Exploring Best Arm with Top Reward-Cost Ratio in Stochastic Bandits

Zhida Qin[†], Xiaoying Gan[†], Jia Liu[‡], Hongqiu Wu[†], Haiming Jin[†], Luoyi Fu[†]
[†]Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai, P. R. China
[‡]Department of Computer Science, Iowa State University, Ames, IA, USA
Email: [†]{zanderqin, ganxiaoying, Wu.Liverpool, jinhaiming, yiluofu}@sjtu.edu.cn, [‡]jialiu@iastate.edu

***Abstract*—The best arm identification problem in multi-armed bandit model has been widely applied into many practical applications, such as spectrum sensing, online advertising, and cloud computing. Although lots of works have been devoted into this area, most of them do not consider the cost of pulling actions, i.e., a player has to pay some cost when she pulls an arm. Motivated by this, we study a ratio-based best arm identification problem, where each arm is associated with a random reward as well as a random cost. For any $\delta \in (0,1)$, with probability at least $1-\delta$, the player aims to find the optimal arm with the largest ratio of expected reward to expected cost using as few samplings as possible. To solve this problem, we propose three algorithms: 1) a genie-aided algorithm GA; 2) the successive elimination algorithm with unknown gaps SEUG; 3) the successive elimination algorithm with unknown gaps and variance information SEUG-V, where gaps denote the differences between the optimal arm and the suboptimal arms. We show that for all three algorithms, the sample complexities, i.e., the pulling times for all arms, grow logarithmically as $\frac{1}{\delta}$ increases. Moreover, compared to existing works, the running of our elimination-type algorithms is independent of the arm-related parameters, which is more practical. In addition, we also provide a fundamental lower bound for sample complexities of any algorithms under Bernoulli distributions, and show that the sample complexities of the proposed three algorithms match that of the lower bound in the sense of $\log \frac{1}{\delta}$. Finally, we validate our theoretical results through numerical experiments.**

I. Introduction

The multi-armed bandit (MAB) problem has been extensively studied in recent decades. Classic MAB formulation is an abstraction of balancing exploration vs. exploitation under uncertainty, which has shown its applications in various areas, such as channel selection [2], [27], medical trials [30], cloud computing [7], and crowdsourcing [4], [20]. Generally, a bandit player faces a set of arms with unknown distributions. Based on the past observations and selection history, the player adaptively pulls arms and receives random reward until some criteria is met or the budget runs out. Depending on different objectives, existing works on MAB can be classified into two categories: i) cumulative regret minimization (the regret denotes the deviations of suboptimal arms to the optimal one); and ii) pure arm exploration. i.e., finding the optimal arm within a fixed budget or a fixed confidence level.

Surprisingly, despite numerous works on pure exploration problems, few works focus on the best arm exploration based on reward to cost ratio, whose application are prevalent in daily life. Take the opportunistic scheduling in wireless communication as an example: A transmitter hopes to choose the best channel among a number of candidates to transmit. However, the channel state information is often unknown to the transmitter and needs to be probed. The quality of each channel is characterized by the SNR (signal to noise ratio). At each probing round, the transmitter sends some pilot symbols to the receiver. The receiver will calculate the SNR value by measuring the corresponding signal and noise power levels (see, e.g., [32] for signal and noise power measurement). Based on the past selection history and observations, the transmitter adaptively choose its sampling in the next round until the channel with the largest empirical SNR value is selected with certain confidence level. This problem can be perfectly formulated as a best arm identification problem with reward to cost ratio. Each channel can be viewed as an arm and its quality depends on the ratio of reward to cost, where its reward corresponds to the signal power and the cost corresponds to the noise power. Besides opportunistic scheduling, there are many other applications that can be modeled as MAB with best reward to cost ratio exploration problem (see more applications in [14]).

Given the significance of the applications, in this paper, we study the best arm identification with reward to cost ratio in stochastic bandit with fixed confidence δ. In our MAB model, each arm is associated with a random reward and a random cost. Specifically, a bandit machine player faces a set of K arms. Once pulling an arm, the player will receive a random reward and pay a random cost from unknown distributions. After a certain pulling times, the player needs to distinguish these K arms and find the one with the top reward to cost ratio with probability at least $1-\delta$. By defining the pulling times of an algorithm as its sample complexity, the objective of the player is to design an algorithm and find the optimal arm within the confidence level and with minimum sample complexity.

To the best of our knowledge, this is the first work to explore the best arm with reward to cost ratio with fixed confidence setting. Considering a ratio-based arm setting brings much challenges for this new problem. This is because during the playing process, the player has to observe both the rewards and costs of arms and estimate their corresponding expectations simultaneously. Such two-dimensional observations will dramatically augment the uncertainty and inevitably result in more samplings to distinguish multiple arms. As a result, the sample complexity for finding the best arm will be significantly increased. Thus, to achieve the designed objective, the player naturally faces *a dilemma*: *how to achieve accurate estimations of both rewards and costs to select the best arm*

and minimize the sample complexity as few as possible? Although many studies have been devoted to the best arm identification problem in MAB model, most of them assume that each arm is only related to a random reward and cannot be directly extended to our settings.

To address these challenges, we adopt the variance information of arms and a refined gaps (i.e., the deviations between suboptimal arms to the optimal arm) to deal with the estimations of expectations for both rewards and costs. The variance information includes the empirical variances of both reward and cost for each arm. The refined gaps, which were proposed in [14], characterize the ratio differences between optimal arm and suboptimal arms from both the reward and cost perspectives. By carefully incorporating these two characteristics into our algorithm design, we achieve the finer-grained empirical means of both rewards and costs with fewer sampling times and develop three algorithms to select the best arm. Further, we prove that our algorithms could not only accurately estimate the expectations of rewards and costs and identify the best arm, but also achieve significantly reduced sample complexity. In addition, compared with the existing Upper Confidence Bound (UCB) type algorithms, our elimination-type algorithms are more challenging as they are independent of the arm-related parameters and more practical.

Our main contributions can be summarized as follows.

- We propose three algorithms with different assumptions of arms' gap knowledge to select the optimal arm with probability at least $1-\delta$ for any $\delta \in (0,1)$. We prove that their sample complexities are all on the order of $O(c\log\frac{1}{\delta})$, where c is a constant depending on problem instances. Our theoretical results are verified by the numerical simulations.
- Compared to existing works, our elimination-type algorithms have two advantages: i) the use of variance information and refined gaps lead to significantly reduced sample complexity; ii) the running of algorithms is independent of the arm distribution parameters, which makes them more applicable to the real world applications.
- Assuming the reward and cost distributions for the arms follow Bernoulli distributions, we provide a fundamental lower bound of the sample complexities of any algorithms. Moreover, we prove that the sample complexity of the lower bound match those of the three proposed algorithms via both theoretical analysis and numerical experiments.

The rest of the paper is organized as follows. In Section II we review the related works. we present the problem formulation in Section III. In Section IV, we propose three algorithms and establish their corresponding sample complexity. A fundamental lower bound is proved in Section V. We conclude the paper in Section VIII.

II. Related Works

The best arm identification problem in MAB has been extensively investigated in the past decades [3], [10], [17], [28]. Generally speaking, these studies can be divided into two settings: fixed confidence and fixed budget settings.

Fixed confidence: In this setting, the player needs to find the best arm under a fixed confidence level δ with minimal sampling times. In the original work [16], Maron *et al.* proposed the Hoeffding races technique to find the best model for training a set of data. Based on this, Mnih *et al.* in [19] and Maurer *et al.* [21] utilized the variance information and proposed Bernstein bounds to further reduce the sample complexity. Later, different variants of MAB models are studied on this setting, such as selecting a top-m arms in one bandit [13], the linear banidts [11], the combinatorial bandits [28] and the bandit with bipartite graphs [33]. Another line of works paid attention to the probably approximately correct (PAC) bound, which aims to find an ϵ-optimal arm with probability of at least $1-\delta$. Even-Dar *et. al* in the seminal work [1], [22] proved that the sample complexity is on the order of $O(\frac{K}{\epsilon^2}\log\frac{1}{\delta})$. Kalyanakrishnan *et. al* in [24], [25] proposed a variant of explore-k problem, and analyzed the upper and lower bounds of sample complexity. Ren. *et. al* in [15] and Chaudhuri *et. al* [8] studied the infinitely multi-armed bandit. Different with their traditional bandit settings, our work is focused on the reward to cost ratio of arms. In addition, different with [24], [25], we apply the changing distributions and KL divergence to analyze the lower bound.

Fixed budget: In this setting, the player aims to find the best arm within a fixed number of sampling. The seminal work is [6], where Madani *et. al* proposed the Budgeted Multi-armed Bandit Problem. Bubeck *et. al* in [18], [29] discussed the relationships between pure exploration and cumulative regret. Later, Audibert *et. al* in [5] proved that the error probabilities are bounded by $O(\exp(-cn))$, where n is the fixed budget and c is arm-related constant. Follow-up works [12], [26], [31] extended the problem into different bandit models. However, none of these works takes the reward to cost ratio of arms into consideration.

The most related line of works to ours are [14] and [9]. Xia. *et. al* in [14] aimed to identify the arm with the largest reward to cost ratio under the fixed budget setting. However, our work differs significantly from theirs in the following key aspects: First, we incorporate the variance information into discarding criteria design of elimination-type algorithm. By doing so, we obtain finer-grained empirical means of both rewards and costs with fewer sampling times. Second, the running of UCB-type algorithm in [14] depends on the arm-related parameters, which may not be realistic in practice. While our algorithms have no such limitations. Last but not the least, our goal is to find optimal arm with a fixed confidence level and a minimal sample complexity. Li *et al* in [9] optimized the cumulative regret based on the reward to cost ratio. In contrast, our work is focused on the best arm identification.

III. Problem Formulation

There are K arms contained in a set $\mathcal{S}$, where $\mathcal{S} = \{1,2,...,K\}$. Each arm $i \in \mathcal{S}$ is characterized by a random reward distribution as well as a random cost distribution, both having a bounded support $[0,1]$. In each round t, if arm i is pulled, the player will receive a random reward value $X_{i,t}^{\mu}$

at a random cost value $X_{i,t}^c$. For each arm i, we assume that $X_{i,t}^\mu$ and $X_{i,t}^c$ are independent with each other, and also independent of the past samples of these two distributions. For arm i, we use μ_i and σ_i^μ to denote the expectation and variance of its reward distribution, i.e., $\mu_i = \mathbb{E}\{X_i^\mu\}$ and $\sigma_i^\mu = \mathbb{E}\{(X_i^\mu - \mu_i)^2\}$. Similarly, the expectation and variance of cost distribution are denoted by c_i and σ_i^c respectively, i.e., $c_i = \mathbb{E}\{X_i^c\}$, $\sigma_i^c = \mathbb{E}\{(X_i^c - c_i)^2\}$. As the reward and cost are bounded by $[0, 1]$, we have $\{\mu_i, c_i, \sigma_i^\mu, \sigma_i^c\} \in (0, 1)$ for all $1 \le i \le K$.

The optimal arm is defined as the one with the largest ratio of the expected reward to the expected cost. For simplicity, we assume that there is only one optimal arm. Without loss of generality, we assume that $\frac{\mu_1}{c_1} > \frac{\mu_2}{c_2} > ... > \frac{\mu_K}{c_K}$. Note that such order information is assumed for the convenience of theoretical analysis. The player in real world has no such information *a priori*. In addition, our algorithms still work without such assumptions.

Next, we define the sample complexity and present our problem definition accordingly.

Definition 1: Sample Complexity: *For a best arm identification problem in MAB model, the sample complexity is defined as the pulling times of all arms when the best arm is selected.*

Problem definition: *In a best arm with reward to cost ratio identification problem, given a set of K arms and a confidence level δ, we aim to find the arm with the largest reward to cost ratio with probability at least $1 - \delta$ and minimum sample complexity.*

IV. Algorithm Design

In this section, we will first present a genie-aided algorithm with known gaps information, where the gaps denotes the deviations between suboptimal arms to the optimal arm. Finally, based on the insights from the algorithm, two successive elimination algorithms with unknown gaps are designed.

A. Preliminaries

Before we present these algorithms, we introduce some useful definitions and notation. First, for any arm $i \in \mathcal{S}$, t is the pulling rounds, let $\hat{\mu}_{i,t}, \hat{c}_{i,t}$ be the empirical means of reward and cost, respectively. i.e.,

$$\hat{\mu}_{i,t} = \frac{1}{t}\sum_{s=1}^{t} X_{i,s}^\mu, \text{and } \hat{c}_{i,t} = \frac{1}{t}\sum_{s=1}^{t} X_{i,s}^c. \tag{1}$$

Next, following [14], we introduce two types of gaps between the optimal arm and a suboptimal arm:

$$\Delta_i = \frac{\mu_1}{c_1} - \frac{\mu_i}{c_i}, \text{and } \xi_i = \frac{c_1 c_i \Delta_i}{\mu_1 + c_1 + \mu_i + c_i}. \tag{2}$$

Clearly, Δ_i denotes the gap between the optimal arm with arm i. ξ_i also characterizes the differences between the best arm and other arms. The meaning of ξ_i is that increasing c_1, μ_i with ξ_i, and decreasing μ_1 and c_i with ξ_i, we can make the ratio of the best arm equals to that of arm i, i.e.,

$$\frac{\mu_1 - \xi_i}{c_1 + \xi_i} = \frac{\mu_i + \xi_i}{c_i - \xi_i}. \tag{3}$$

Note that for $i \ge 2$, we can get $\mu_1 > \xi_i$ and $c_i > \xi_i$. In addition, we define that $\xi_1 = \min_{i \ge 2} \xi_i$ and $H_1 = \sum_{i=1}^{K} \xi_i^{-2}$.

Finally, following the Theorem 11 in [21], we can show the following useful Bernstein bound:

Bernstein bound: For an random variable X with t samples, let $\hat{X}_t$ be the average over these t sample values and σ_X is the variance of X. Then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$ we have

$$\mathbb{E}[X] \le \hat{X}_t + \sqrt{\frac{2\sigma_X \log\frac{2}{\delta}}{t}} + \frac{7\log\frac{2}{\delta}}{3(t-1)}. \tag{4}$$

B. Algorithm Design with Known Gaps

In this subsection, we focus on a genie-aided scenario that the expectations of reward and cost are known, while their matching to arms are still unknown. Although such a scenario maybe not realistic in real world applications, it provide the insights for the subsequent algorithm design with unknown gaps. Specifically, consider a genie-aided algorithm based on the hidden information $\{\xi_i\}_{i=1}^K$.

Algorithm 1 Genie-Aided Algorithm

Input: The K arms set $\mathcal{S}$, and a confidence level δ.
Output: The optimal arm i^*.
1: **for** $i = 1$ to K **do**
2: Pull arm i for $t_i = \lceil \frac{1}{2\xi_i^2} \log \frac{2K}{\delta} \rceil$ times.
3: Calculate $\hat{\mu}_i$ and $\hat{c}_i$ based on Eq. (1).
4: **end for**
5: **return** $i^* = \arg\max_{i \in \mathcal{S}} \frac{\hat{\mu}_i}{\hat{c}_i}$.

In the genie-aided algorithm, by using the Hoeffding bound and the gaps knowledge, one can compute the sampling times for each arm and accurately estimate the expectations of both reward and cost within the confidence level. Therefore, the optimal arm can be selected by comparing the empirical means of reward and cost for each arm. For simplicity, we use GA to denote **G**enie-**A**ided algorithm in the rest of this paper. The following theorem shows the sample complexity of the GA algorithm.

Theorem 1: *To return an optimal arm with probability at least $1 - \delta$, the sample complexity for GA algorithm is $O(H_1 \log \frac{2K}{\delta})$, where $H_1 = \sum_{i=1}^{K} \xi_i^{-2}$.*

Proof: If the output of naive selection algorithm is a suboptimal arm, we can know that the following event $\mathcal{E}_{\text{GA}}$ happens,

$$\mathcal{E}_{\text{GA}} = \bigcup_{i=2}^{K} \Big(\frac{\hat{\mu}_i}{\hat{c}_i} \ge \frac{\hat{\mu}_1}{\hat{c}_1} \Big), \tag{5}$$

which means at least one of the following two events is true,

$$\frac{\hat{\mu}_i}{\hat{c}_i} \ge \frac{\mu_1 - \xi_i}{c_1 + \xi_i} \text{ or } \frac{\hat{\mu}_1}{\hat{c}_1} \le \frac{\mu_i + \xi_i}{c_i - \xi_i}. \tag{6}$$

Intersecting the event $\mathcal{E}_{\text{GA}}$ and the events in Eq. (6), we can

obtain that

$$\begin{aligned}\mathcal{E}_{\text{GA}} &= \bigcup_{i=2}^{K}\Big\{\frac{\hat{\mu}_i}{\hat{c}_i} \geq \frac{\hat{\mu}_1}{\hat{c}_1}\Big\} \\ &\subseteq \bigcup_{i=2}^{K}\Big\{\Big(\frac{\hat{\mu}_1}{\hat{c}_1} \leq \frac{\mu_1-\xi_i}{c_1+\xi_i}\Big)\bigcup\Big(\frac{\hat{\mu}_i}{\hat{c}_i} \geq \frac{\mu_i+\xi_i}{c_i-\xi_i}\Big)\Big\}.\end{aligned} \tag{7}$$

Decomposing Eq. (7), we obtain:

$$\begin{aligned}\mathcal{E}_{\text{GA}} &\subseteq \bigcup_{i=2}^{K}\big\{(\hat{\mu}_1 \leq \mu_1-\xi_i)\cup(\hat{c}_1 \geq c_1+\xi_i) \\ &\quad \cup(\mu_i \geq \mu_i+\xi_i)\cup(\hat{c}_i \leq c_i-\xi_i)\big\} \\ &= \bigcup_{i=2}^{K}\big\{(\hat{\mu}_i \geq \mu_i+\xi_i)\cup(\hat{c}_i \leq c_i-\xi_i)\big\} \\ &\quad \cup(\hat{\mu}_1 \leq \mu_1-\xi_i)\cup(\hat{c}_1 \geq c_1+\xi_i).\end{aligned} \tag{8}$$

In the GA algorithm, each arm i is pulled $t_i = \lceil \frac{2}{\xi_i^2}\log\frac{2K}{\delta}\rceil$ times, which is deterministic. Thus, according to the Hoeffding inequality, we have that

$$\begin{aligned}\mathbb{P}(\mathcal{E}_{\text{GA}}) &\leq \bigcup_{i=2}^{K}\big\{\mathbb{P}(\mu_i \geq \mu_i+\xi_i)+\mathbb{P}(\hat{c}_i \leq c_i-\xi_i)\big\} \\ &\quad +\mathbb{P}(\hat{\mu}_1 \leq \mu_1-\xi_i)+\mathbb{P}(\hat{c}_1 \geq c_1+\xi_i) \\ &\leq 2\sum_{k=1}^{K}\exp(-2\xi_i^2 t_i) = \delta,\end{aligned} \tag{9}$$

which means that under the GA algorithm, the optimal arm will be selected with probability at least $1-\delta$. The proof is complete. ∎

Remark 1. From the Theorem 1, we can see that the sample complexity of the GA algorithm is on the order of $O(\log\frac{1}{\delta})$, which grows logarithmically with respect to $\frac{1}{\delta}$. In addition, the sample complexity of the GA algorithm depends the parameters ξ_i $(1 \leq i \leq K)$. As we can see later, ξ_i also plays an important role in the proofs of other algorithms. In fact, the quantity $H_1 = \sum_{i=1}^{K}\xi_i^{-2}$ characterizes of the hardness of the problem. We define H_1 as the complexity hardness of the GA algorithm.

The limitation of the GA algorithm is that a player needs to know all gaps in advance, which is not realistic in practice. To overcome this limitation, we will develop two algorithms without knowing $\{\xi_i\}_{i=1}^{K}$ in the following subsection.

C. Algorithm Design with Unknown Gaps

In this subsection, we propose the successive elimination algorithm without prior gaps information. The basic idea of elimination-type algorithms (e.g. [1], [12]) is to sequentially eliminate suboptimal arms. We adopt arm-elimination into our algorithm design to overcome the limitations of the GA algorithm. The successive elimination algorithm are conducted as follows. In each round, all the arms are uniformly pulled one time. Then, these arms whose ratios are smaller than a certain threshold from the optimal arm will be eliminated. Upon the completion of the algorithm, the final surviving arm is the presented optimal one. Note that in our algorithm, the threshold depends on the round number as well as the δ, and is independent of the gaps parameter $\{\xi_i\}_{i=1}^{K}$.

We will firstly present a β_t-modulated unified algorithmic framework in Algorithm 2, where β_t is the threshold related parameters. Based on this framework, we develop two elimination-type algorithms with their specified β_t.

Algorithm 2 β_t-Modulated Successive Elimination Algorithm Framework

Input: The arm set $\mathcal{S}$ and a confidence level δ.
Output: $\mathcal{S}$.

1: Set $t=1$.
2: **while** $|\mathcal{S}|>1$ **do**
3: Pull each arm in $\mathcal{S}$ once. For any arm $i\in\mathcal{S}$, update the $\hat{\mu}_{i,t}$ and $\hat{c}_{i,t}$.
4: Let $\frac{\mu_t^*}{c_t^*} \leftarrow \max_i \frac{\hat{\mu}_{i,t}}{\hat{c}_{i,t}}$ and $i^* \leftarrow \arg\max_i \frac{\hat{\mu}_{i,t}}{\hat{c}_{i,t}}$.
5: For each algorithm, compute the specified β_t.
6: **if** $\mu_t^* - 2\beta_t \leq 0$ **then**
7: Go to Line 16
8: **end if**
9: **for** each arm i in $\mathcal{S}\setminus\{i^*\}$ **do**
10: **if** $\hat{c}_{i,t} - 2\beta_t \leq 0$ **then**
11: Back to Line 9
12: **else if** $\frac{\hat{\mu}_t^*-2\beta_t}{\hat{c}_t^*+2\beta_t} > \frac{\hat{\mu}_{i,t}+2\beta_t}{\hat{c}_{i,t}-2\beta_t}$ **then**
13: $\mathcal{S}\leftarrow\mathcal{S}\setminus\{i\}$.
14: **end if**
15: **end for**
16: $t\leftarrow t+1$
17: **end while**
18: **return** $\mathcal{S}$

In Algorithm 2 Line 4, at each round, if the number of optimal arm is larger than 1, we randomly pick one of them. Moreover, in Line 5 of Algorithm 2, the expression $\mu_t^* - 2\beta_t$ may be smaller than zero. This means that zero is within the lower confidence bound of best arm reward with high probability. Thus, we should continue to pull all arms and obtain more accurate estimations of rewards. The same explanations are also applicable to the Line 10 in Algorithm 2.

The first algorithm is named as SEUG (**S**uccessive **E**limination with **U**nknown **G**aps) algorithm. The specified β_t of SEUG is defined as $\beta_t = \sqrt{\frac{\log\frac{4K}{\delta}}{2t}}$ in Line 5 of Algorithm 2. In SEUG, as we will prove later, the β_t and the discarding criteria are carefully designed to make sure that the best arm can be selected with probability at least $1-\delta$. We present the sample complexity of SEUG algorithm in the following theorem.

Theorem 2: *To return an optimal arm with probability at least* $1-\delta$*, the sample complexity for SEUG algorithm is* $O(H_1\log\frac{4K}{\delta})$*, where* $H_1 = \sum_{i=1}^{K}\xi_i^{-2}$.

Proof: Our proof is organized in three steps. First, we prove that under a certain event, the optimal arm can be finally selected with probability at least $1-\delta$. Second, we show an

upper bound of pulling times for all the suboptimal arms. In the last step, based on the results from the previous two steps, we calculate the sample complexity.

Step 1: We define an event $\mathcal{E}_1$ as:

$$\mathcal{E}_1 \triangleq \{| \hat{\mu}_{i,t} - \mu_i | \leq \beta_t, | \hat{c}_{i,t} - c_i | \leq \beta_t, \forall i \in \mathcal{S}, \forall t \in \mathcal{T}\},$$

where the set $\mathcal{T} \triangleq \{1, 2, ..., T\}$.

According to the Hoeffding inequality and the union bound, the probability for event $\mathcal{E}_1$ can be lower bounded by:

$$\mathbb{P}(\mathcal{E}_1) \geq 1 - 4TK \exp(-2t\beta_t^2) \geq 1 - \delta. \tag{10}$$

Next, we prove that under the event $\mathcal{E}_1$, the optimal arm with the best ratio of reward to cost can be identified. We assume i is an arbitrary suboptimal arm and is deleted in round t_i. It is easy to see that for any suboptimal arm i, under the event $\mathcal{E}_1$ and in any round t, we have that

$$\frac{\hat{\mu}_{1,t} + \beta_t}{\hat{c}_{1,t} - \beta_t} > \frac{\mu_1}{c_1} > \frac{\mu_i}{c_i} > \frac{\hat{\mu}_{i,t} - \beta_t}{\hat{c}_{i,t} + \beta_t}. \tag{11}$$

The inequality in (11) means that with probability at least $1 - \delta$, the optimal arm with the highest reward to cost ratio will never be eliminated. Moreover, since β_t decreases as t increases, all suboptimal arms will be eliminated when t is sufficiently large.

Step 2: Next, we establish an upper bound of the pulling times for all suboptimal arms. Assume that a suboptimal arm i is deleted in round t_i. Under the event $\mathcal{E}_1$, we will show that

$$t_i \geq \frac{2 \log \frac{4K}{\delta}}{\xi_i^2}. \tag{12}$$

We prove Eq. (12) by contradiction. Assume that

$$t_i < \frac{2 \log \frac{4K}{\delta}}{\xi_i^2}, \tag{13}$$

which means that $\xi_i < 2\beta_{t_i}$. According to Algorithm 2, if arm i is deleted from the set $\mathcal{S}$ at t_i, the condition $\frac{\hat{\mu}_1 - \beta_{t_i}}{\hat{c}_1 + \beta_{t_i}} \geq \frac{\hat{\mu}_{i,t_i} + \beta_{t_i}}{\hat{c}_{i,t_i} - \beta_{t_i}}$ is satisfied. Combined with the event $\mathcal{E}_1$, we can conclude that

$$\frac{\mu_1 - 2\beta_{t_i}}{c_1 + 2\beta_{t_i}} \geq \frac{\hat{\mu}_1 - \beta_{t_i}}{\hat{c}_1 + \beta_{t_i}} \geq \frac{\hat{\mu}_{i,t_i} + \beta_{t_i}}{\hat{c}_{i,t_i} - \beta_{t_i}} \geq \frac{\mu_i + 2\beta_{t_i}}{c_i - 2\beta_{t_i}}. \tag{14}$$

Further, combined with Eq. (2), we can obtain that

$$\frac{\mu_1 - 2\beta_{t_i}}{c_1 + 2\beta_{t_i}} \geq \frac{\mu_1 - \xi_i}{c_1 + \xi_i} = \frac{\mu_i + \xi_i}{c_i - \xi_i} \geq \frac{\mu_i + 2\beta_{t_i}}{c_i - 2\beta_{t_i}}. \tag{15}$$

From both the sides of the equality in Eq. (15), we can derive that

$$\xi_i \geq 2\beta_{t_i} \Rightarrow t_i \geq \frac{2 \log \frac{4K}{\delta}}{\xi_i^2}, \tag{16}$$

which contradicts Eq. (13). Thus, Eq. (12) is proved.

Step 3: Let τ_1 denote the sample complexity of SEUG algorithm. Thus, $\tau_1 = \sum_{i=1}^{K} t_i = O(H_1 \log \frac{4K}{\delta})$, where $H_1 = \sum_{i=1}^{K} \xi_i^{-2}$. The proof is complete. ■

Remark 2. From Theorem 2, we can see that the sample complexity of SEUG also grows logarithmatically with $\frac{1}{\delta}$, which is of the same order as that of the GA algorithm. In addition, we can see that both of their sample complexities depend on H_1. However, in practice, the SEUG algorithm needs much more samples to identify the best arm. This is because that the SEUG algorithm has no prior knowledge about H_1. Hence, to better estimate the differences between arms and eliminate uncertainty, the player will sample more times on each arm to narrow the empirical means of both reward and cost, which results in a larger sample complexity. This phenomenon will also be observed by the numerical simulations.

As stated in the Remark 2, the uncertainty of gaps in SEUG algorithm results in much higher sample complexity. To overcome this limitation, we consider to utilize the variance information of both reward and cost to achieve lower complexity. Intuitively, a suboptimal arm with small variances of reward or cost needs a smaller number of rounds to be deleted. To this end, we propose a variant of SEUG algorithm, which incorporates the variances into the β_t-index and hence is called SEUG-V. Specifically, for arm i, we define $\hat{\sigma}_{i,t}^{\mu} = \frac{1}{t-1} \sum_{s=1}^{t} (\mu_{i,s} - \hat{\mu}_i)^2$, $\hat{\sigma}_{i,t}^{c} = \frac{1}{t-1} \sum_{s=1}^{t} (c_{i,s} - \hat{c}_i)^2$ to be empirical variances for reward and cost, respectively. The specified β_t for arm i of SEUG-V in Algorithm 2 is defined as

$$\beta_{i,t}^{\mu} = \sqrt{\frac{\log \frac{2K}{\delta} \hat{\sigma}_{i,t}^{\mu}}{t}} + \frac{7 \log \frac{2K}{\delta}}{3(t-1)}, \tag{17}$$

$$\beta_{i,t}^{c} = \sqrt{\frac{\log \frac{2K}{\delta} \hat{\sigma}_{i,t}^{c}}{t}} + \frac{7 \log \frac{2K}{\delta}}{3(t-1)}. \tag{18}$$

Based on the above definitions, in Algorithm 2, by substituting the expression $(\hat{\mu}_t^* - 2\beta_t)$ with $(\hat{\mu}_t^* - 2\beta_{i^*,t}^{\mu})$ in Line 6 and 12, the expression $(\hat{c}_t^* + 2\beta_t)$ with $(\hat{c}_t^* + 2\beta_{i^*,t}^{c})$ in Line 12, the expression $(\hat{\mu}_{i,t} + 2\beta_t)$ with $(\hat{\mu}_{i,t} + 2\beta_{i,t}^{\mu})$ in Line 12, the expression $(\hat{c}_{i,t} - 2\beta_t)$ with $(\hat{c}_{i,t} - 2\beta_{i,t}^{c})$ in Line 12, we can obtain the SEUG-V algorithm. The sample complexity is presented in the following Theorem 3.

Theorem 3: *To return an optimal arm with probability at least $1 - \delta$, the sample complexity for SEUG-V algorithm is $O(H_2 \log \frac{2K}{\delta})$, where $H_2 = \sum_{i=1}^{K} \frac{\log \frac{2K}{\delta} (\sigma_i^{max} + \sqrt{(\sigma_i^{max})^2 + \frac{7\xi_i}{3}})^2}{\xi_i^2}$.*

Proof: The proof process is similar to the proof of SEUG algorithm. We divide this process into three steps and omit some details for brevity.

Step 1: We define an event $\mathcal{E}_2$ as:

$$\mathcal{E}_2 = \{| \hat{\mu}_{i,t} - \mu_i | \leq \beta_{i,t}^{\mu}, | \hat{c}_{i,t} - c_i | \leq \beta_{i,t}^{c}, \forall i \in \mathcal{S}, \forall t \in \mathcal{T}\}.$$

According to the empirical Bernstein bound (Theorem 11, [21]) and the union bound, $\mathcal{E}_2$ holds with probability at least $1 - \delta$. Thus, similar to the proof of SEUG algorithm, step 1, under event $\mathcal{E}_2$, we have

$$\frac{\hat{\mu}_{1,t} + \beta_{1,t}^{\mu}}{\hat{c}_{1,t} - \beta_{1,t}^{c}} > \frac{\mu_1}{c_1} > \frac{\mu_i}{c_i} > \frac{\hat{\mu}_{i,t} - \beta_{i,t}^{\mu}}{\hat{c}_{i,t} + \beta_{i,t}^{c}}. \tag{19}$$

which means that under event $\mathcal{E}_2$, the best arm will never be deleted with probability at least $1 - \delta$. Moreover, with the growth of rounds, all suboptimal arms will be deleted.

Step 2: Under event $\mathcal{E}_2$, we present an upper bound for the pulling times of each arm $i \in \mathcal{S}$. By contradiction, to delete

a suboptimal arm i, the following condition has been to be satisfied:

$$\xi_i \geq 2\max\{\beta^{\mu}_{1,t_i}, \beta^{c}_{1,t_i}, \beta^{\mu}_{i,t_i}, \beta^{c}_{i,t_i}\}, 2 \leq i \leq K. \quad (20)$$

Similar to the step 2 in SEUG algorithm, assume that arm i is deleted in round t_i^v. Combined with Eqs. (17) and (18), we have

$$t_i^v \geq \frac{\log \frac{2K}{\delta}(\sigma_i^{\max} + \sqrt{(\sigma_i^{\max})^2 + \frac{7\xi_i}{3}})^2}{\xi_i^2}, 2 \leq i \leq K, \quad (21)$$

where $\sigma_i^{\max} = \max\{\sigma_1^{\mu}, \sigma_1^{c}, \sigma_i^{\mu}, \sigma_i^{c}\}$. Let τ_v to be the sample complexity of SEUG-V algorithm. we have,

$$\tau_v = t_1^v + \sum_{i=2}^{K} t_i^v = O(H_2 \log \frac{2K}{\delta}), \quad (22)$$

where $H_2 = \sum_{i=1}^{K} \frac{\log \frac{2K}{\delta}(\sigma_i^{\max} + \sqrt{(\sigma_i^{\max})^2 + \frac{7\xi_i}{3}})^2}{\xi_i^2}$. The proof is complete. ∎

Remark 3. As we can see in Theorem 3, the SEUG-V algorithm introduces an new notion of hardness H_2, which is related to both $\{\xi_i\}_{i=1}^K$ and variance information. It is easy to check that $H_1 \geq H_2$ if both rewards and costs are in the range of $[0, 1]$. Therefore, compared to the SEUG algorithm, the sample complexity will be significantly reduced in SEUG-V. However, since it still cannot accurately estimate the gaps, its sample complexity is larger than that of GA algorithm. Such results will also be validated by the numerical simulations.

V. Lower Bound of Sample Complexity

In this section, we present a fundamental lower bound for the sample complexity. We assume the both the rewards and costs of all the arms follow the Bernoulli distribution, which has been widely used in the lower bound analysis in multi-armed bandit problems. Similar to [14], we define a series of notation as follows.

First, we introduce the Kullback-Leibler divergence (KL-divergence) under the Bernoulli distributions. For two Bernoulli distributions with parameters p, q, their KL-divergence is defined as

$$\mathrm{D}(p\|q) = p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q}, p, q \in (0,1). \quad (23)$$

Second, for any $i \geq 2$ and $\gamma > 0$ we define a constant $\epsilon_i(\gamma)$ as follows

$$\epsilon_i(\gamma) = \frac{c_1 c_i \Delta_i}{\gamma\mu_1 + c_1}. \quad (24)$$

Intuitively, $\epsilon_i(\gamma)$ is an adaptive version of Δ_i: By increasing μ_i by $\epsilon_i(\gamma)$ and decreasing c_i by $\gamma\epsilon_i(\gamma)$, one can make arm i to be the best arm, i.e., $\frac{\mu_1}{c_1} = \frac{\mu_i + \epsilon_i(\gamma)}{c_i - \gamma\epsilon_i(\gamma)}$.

In MAB literature (e.g., [5], [23]), the key element to establish a distribution-dependent lower bound lies in the changes of distributions. Specifically, it is related to the probabilities of the same event under two different bandit models. Thus, we use ν and v to denote two bandit models respectively, which are both a product of Bernoulli distributions. Precisely,

$$\begin{aligned} \nu &= \otimes_{i=1}^{K}\mathrm{Ber}(\nu_i^{\mu}) \times \otimes_{i=1}^{K}\mathrm{Ber}(\nu_i^{c}), \\ v &= \otimes_{i=1}^{K}\mathrm{Ber}(v_i^{\mu}) \times \otimes_{i=1}^{K}\mathrm{Ber}(v_i^{c}). \end{aligned} \quad (25)$$

In Eq. (25), $\mathrm{Ber}(p)$ denotes the Bernoulli distribution with parameter p. We use $\otimes$ and $\times$ to represent the products of distributions. The term $\otimes_{i=1}^{K}\mathrm{Ber}(\nu_i^{\mu})$ denotes the products of K independent Bernoulli distribution, and the i-th one has parameter ν_i^{μ}. Moreover, $\{\nu_i^{\mu}, \nu_i^{c}, v_i^{\mu}, v_i^{c}\} \in (0,1)$ for $1 \leq i \leq K$.

In the best arm identification problem in our setting, the player needs to design a mechanism to select the arm with optimal reward to cost ratio. Similar to [23], we define a algorithm $\mathcal{A}$ with a triplet $\mathcal{A} = \{a_t, \tau, a_\tau^*\}$, which is stated as follows:

1) *The sampling rule.* Based on the historical observations, a sampling rule decides those arms that are chosen to be sampled at round t. Thus, the arm set $\mathcal{S}_t$ is $\mathcal{F}_t$-measurable, where $\mathcal{F}_t = \sigma\{a_1, X^{\mu}_{a_1}, X^{c}_{a_1}, ..., a_t, X^{\mu}_{a_t}, X^{c}_{a_t}\}$ is a σ-algebra of historical observations and selected arms.

2) *The stopping rule.* A stopping rule determines the end of the sampling process and is represented by a stopping time τ. τ is associated with $\mathcal{F}_t$ and $\mathbb{P}(\tau < +\infty) = 1$.

3) *The recommendation rule.* A recommendation rule selects the optimal based on the observation of $\mathcal{F}_\tau$. The final selected arm at τ is denoted by a_τ^*.

A. A Fundamental Lower Bound

Based on the above definitions, we assume $T_i(t) = \sum_{s=1}^{t}\mathbf{1}(i \in a_s)$ is the number of sampling times for arm i up to round t. Specifically, for an algorithm $\mathcal{A}$ with the stopping time τ, $T_i(\tau)$ is the number of total samples for arm i when $\mathcal{A}$ is stopped. For two bandit models ν, v and the history observations, we can obtain the following lemma.

Lemma 1: *For two bandit models ν, v with K arms, assume τ is the stopping time based on the historical observation $\mathcal{F}_t$. For each event $\mathcal{E} \in \mathcal{F}_\tau$,*

$$\sum_{i=1}^{K}\mathbb{E}[T_i(\tau)](\mathrm{D}(\nu_i^{\mu}\|v_i^{\mu}) + \mathrm{D}(\nu_i^{c}\|v_i^{c})) \geq \mathrm{D}(\mathbb{P}_\nu(\mathcal{E})\|\mathbb{P}_v(\mathcal{E})). \quad (26)$$

Proof: For an algorithm $\mathcal{A}$ with stopping time τ and the corresponding observations and historical selections $\mathcal{F}_\tau$, we introduce the log-likelihood ratio L_τ of $\mathcal{F}_\tau$. According to Theorem 8, [9], we have

$$\begin{aligned} L_\tau = &\sum_{i=1}^{K}\sum_{s=1}^{T_i(\tau)}\log\frac{\nu_i^{\mu}X_{i,s}^{\mu} + (1-\nu_i^{\mu})(1-X_{i,s}^{\mu})}{v_i^{\mu}X_{i,s}^{\mu} + (1-v_i^{\mu})(1-X_{i,s}^{\mu})} \\ &+ \sum_{i=1}^{K}\sum_{s=1}^{T_i(\tau)}\log\frac{\nu_i^{c}X_{i,s}^{c} + (1-\nu_i^{c})(1-X_{i,s}^{c})}{v_i^{c}X_{i,s}^{c} + (1-v_i^{c})(1-X_{i,s}^{c})}, \end{aligned} \quad (27)$$

where $X_{i,s}^{\mu}$ and $X_{i,s}^{c}$ are the reward and cost sampled in round s, respectively. According to the Lemmas 18, 19 in [23], for any event $\mathcal{E} \in \mathcal{F}_\tau$, one have that

$$\begin{aligned} \mathbb{P}_v(\mathcal{E}) &= \mathbb{E}_\nu[\exp(-L_\tau)\mathbf{1}(\mathcal{E})] \\ &= \mathbb{E}_\nu[\exp(-L_\tau)\mathbf{1}(\mathcal{E})|\mathbf{1}(\mathcal{E}) = 1]\mathbb{P}_\nu(\mathcal{E}) \\ &= \mathbb{E}_\nu[\exp(-L_\tau)|\mathcal{E}]\mathbb{P}_\nu(\mathcal{E}) \\ &\geq \exp(\mathbb{E}_\nu[-L_\tau|\mathcal{E}])\mathbb{P}_\nu(\mathcal{E}). \end{aligned} \quad (28)$$

Thus, we can obtain that:

$$\exp(\mathbb{E}_\nu[-L_\tau|\mathcal{E}]) \geq \log \frac{\mathbb{P}_\nu(\mathcal{E})}{\mathbb{P}_v(\mathcal{E})}. \tag{29}$$

For any event $\mathcal{E} \in \mathcal{F}_\tau$, $\overline{\mathcal{E}} \in \mathcal{F}_\tau$. Thus, the following equality also holds

$$\exp(\mathbb{E}_\nu[-L_\tau|\overline{\mathcal{E}}]) \geq \log \frac{\mathbb{P}_\nu(\overline{\mathcal{E}})}{\mathbb{P}_v(\overline{\mathcal{E}})}. \tag{30}$$

Combining Eqs. (29) and (30), we have,

$$\begin{aligned}\mathbb{E}_\nu[L_\tau] &= \mathbb{E}[L_\tau|\mathcal{E}]\mathbb{P}_\nu(\mathcal{E}) + \mathbb{E}_\nu[L_\tau|\overline{\mathcal{E}}]\mathbb{P}_\nu(\overline{\mathcal{E}}) \\ &\geq \mathbb{P}_\nu(\mathcal{E}) \log \frac{\mathbb{P}_\nu(\mathcal{E})}{\mathbb{P}_v(\mathcal{E})} + \mathbb{P}_\nu(\overline{\mathcal{E}}) \log \frac{\mathbb{P}_\nu(\overline{\mathcal{E}})}{\mathbb{P}_v(\overline{\mathcal{E}})}, \\ &= \mathrm{D}(\mathbb{P}_\nu(\mathcal{E})\|\mathbb{P}_v(\mathcal{E})).\end{aligned} \tag{31}$$

Moreover, according to Eq. (27) and Lemma 19 in [23],

$$\begin{aligned}\mathbb{E}_\nu[L_\tau] &= \mathbb{E}_\nu\Big[\sum_{i=1}^{K}\sum_{s=1}^{T_i(\tau)} \log \frac{\nu_i^\mu X_{i,s}^\mu + (1-\nu_i^\mu)(1-X_{i,s}^\mu)}{v_i^\mu X_{i,s}^\mu + (1-v_i^\mu)(1-X_{i,s}^\mu)}\Big] \\ &+ \mathbb{E}_\nu\Big[\sum_{i=1}^{K}\sum_{s=1}^{T_i(\tau)} \log \frac{\nu_i^c X_{i,s}^c + (1-\nu_i^c)(1-X_{i,s}^c)}{v_i^c X_{i,s}^c + (1-v_i^c)(1-X_{i,s}^c)}\Big] \\ &= \sum_{i=1}^{K} \mathbb{E}[T_i(\tau)](\mathrm{D}(\nu_i^\mu\|v_i^\mu) + \mathrm{D}(\nu_i^\mu\|v_i^\mu)).\end{aligned} \tag{32}$$

Combing Eqs. (31) and Eq. (32), we can obtain the stated results in Lemma 1. The proof is complete. ∎

Next, we state the lower bound of sample complexity in the following theorem.

Theorem 4: *Let* $\Gamma_j = \{\gamma_j | j \in \{2,...,K\}, \mu_j + \epsilon_j(\gamma) < 1, c_j - \gamma\epsilon_j(\gamma) > 0, \gamma_j \geq 0\}$. *For the best arm identification problem with Bernoulli distributions, the sample complexity* τ *of any algorithm* $\mathcal{A}$ *under fixed confidence level* δ *satisfies*

$$\mathbb{E}[\tau] \geq \sum_{j=1}^{K} \frac{\mathrm{D}(1-\delta\|\delta)}{\mathrm{D}_j^*}, \tag{33}$$

where $\mathrm{D}_j^* = \sup_{\gamma_j \in \Gamma_j}\{\mathrm{D}(\mu_j\|\mu_j+\epsilon_j(\gamma)) + \mathrm{D}(c_j\|c_j-\gamma\epsilon_j(\gamma))\}$ *and* $\mathrm{D}_1^* = \min_{j=\{2,...,K\}} \mathrm{D}_j^*$

Proof: To derive the lower bound, we specialize the bandit model ν, v, and $\mathcal{E}$. For an algorithm $\mathcal{A}$ with stopping time τ and the corresponding $\mathcal{F}_\tau$, we have

$$\begin{aligned}&\nu = \otimes_{i=1}^{K}\mathrm{Ber}(\mu_i) \times \otimes_{i=1}^{K}\mathrm{Ber}(c_i); \\ &v = v^\mu \otimes v^c, \text{where} \\ &v^\mu = \otimes_{i=1}^{j-1}\mathrm{Ber}(\mu_i) \otimes \mathrm{Ber}(\mu_j + \epsilon_j(\gamma) + \alpha) \otimes_{i=j+1}^{K} \mathrm{Ber}(\mu_i); \\ &v^\mu = \otimes_{i=1}^{j-1}\mathrm{Ber}(c_i) \otimes \mathrm{Ber}(c_j - \gamma\epsilon_j(\gamma) - \alpha) \otimes_{i=j+1}^{K} \mathrm{Ber}(c_i),\end{aligned} \tag{34}$$

where the γ and α are constrained by $\mu_j + \epsilon_j(\gamma) + \alpha \in (0,1)$, $c_j - \gamma\epsilon_j(\gamma) - \alpha \in (0,1)$ and $j \geq 2$.

Based on above definitions, we can see that the optimal arm in bandit model v is arm j instead of arm 1. Hence, for any algorithms under fixed confidence δ, assume event $\mathcal{E} \triangleq \{1 = \arg\max_i \frac{\hat{\mu}_{i,\tau}}{\hat{c}_{i,\tau}}\} \in \mathcal{F}_\tau$. Then $\mathbb{P}_\nu(\mathcal{E}) \geq 1-\delta$, and $\mathbb{P}_v(\mathcal{E}) \leq \delta$. Plugging notations into (26), we have that

$$\begin{aligned}\mathbb{E}[T_j(\tau)] &\geq \frac{\mathrm{D}(\mathbb{P}_\nu(\mathcal{E})\|\mathbb{P}_v(\mathcal{E})}{\mathrm{D}(\nu_j^\mu\|v_j^\mu) + \mathrm{D}(\nu_j^c\|v_j^c)} \\ &\geq \frac{\mathrm{D}(1-\delta\|\delta)}{\mathrm{D}(\mu_j\|\mu_j+\epsilon_j(\gamma)+\alpha)+\mathrm{D}(c_j\|c_j-\gamma\epsilon_j(\gamma)-\alpha)}.\end{aligned} \tag{35}$$

For $j \in \{2,...,K\}$, let $\Gamma_j = \{\gamma_j | j \in \{2,...,K\}, \mu_j + \epsilon_j(\gamma) < 1, c_j - \gamma\epsilon_j(\gamma) > 0, \gamma_j \geq 0\}$. Let $\mathrm{D}_j^* = \sup_{\gamma_j \in \Gamma_j}\{\mathrm{D}(\mu_j\|\mu_j + \epsilon_j(\gamma)) + \mathrm{D}(c_j\|c_j - \gamma\epsilon_j(\gamma))\}$. Taking $\alpha \to 0$, we have

$$\mathbb{E}[T_j(\tau)] \geq \frac{\mathrm{D}(1-\delta\|\delta)}{\mathrm{D}_j^*}. \tag{36}$$

For $j = 1$, let $D_1^* = \min_{j\in\{2,...,K\}} D_j^*$. Summing (36) over all $j \in \{1,...,K\}$, we obtain a lower bound of stopping time τ:

$$\mathbb{E}[\tau] = \sum_{j=1}^{K} \mathbb{E}[T_j(\tau)] \geq \sum_{j=1}^{K} \frac{\mathrm{D}(1-\delta\|\delta)}{\mathrm{D}_j^*}. \tag{37}$$

The proof is complete. ∎

Remark 4. In essence, the lower bound in Theorem 4 matches the sample complexity for the three proposed algorithms on the order of $\log\frac{1}{\delta}$. According to [23], in (33) of Theorem 4, for any $\delta \in (0,1)$, the expression $\mathrm{D}(1-\delta\|\delta)) \geq \log\frac{1}{2.4\delta}$. Moreover, the denominator $\sum_{j=1}^{K} 1/D_j^*$ depends on the specific distribution parameters and is bounded. Thus, the lower bound is on the order of $\Omega(\log\frac{1}{\delta})$, which matches the upper bounds from the proposed algorithms. This shows that both of our algorithms are order-optimal with respect to δ.

B. Discussions

In this subsection, we discuss the asymptotic property of the lower bound when the number of arms K grows. In Theorem 4, when δ is fixed, the numerator of the lower bound is a constant, while the denominator is a linear function of K. Thus, such lower bound grows linearly with K. This result, however, is not consistent with the results of the three proposed algorithms, as their sample complexities are on the order of $O(K\log K)$. This loose lower bound is caused by the imprecise estimation of the probability $\mathbb{P}_v(\mathcal{E})$ in the inequality in (35). In our bandit model v, arm j is the optimal arm, and arm 1 is one of $K-1$ suboptimal arms. Under any algorithms with a fixed confidence δ, arm j will be finally selected with probability at least $1-\delta$. Moreover, since arm 1 is one of the $K-1$ suboptimal arms, the upper bound of $\mathbb{P}_v(\mathcal{E})$, i.e., the probability for arm 1 to be finally selected, is smaller than δ. Such an overestimation leads to the mismatch of the lower bound and the upper bounds of three algorithms.

Next, we show that the asymptotic performance with regard to K depends on the specific bandit models. We apply two special examples to analyze the impacts of different bandit models on the asymptotic performance of the sample complexity of lower bound with respect to K. By this way, we show that under certain conditions, the lower bound matches the upper bound as K grows. In the first example, we assume in bandit model v, arm 1 dominates over the other $K-1$ suboptimal arms, in the sense that the ratio of the expected reward to expected cost is much larger than the other suboptimal arms. Therefore, $\mathbb{P}_v(\mathcal{E}) \to \delta$ as K grows. Then, as we have stated in the previous paragraph, the sample complexities of any algorithms grow linearly with respect to K. On the contrary, in the second example, we assume that in bandit model v, the expected rewards and the expected costs for all the

$K-1$ suboptimal arms are very similar. Thus, their ratios of reward to cost are nearly the same. Under such circumstance, $\mathbb{P}_v(\mathcal{E}) \to \frac{\delta}{K-1}$ with high probability. One can easily check that when δ is fixed, $\mathrm{D}(1-\delta\|\frac{\delta}{K-1})$ grows logarithmatically with K. Moreover, the expression $\sum_{j=1}^{K}\frac{1}{D_j^*}$ is a linear function of K. Consequently, the lower bound in Theorem 4 has a growth rate $K\log K$ and matches the upper bound of the proposed algorithms. These theoretical results will also be validated by numerical simulations.

VI. Numerical Results

A. Parameter Settings

In this section, we evaluate our designed algorithms by numerical simulations. Each point in the figures is averaged over 100 realizations. We consider a 10-armed model and a 30-armed model. Gaussian and Bernoulli distributions are used in simulations. Thus, four types of bandit models are considered in our settings:

- A 10-armed bandit model with Gaussian distribution.
- A 30-armed bandit model with Gaussian distribution.
- A 10-armed bandit model with Bernoulli distribution.
- A 30-armed bandit model with Bernoulli distribution.

For all bandit models, arm 1 is optimal. Table I shows the detailed parameters for each arm in the 10-armed bandits. The 30-armed bandits are generated by duplicating the last five arms in 10-arm bandit for five times, while arm 1 is still the optimal arm. The variances of rewards and costs for all arms with Gaussian distribution are 0.1, while the variance for each arm under Bernoulli settings depends on its specific expectations.

TABLE I
THE EXPECTATIONS OF EACH ARM IN 10-ARMED BANDIT MODEL.

No.	1	2	3	4	5	6	7	8	9	10
μ	0.7	0.5	0.9	0.8	0.1	0.5	0.8	0.2	0.3	0.1
c	0.1	0.3	0.6	0.7	0.1	0.5	0.8	0.5	0.9	0.9

B. Comparisons of Different Algorithms

Fig. 1(a) and Fig. 1(b) illustrate the sample complexities of the three proposed algorithms under Gaussian distribution. From these two figures, we can make two observations: i) for all three algorithms, the sample complexity increases as δ decreases; ii) GA algorithm achieves best performance while SEUG algorithm has the highest sample complexity. Although SEUG-V algorithm is not as good as GA algorithm, it is much better than SEUG algorithm. The first observation is intuitive since that in the GA algorithm, the player has the prior knowledge of gaps for each suboptimal arm. Therefore, the sampling time for each arm could be accurately estimated by the Hoeffding bound. On the contrary, in the SEUG algorithm, the player has no such knowledge. Consequently, it has to spend more time on pulling each arm to narrow the confidence levels of empirical means, which results in the much larger sample complexity. The second observation is because that the incorporation of variance information achieves a finer-grained empirical means of both rewards and costs, which further leads to significantly reduced sample complexity. Moreover, as we will see in Fig. 3(a), the smaller variances, the fewer sample complexity.

Fig. 1(c) and Fig. 1(d) show the sample complexities under Bernoulli distribution for both the 10-armed bandit model and 30-armed bandit model. Similarly, we can see that the SEUG algorithm has the largest sample complexity, while the GA algorithm achieves the smallest one. The SEUG-V algorithm is better than the SEUG algorithm and worse than the GA algorithm, which agrees with our theoretical results.

C. Lower Bound Analysis

In this subsection, we demonstrate the tightness of the lower bound of Theorem 4 in Fig. 2 with respect to $\frac{1}{\delta}$ and K, respectively, where LB is the abbreviation of Lower Bound. We take the average of 50 independent simulation runs for all three algorithms. In this experiment, we apply a different bandit model from Table I, which are presented in the Table II. We can see that arm 1 is still the optimal arm and the suboptimal arms ratios are more similar, which is consistent with the second example in Section V-B.

TABLE II
THE PARAMETERS OF A 10-ARMED BANDIT MODEL IN LOWER BOUND ANALYSIS.

No.	1	2	3	4	5	6	7	8	9	10
μ	0.7	0.5	0.6	0.7	0.5	0.8	0.7	0.6	0.5	0.4
c	0.3	0.4	0.5	0.6	0.5	0.9	0.8	0.7	0.4	0.5

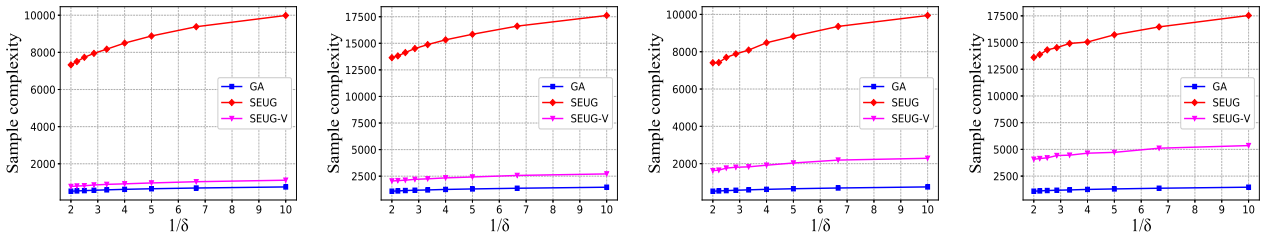
Fig. 2(a) shows the asymptotic performance of the lower bound and algorithms with respect to $\frac{1}{\delta}$ with $K=10$. An zoom-in view figure for the lower bound curve is shown in Fig. 2(b). We can see that the sample complexities for all the three algorithms and the lower bound follow the growth rate $\log\frac{1}{\delta}$.

Fig. 2(c) shows the asymptotic performance of the lower bound and algorithms with respect to K. When $K>10$, we simply set parameters $\mu_i=0.3$ and $c_i=0.4$ for any $i>10$. The LB-ideal curve corresponds to the seconde example we discussed in Section V-B, where the suboptimal arms in a bandit model are very similar with each other. Fig. 2(d) is an zoom-in view of LB and LB-ideal. From Fig. 2(c) and 2(b), we can see that the sample complexities of algorithms and LB-ideal have the same asymptotic performance, which is $O(K\log K)$. In addition, the fundamental lower bound grows linearly with K. These results verify our theoretical analyses.

D. Impacts of Variances

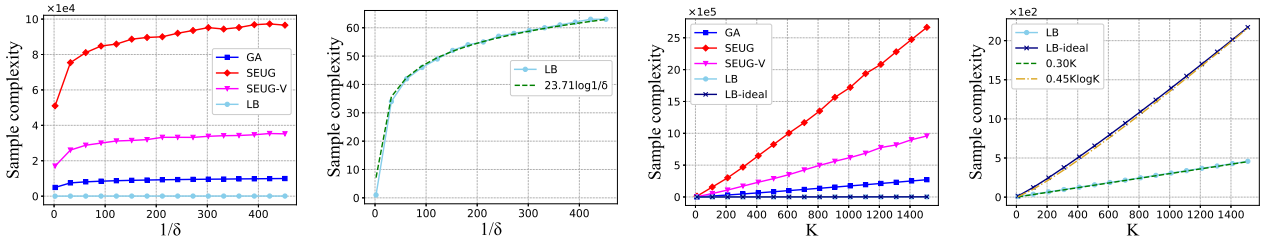
Fig. 3 shows the impacts of variances on sample complexity. The experiments are based on the 10-armed model in Table I with both Gaussian and Bernoulli distributions with $\delta=0.15$. Similarly, each point on the figures is averaged over 100 realizations.

Fig. 3(a) presents the impacts of variances under Gaussian distribution. The parameter settings for each arm are the same as that of Table I. The variances of reward and cost for optimal arm remain 0.1, while those of the other suboptimal arms change from 0.1 to 0.5. We can see that the SEUG algorithm



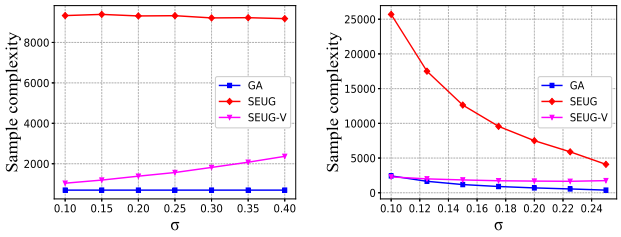
(a) K=10, Gaussian (b) K=30, Gaussian (c) K=10, Bernoulli (d) K=30, Bernoulli

Fig. 1. The sample complexity under different bandit model and distributions.



(a) Comparisons vs. $\frac{1}{\delta}$ (b) Lower bound vs. $\frac{1}{\delta}$ (c) Comparisons vs. K (d) Lower bound vs. K

Fig. 2. The asymptotic performance of lower bound.



(a) Gaussian (b) Bernoulli

Fig. 3. The impacts of variances on sample complexity.

achieves the largest sample complexity while GA algorithm has the smallest. The performance SEUG-V algorithm is better than SEUG and poorer than GA. Such results confirm our theoretical analyses. In addition, as variance grows, the sample complexity of GA and SEUG algorithms remain nearly constant, while the sample complexity of SEUG-V algorithm grows. The reason is that the sample complexities of GA and SEUG algorithms is independent of variances, while the sample complexity of SEUG-V algorithm is a function of variances. Essentially, a larger variance means that each sample value has a large deviation with respect to its expectation, which makes it harder for the average of sample values to converge to its expectation and requires more pulling times.

Fig. 3(b) shows the impacts of variances for Bernoulli distributions. In Bernoulli distributions, the variance depends on the expectation and has the largest value 0.25. Thus, we assume variances of these suboptimal arms range from 0.1 to 0.25. Similar to the Gaussian setting, SEUG-V is better than SUEG and poorer than GA. Moreover, we can see that as variances grow, the sample complexity of the SEUG algorithm decreases sharply while those of the other two algorithms decrease more gradually. This is because we assume the variances grow with the expectations. Thus, for each suboptimal arm with Bernoulli distribution, larger variances represent larger expectations for both reward and cost. Since we keep the reward and cost expectations of the optimal arm unchanged, the gaps ξ_i between the optimal arm and suboptimal arms will become larger with the growth of variances. Consequently, the sample complexity will be reduced.

VII. Acknowledgment

This work was supported in part by National Key R&D Program of China 2018AAA0101200, NSFC China (No. 61672342, 61671478, 61532012, 61822206, 61829201, 61960206002), the Science and Technology Innovation Program of Shanghai (Grant 18XD1401800, 17511105103, 18510761200), in part by Shanghai Key Laboratory of Scalable Computing and Systems, NSF grants ECCS-1818791, CCF-1758736, and CNS-1758757.

VIII. Conclusion

In this paper, we study the exploring best arm with reward to cost problem with fixed confidence budget δ. Each arm in our settings is associated with a random reward and a random cost, and the best arm is defined as the one with the largest ratio of the expected reward to expected cost. We proposed three algorithms to address this problem and proved that the sample complexities for all algorithms grow logarithmically as δ grows. In addition, we derived a fundamental lower bound of sample complexities for any algorithms under Bernoulli distributions, and proved that the sample complexities of the proposed three algorithms match that of the lower bound.

Under certain settings, in our future work, we will consider how to derive the probable approximately correct bound for our arm settings, i.e., to find a ϵ-optimal arm? We will also investigate how to characterize the relationships and find the best arm if the reward and cost are dependent with each other.

REFERENCES

[1] E. Even-Dar, S. Mannor, and Y. Mansour, "PAC bounds for multi-armed bandit and markov decision processes", in COLT 2002.

[2] Y. Xue, P. Zhou, S. Mao, D. Wu and Y. Zhou, "Pure-exploration bandits for channel selection in mission-critical wireless communications", *IEEE Trans. on Vehicular Technology*, vol. 67, No. 11, pages 10995-11007, Nov. 2018.

[3] J. Scarlett, I. Bogunovic and V. Cevher, "Overlapping multi-bandit best arm identification", in ISIT 2019.

[4] F. Li, J. Liu and B. Ji, "Combinatorial sleeping bandits with fairness constraints", in INFOCOM 2019.

[5] J.-Y. Audibert, S. Bubeck, and R. Munos, "Best arm identification in multi-armed bandits", in COLT 2010.

[6] O. Madani, D. Lizotte and R. Greiner, "The budgeted multi-armed bandit problem", in COLT 2004.

[7] L. Chen and J Xu, "Task replication for vehicular cloud: contextual combinatorial bandit with delayed feedback", in INFOCOM 2019.

[8] A. R. Chaudhuri and S. Kalyanakrishnan "PAC identification of a bandit arm relative to a reward quantile", in AAAI 2017

[9] H. Li and Y. Xia, "Infinitely many-armed bandits with budget constraints", in AAAI 2017.

[10] X. Yu, H. Shao, M.R. Lyu and I. King, " Pure exploration of multi-armed bandits with heavy-tailed payoffs", in UAI 2018.

[11] W. Cao, J. Li, Y. Tao, Z. Li, "On top-k selection in multi-armed bandits and hidden bipartite graphs", in NeurIPS 2015.

[12] V. Gabillon, M. Ghavamzadeh and A. Lazaric, "Multi-bandit best arm identification", in NeurIPS 2011.

[13] V. Gabillon, M. Ghavamzadeh and A. Lazaric, "Best arm identification: a unified approach to fixed budget and fixed confidence", in NeurIPS 2012.

[14] Y. Xia, T. Qin, N. Yu, T. Liu, "Best action selection in a stochastic environment", in AAMAS 2016.

[15] W. Ren, J. Liu, and N. Shroff, "Exploring k out of top ρ fraction of arms in stochastic bandits", in AISTATS 2019.

[16] O. Maron and A. W. Moore, "Hoeffding races: accelerating model selection search for classification and function approximation", in NeurIPS 1993.

[17] K. Jamieson and R. Nowak, " Best-arm Identification Algorithms for Multi-Armed Bandits in the Fixed Confidence Setting", in CISS 2014.

[18] S. Bubeck, R. Munos, and G. Stoltz, "Pure exploration in multi-armed bandits problems", In Proc. of the 20th International Conference on Algorithmic Learning Theory, 2009.

[19] V. Mnih, C. Szepesvári and J.-Y. Audibert, "Empirical bernstein stopping", in ICML 2008.

[20] Y. Zhou, X. Chen and J. Li, "Optimal PAC multiple arm identification with applications to crowdsourcing", in ICML 2014.

[21] A. Maurer and Massimiliano Pontil, "Empirical bernstein bounds and sample variance penalization", in COLT 2009.

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

[23] E. Kaufmann, O. Cappé and A. Garivier, "On the complexity of best-arm identification in multi-armed bandit models", *Journal of Machine Learning Research*, vol. 17, pp. 1-42, 2016.

[24] S. Kalyanakrishnan and P. Stone, "Efficient selection of multiple bandit arms: theory and practice", in ICML 2010.

[25] S. Kalyanakrishnan, A. Tewari, P. Auer, and P. Stone, "Pac subset selection in stochastic multiarmed bandits", in ICML 2012.

[26] A. Carpentier and M. Valko, "Simple regret for infinitely many armed bandits", in ICML, 2015.

[27] M. Hashemi, N. B. Shroff, C. E. Koksal and A. Sabharwal, "Efficient beam Alignment in millimeter wave systems using contextual bandits", in INFOCOM 2018.

[28] Y. Wu, A. Gyorgy, and C. Szepesvari, "On identifying good options under combinatorially structured feedback in finite noisy environments", in ICML 2015, pages 1283-1291.

[29] S. Bubeck, R. Munos, and G. Stoltz, "Pure exploration finitely-armed and continuous-armed bandits", in *Theoretical Computer Science*, vol. 412, issue 19, pages 1832-1852, 2011.

[30] T. L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules", in *Advances in Applied Mathematics*, vol. 6, no. 1, March, 1985.

[31] Da. P. Zhou and C. J. Tomlin, "Budget-constrained multi-armed bandits with multiple plays", in AAAI 2018.

[32] B. Stec and W. Susek, "Theory and measurement of signal-to-noise ratio in continuous-wave noise radar", in *sensors*, vol. 18, no. 5, pp. 1445, May 2018.

[33] M. Soare, A. Lazaric and R. Munos, "Best-arm identification in linear bandits", in NeurIPS 2014.