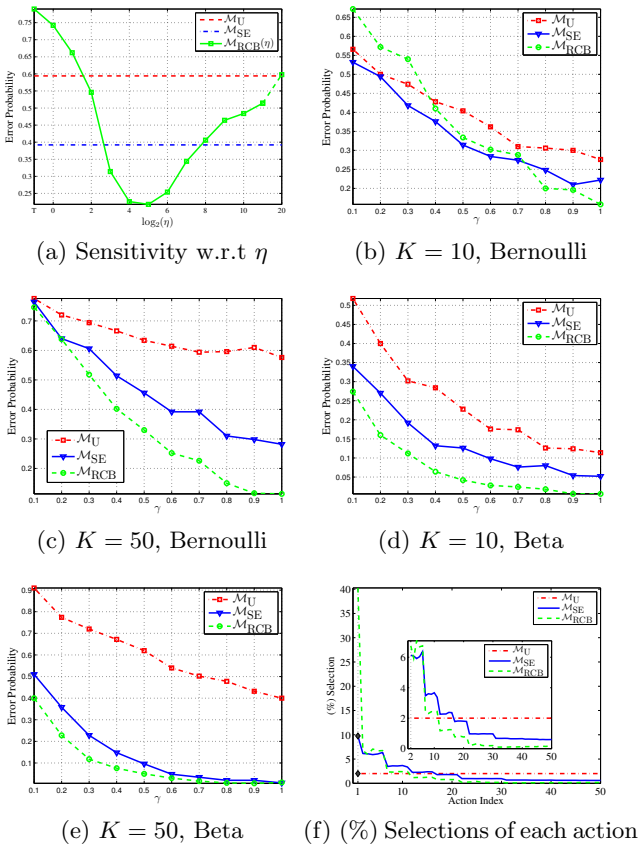
$$\mathscr{E}_n \ge \Omega\Big(\exp\{-\min_{i\ge 2}D\Big(\mu_i - \frac{\mu_1\mu_i\Delta_i}{\lambda_1}\|\mu_i\Big)n\}\Big). \tag{46}$$

Note that for simplicity, (45) is obtained with an additional setting that $\lambda_1/\mu_1 > \lambda_2/\mu_2 > \lambda_i/\mu_i$ for any $i \ge 3$. We can also refine the discussions in Section 5 to get their counterparts for the two degenerated cases.

Furthermore, by setting $\mu_i = 1$ for all i, our best action selection problem will degenerate to the conventional best arm identification problem in [1]. The coefficients before n in $\exp\{\cdots n\}$ of our upper bounds for $\mathscr{E}_n(\mathcal{M}_{\text{SE}})$ and $\mathscr{E}_n(\mathcal{M}_{\text{RCB}})$ are up to constant factors to those in [1].

[7]We do not consider the case that both the rewards and the costs are deterministic because it is trivial to identify the best action after taking each action once.

7. EMPIRICAL PERFORMANCE



(a) Sensitivity w.r.t η (b) $K = 10$, Bernoulli

(c) $K = 50$, Bernoulli (d) $K = 10$, Beta

(e) $K = 50$, Beta (f) (%) Selections of each action

Figure 1: Experimental Results

In this section, we test the empirical performance of the proposed mechanisms through numerical simulations. We repeat each experiment for 500 times and use the relative frequencies (normalized by 500) that a mechanism does not recommend the best action as a proxy of $\mathscr{E}_n$.

We first simulated a 10-action Bernoulli setting and a 50-action Bernoulli setting. In the 10-action setting, for any $i \in \{1, 2, \cdots, 5\}$, $\lambda_i = 0.4+0.1i, \mu_i = \lambda_i - 0.1$; $\lambda_6 = \mu_6 = 0.9$; for any $i \in \{7, 8, 9, 10\}$, $\lambda_i = 1.5 - 0.1i$, $\mu_i = \lambda_i + 0.1$. In the 50-action setting, the parameters are randomly duplicated while keeping the best action unique. We set $N_{10} = 10000$ and $N_{50} = 50000$. For the x-action setting, the number of selection rounds is set as γN_x, where $x \in \{10, 50\}$ and $\gamma \in \{0.1, 0.2, \cdots, 1\}$.

$\mathcal{M}_{\mathrm{RCB}}$ is associated with a hyper-parameter α. According to Theorem 3, and the previous literature [1, 13], we set $\alpha = \eta \frac{n}{H_1}$, where η is a hyper-parameter. We first study the sensitivity of the error probability w.r.t η for $\mathcal{M}_{\mathrm{RCB}}$. We take the 50-action Bernoulli setting with selection rounds $0.7N_{50}$ for example. The η is chosen from $\{T, 2^0, 2^1, 2^2, \cdots, 2^{10}, 2^{11}, 2^{20}\}$, where $T = 9/16$. T is the parameter suggested in Theorem 3. From Figure 1(a), we can see that $\mathscr{E}_n(\mathcal{M}_{\mathrm{RCB}})$ is sensitive to the choice of the hyper-parameter, which is a drawback of $\mathcal{M}_{\mathrm{RCB}}$. However, on the other side, this is the flexibility of $\mathcal{M}_{\mathrm{RCB}}$, since by tuning the η, $\mathcal{M}_{\mathrm{RCB}}$ can achieve very good results. One should not set η too small (like $T, 1, 2$) when n is not large enough, for which the mechanism can be easily trapped in the empirical best selection and make little exploration. The η should not be set too large (like $2^{10}, 2^{20}$), for which $\mathcal{M}_{\mathrm{RCB}}$ behaves like $\mathcal{M}_{\mathrm{U}}$. Therefore we choose $\eta = 16$ for all the following experiments.

The results for the two Bernoulli settings are shown in Figure 1(b) and Figure 1(c). From the figures we have several observations. (1) As n increases, the error probabilities of all the mechanisms decrease. (2) $\mathcal{M}_{\mathrm{SE}}$ outperforms $\mathcal{M}_{\mathrm{U}}$. Though $\mathcal{M}_{\mathrm{SE}}$ does not need to know anything about the bandit, the error probabilities of $\mathcal{M}_{\mathrm{SE}}$ are not much worse than $\mathcal{M}_{\mathrm{RCB}}$. (3) $\mathcal{M}_{\mathrm{RCB}}$ performs the best among the three mechanisms. (4) $\mathcal{M}_{\mathrm{U}}$ is not as good as other two mechanisms since it wastes too many selections on the actions that are certainly not the best one. Note that the error probability of $\mathcal{M}_{\mathrm{U}}$ is acceptable for a sufficiently large n if the number of actions is not large (like Figure 1(b)), because it can still identify the best action with high probability even if it wastes some rounds of selections. (5) Fixing n, the error probabilities of the three mechanisms increase w.r.t K. Take $n = 10^4$ for example: the error probabilities of $\mathcal{M}_{\mathrm{U}}, \mathcal{M}_{\mathrm{SE}}, \mathcal{M}_{\mathrm{RCB}}$ are $0.28, 0.22, 0.16$ for $K = 10$, and $0.72, 0.64, 0.63$ for $K = 50$. This is not surprising, since more actions make the optimal one harder to be distinguished from others.

We also simulate a 10-action beta setting and a 50-action beta setting, whose rewards and costs follow beta distributions. From Figure 1(d) and Figure 1(e), we can get similar observations as those for the Bernoulli settings.

At last we study the selection frequency of each action for the 50-action Bernoulli setting with $n = 0.7N_{50}$ (See Figure 1(f)). The x-axis is the action index ordered by the expected reward to the expected cost ratio in non-increasing orders. (The x-axis of the small figure starts from action 2.) We can see that both $\mathcal{M}_{\mathrm{SE}}$ and $\mathcal{M}_{\mathrm{RCB}}$ spend most of the selections on the best action and those close to the best one, which are the most uncertain actions. $\mathcal{M}_{\mathrm{RCB}}$ selects the best action most, thus the head of its curve is the heaviest. Next comes $\mathcal{M}_{\mathrm{SE}}$. The curve of $\mathcal{M}_{\mathrm{U}}$ is flat since it allocates each action the same selection rounds.

Overall, the experimental results suggest that $\mathcal{M}_{\mathrm{U}}$ is not so effective as $\mathcal{M}_{\mathrm{SE}}$ and $\mathcal{M}_{\mathrm{RCB}}$ for the best action selection problem in a stochastic environment, and $\mathcal{M}_{\mathrm{RCB}}$ performs the best by carefully setting the hyper parameter.

8. CONCLUSION AND FUTURE WORK

We designed three mechanisms for the best action selection problem in a stochastic environment. We theoretically upper bounded the error probabilities of the proposed three mechanisms and empirically tested their performances. We also provided a lower bound of the error probability of any consistent mechanism with the Bernoulli setting.

We will explore the following directions in the future. First, we will study the setting that the rewards and costs of each action are dependent. Second, we will study the multiple actions identification problem, i.e., select the top-$m(\geq 1)$ best actions like the problems studies in [8, 27]. Third, we defined the best action as the one with the largest ratio of the expected reward to expected cost. It is interesting to consider other definitions of the best action.

Acknowledgments

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

REFERENCES

[1] J.-y. Audibert, S. Bubeck, and R. Munos. Best arm identification in multiarmed bandits. In *23rd annual Conference on Learning Theory*, 2010.

[2] A. Badanidiyuru, R. Kleinberg, and A. Slivkins. Bandits with knapsacks. In *Foundations of Computer Science (FOCS), 2013 IEEE 54th Annual Symposium on*, pages 207–216. IEEE, 2013.

[3] S. Bhat, S. Jain, S. Gujar, and Y. Narahari. An optimal bidimensional multi-armed bandit auction for multi-unit procurement. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 1789–1790, 2015.

[4] A. Biswas, S. Jain, D. Mandal, and Y. Narahari. A truthful budget feasible multi-armed bandit mechanism for crowdsourcing time critical tasks. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 1101–1109, 2015.

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

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

[7] S. Bubeck, R. Munos, and G. Stoltz. Pure exploration in finitely-armed and continuous-armed bandits. *Theoretical Computer Science*, 412(19):1832–1852, 2011.

[8] S. Bubeck, T. Wang, and N. Viswanathan. Multiple identifications in multi-armed bandits. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, 2013.

[9] P.-A. Chen and C.-J. Lu. Playing congestion games with bandit feedbacks. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 1721–1722, 2015.

[10] W. Ding, T. Qin, X.-D. Zhang, and T.-Y. Liu. Multi-armed bandit with budget constraint and variable costs. In *The 27th AAAI Conference on Artificial Intelligenc*, 2013.

[11] E. Even-Dar, S. Mannor, and Y. Mansour. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *The Journal of Machine Learning Research*, 7:1079–1105, 2006.

[12] V. Gabillon, M. Ghavamzadeh, and A. Lazaric. Best arm identification: A unified approach to fixed budget and fixed confidence. In *Advances in Neural Information Processing Systems*, pages 3212–3220, 2012.

[13] V. Gabillon, M. Ghavamzadeh, A. Lazaric, and S. Bubeck. Multi-bandit best arm identification. In *Advances in Neural Information Processing Systems*, pages 2222–2230, 2011.

[14] S. Jain, S. Gujar, O. Zoeter, and Y. Narahari. A quality assuring multi-armed bandit crowdsourcing mechanism with incentive compatible learning. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems (AAMAS)*, pages 1609–1610, 2014.

[15] S. Kalyanakrishnan, A. Tewari, P. Auer, and P. Stone. Pac subset selection in stochastic multi-armed bandits. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pages 655–662, 2012.

[16] E. Kaufmann, O. Cappé, and A. Garivier. On the complexity of best arm identification in multi-armed bandit models. *arXiv preprint arXiv:1407.4443*, 2014.

[17] S. Mannor and J. N. Tsitsiklis. The sample complexity of exploration in the multi-armed bandit problem. *The Journal of Machine Learning Research*, 5:623–648, 2004.

[18] T. Qin, W. Chen, and T.-Y. Liu. Sponsored search auctions: Recent advances and future directions. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(4):60, 2015.

[19] S. Sen, A. Ridgway, and M. Ripley. Adaptive budgeted bandit algorithms for trust development in a supply-chain. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 137–144, 2015.

[20] M. Soare, A. Lazaric, and R. Munos. Best-arm identification in linear bandits. In *Advances in Neural Information Processing Systems*, pages 828–836, 2014.

[21] R. Stranders, L. Tran-Thanh, F. M. D. Fave, A. Rogers, and N. R. Jennings. Dcops and bandits: Exploration and exploitation in decentralised coordination. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 289–296. International Foundation for Autonomous Agents and Multiagent Systems, 2012.

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

[23] L. Tran-Thanh, A. Chapman, A. Rogers, and N. R. Jennings. Knapsack based optimal policies for budget-limited multi-armed bandits. In *The 26th AAAI Conference on Artificial Intelligence*, 2012.

[24] Y. Wu, A. Gyorgy, and C. Szepesvari. On identifying good options under combinatorially structured feedback in finite noisy environments. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 1283–1291, 2015.

[25] Y. Xia, W. Ding, X.-D. Zhang, N. Yu, and T. Qin. Budgeted bandit problems with continuous random costs. In *Proceedings of the 7th Asian Conference on Machine Learning*, 2015.

[26] Y. Xia, H. Li, T. Qin, N. Yu, and T.-Y. Liu. Thompson sampling for budgeted multi-armed bandits. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence*, pages 3960–3966, 2015.

[27] Y. Zhou, X. Chen, and J. Li. Optimal pac multiple arm identification with applications to crowdsourcing. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 217–225, 2014.