Cost Aware Best Arm Identification

Kellen Kanarios
kellenkk@umich.edu
University of Michigan, Ann Arbor

Qining Zhang
qiningz@umich.edu
University of Michigan, Ann Arbor

Lei Ying
leiying@umich.edu
University of Michigan, Ann Arbor

Abstract

In this paper, we study a best arm identification problem with dual objects. In addition to the classic reward, each arm is associated with a cost distribution and the goal is to identify the largest reward arm using the minimum expected cost. We call it *Cost Aware Best Arm Identification* (CABAI), which captures the separation of testing and implementation phases in product development pipelines and models the objective shift between phases, i.e., cost for testing and reward for implementation. We first derive a theoretical lower bound for CABAI and propose an algorithm called CTAS to match it asymptotically. To reduce the computation of CTAS, we further propose a simple algorithm called *Chernoff Overlap* (CO), based on a square-root rule, which we prove is optimal in simplified two-armed models and generalizes well in numerical experiments. Our results show that (i) ignoring the heterogeneous action cost results in sub-optimality in practice, and (ii) simple algorithms can deliver near-optimal performance over a wide range of problems.

1 Introduction

The stochastic multi-armed bandit (MAB) (Thompson, 1933; Robbins, 1952) is a classic model which has widespread applications, from content recommendation (Kohli et al., 2013), resource allocation (Liu et al., 2020), clinical trials (Villar et al., 2015), to efficient ad placement. A multi-armed bandit problem involves an agent and an environment, which is represented by a set of actions (arms) with distinct underlying reward distributions. At each round, the agent will choose one of the arms and then obtain a random reward generated from the associated distribution. Most existing studies formulate MAB either as the best arm identification (BAI) problem (Kaufmann et al., 2016; Garivier & Kaufmann, 2016) where the agent intends to identify the highest reward arm as quickly as possible or as a regret minimization problem (Auer, 2002; Garivier & Cappé, 2011) where the goal is to maximize the cumulative reward over a time-horizon. Both formulations have been well-developed and successful in balancing the trade-off between exploration and exploitation.

However, unlike the classic MAB model, most real-world product development pipelines are usually separated into two phases: testing (survey) and implementation (release). Here, testing refers to the process where one intends to find the best product among a set of potential candidates through sequential trials, e.g., A/B testing for clinical decisions. The implementation phase refers to the selected best product being used in a wider population after testing, usually involving mass production. Different performance measures are emphasized in different phases. For example, the cost of prototype medicine may be of primary concern in testing while the efficacy is more important in implementation since the production cost is either decreased via mass production or covered through insurance. Similarly, when choosing the best platform for online advertising, the total payment to different platforms for advertising may be of essential concern in testing, while the click-through and

conversion rate are what matters in implementation after the best platform is selected. Unfortunately, neither BAI nor regret minimization captures the aforementioned differences, which makes algorithms developed for traditional MAB not directly applicable. Moreover, trying out different candidates during testing may require different costs, which again is not captured in classic MAB.

Cost Aware Best Arm Identification: in this paper, we propose a new MAB problem called *Cost Aware Best Arm Identification* (CABAI), where besides reward (the main object for implementation), each arm is in addition associated with a cost distribution to model the object for testing. Each time the agent chooses an arm, it will observe a random reward-cost pair, which is independently generated from the reward distribution and cost distribution respectively. The goal in CABAI is to identify the highest reward arm using the minimum cost possible, which breaks down to the following questions: (1) how should we sample arms during testing with unknown cost, and how does the rule differ from BAI (**sampling rule**)? (2) when is the best arm identifiable (**stopping rule**)? (3) which arm should we choose for implementation (**decision rule**)? We will show that the design of algorithms for CABAI is related to BAI, but have fundamental differences so that BAI optimal algorithms do not necessarily achieve good performance in CABAI. This also implies that directly applying BAI algorithms and neglecting the heterogeneous nature of arms in practice will result in sub-optimality. We address the following questions in our paper: (1) What are the fundamental limits of CABAI? (2) How should we design efficient algorithms to achieve the limit?

Our Contributions: We propose CABAI and show that traditional BAI algorithms no longer perform well. As summarized in Table 1, the optimal proportions of arm pulls have essential differences between traditional BAI and CABAI, i.e., TAS (Garivier & Kaufmann, 2016), which is optimal in BAI, allocates almost the same amount of pulls to the first two arms, while the optimal proportion of arm pulls for CABAI emphasize more on the low-cost arm, as achieved by our proposed CTAS algorithm. We first prove a non-asymptotic lower bound on the minimum cumulative cost required to identify the best arm. Then, we propose an algorithm called *Cost-Aware Track and Stop* (CTAS) to match the lower bound asymptotically. However, the CTAS algorithm is required to solve a bilevel optimization problem at each time step, which exerts relatively high computational complexity and prevents its direct use in practice. To overcome this issue, we further propose a low-complexity algorithm called *Chernoff Overlap* (CO) which exhibits desirable empirical performance and remains theoretically optimal in simplified bandit models.

Algorithm	Optimal?	$w_1(t)$ $(1.5, 1)$	$w_2(t)$ $(1, 0.1)$	$w_3(t)$ $(0.5, 0.01)$
TAS	×	0.46	0.46	0.08
CTAS	✓	0.23	0.72	0.05

Table 1: The expected rewards are $\boldsymbol{\mu} = [1.5, 1, 0.5]$ and the expected costs are $\boldsymbol{c} = [1, 0.1, 0.01]$. For arm i, $w_i(t)$ is the proportion of arm pulls up to time t, i.e., $w_i(t) = N_i(t)/t$ where N_i is the number of pulls. Noticeably, CABAI emphasizes more on low-cost arms to complement high-cost arms.

1.1 Related Work

We review existing MAB results most relevant to our paper. A detailed discussion is in the appendix.

BAI with Fixed Confidence: BAI has been studied for many years and was originally proposed in Bechhofer (1958). In this paper, we consider a subset known as the fixed confidence setting, where the agent aims to minimize the sample complexity while ensuring the best arm is identified with probability at least $1 - \delta$. Here, δ is a pre-specified confidence level, and such algorithms are called δ-PAC. In Kaufmann et al. (2016), the authors introduce a non-asymptotic lower bound for this setting. Subsequently, they propose the Track and Stop algorithm (TAS) that matches this lower bound asymptotically. The TAS algorithm has since been extended to various other settings (Jourdan et al., 2023; Garivier & Kaufmann, 2021; Kato & Ariu, 2021). Before it, researchers proposed "confidence-based" algorithms, e.g., KL-LUCB (Kaufmann & Kalyanakrishnan, 2013), UGapE (Gabil-

lon et al., 2012), which rely on constructing high-probability confidence intervals. They are more computationally feasible than TAS inspired algorithms, but with few theoretical guarantees.

BAI with Safety Constraints: A formulation similar to our paper is BAI with safety constraints (Wang et al., 2022). As a motivating example, they consider the clinical trial setting, where each drug is associated with a dosage and the dosage has an associated safety level. They attempt to identify the best drug and dosage for fixed confidence without violating the safety level. Similarly, Hou et al. (2023) attempts to identify the best arm subject to a constraint on the variance. Our formulation is distinct from them because the agent is free to perform any action. In Chen et al. (2022b;a), they formulate safety constraints as a constrained optimization problem. They explore and show that allowing minimal constraint violations can provide significant improvement in the regret setting. This is distinct from the BAI setting explored in this paper.

Multi-fidelity BAI: An alternative formulation that considers cost is the multi-fidelity formulation introduced in Kandasamy et al. (2016) and recently considered in the best arm identification regime (Poiani et al., 2022; Wang et al., 2023). In this setting, along with choosing an arm, the agent chooses the desired fidelity or "level of accuracy" of the mean estimate. Each fidelity incurs a cost, where higher fidelity incurs a larger cost but provides more accurate estimate. This setting clearly differs from ours because the cost of each fidelity is known a priori and is controllable through choice of fidelity.

2 Preliminaries

We study a model similar to the fixed-confidence BAI in stochastic K-armed bandits. We denote the set of arms as $\mathcal{A} \coloneqq \{1, 2, \cdots, K\}$. Each arm a is associated with a reward distribution $\boldsymbol{\nu}_\mu = \{\nu_{\mu_1}, \ldots, \nu_{\mu_K}\}$ with expectations $\boldsymbol{\mu} \coloneqq \{\mu_1, \mu_2, \cdots, \mu_K\}$. We assume $\boldsymbol{\nu}_\mu$ are independent and make the natural exponential family assumption standard in BAI literature (Kaufmann et al., 2016):

Assumption 1 (Natural Exponential Family)**.** *For any a, ν_{μ_a} belongs to family $\mathcal{P}$ which can be parameterized by the expectation with finite moment generating function, i.e.,*

$$\mathcal{P} = \{\nu_\mu | \mu \in [0, 1], \nu_\mu = h(x) \exp(\theta_\mu x - b(\theta_\mu))\},$$

where θ_μ is a function of μ, and $b(\theta)$ is convex and twice differentiable.

For two different expectations μ and μ' with the same exponential family, we use $d(\mu, \mu')$ to denote the KL-divergence from ν_μ to $\nu_{\mu'}$. Note that Assumption 1 is very general and includes a large class of distributions such as Gaussian (with known variance), Bernoulli, and Poisson distributions by considering the following choice of parameters:

$$\begin{aligned}
\text{Bernoulli}: &\quad \theta_\mu = \log\left(\frac{\mu}{1-\mu}\right), && b(\theta_\mu) = \log\left(1 + e^{\theta_\mu}\right), && h(x) = 1 \\
\text{Poisson}: &\quad \theta_\mu = \log(\mu), && b(\theta_\mu) = e^{\theta_\mu}, && h(x) = \frac{1}{x!} e^{-x} \\
\text{Gaussian}: &\quad \theta_\mu = \frac{\mu}{\sigma^2}, && b(\theta_\mu) = \frac{\sigma^2 \theta_\mu^2}{2}, && h(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2\sigma^2}
\end{aligned}$$

Unique to this work, we assume that each arm has a cost with distribution $\boldsymbol{\nu}_c \coloneqq \{\nu_{c_1}, \ldots, \nu_{c_K}\}$ and expectations $\boldsymbol{c} \coloneqq \{c_1, \ldots, c_K\}$. We assume they satisfy the positivity assumption, which is natural in our motivating examples in real world such as ad placement or clinical trials, where the cost of each action is always bounded and all actions are not free.

Assumption 2 (Bounded Positivity)**.** *For any arm a, we assume the support of the cost distribution ν_{c_a} is positive and bounded away from 0, i.e., $\mathsf{supp}(\nu_{c_a}) \in [\ell, 1]$, where ℓ is a positive constant.*

Problem Formulation: We define the best arm $a^*(\boldsymbol{\mu})$ to be the action which has the highest expected reward, i.e., $a^*(\boldsymbol{\mu}) = \arg\max_{a \in \mathcal{A}} \mu_a$, and we assume there is a unique best arm. The

results can be generalized to scenarios with multiple best arms given the number of best arms. At each round (time) $t \in \mathbb{N}^+$, we interact with the environment by choosing an arm $A_t \in \mathcal{A}$. After that, a (reward, cost) signal pair (R_t, C_t) is independently sampled from the joint distribution $\nu_{\mu_{A_t}} \times \nu_{c_{A_t}}$ of the action that we choose. For any time t, we use $N_a(t)$ to denote the number of times that arm a has been pulled, and we use $\widehat{\mu}_a(t)$ and $\widehat{c}_a(t)$ to denote the empirical average reward and cost:

$$\widehat{\mu}_a(t) = \frac{1}{N_a(t)} \sum_{k=1}^{t} R_k \cdot \mathbf{1}_{\{A_k = a\}}, \quad \widehat{c}_a(t) = \frac{1}{N_a(t)} \sum_{k=1}^{t} C_k \cdot \mathbf{1}_{\{A_k = a\}}.$$

For any policy π, it consists of three components: (1) a sampling rule $(A_t)_{t \geq 1}$ to select arms to interact at each round; (2) a stopping time τ_δ which terminates the interaction; and (3) an arm decision rule $\widehat{a}$ to identify the best arm. As a convention of BAI with fixed confidence, we require our policy π to be δ-*PAC* (Probably Approximately Correct) (Kaufmann et al., 2016), which means the algorithm should terminate in finite time and the probability of choosing the wrong best arm should be lower than the confidence level δ. The definition of δ-PAC is as follows:

Definition 1. *An algorithm π is δ-PAC if for any reward and cost instances $(\boldsymbol{\mu}, \boldsymbol{c})$, it outputs the best arm $a^*(\boldsymbol{\mu})$ with probability at least $1 - \delta$ and in finite time almost surely, i.e.,*

$$\mathbb{P}_{\boldsymbol{\mu} \times \boldsymbol{c}}(\widehat{a} \neq a^*(\boldsymbol{\mu})) \leq \delta, \quad \mathbb{P}_{\boldsymbol{\mu} \times \boldsymbol{c}}(\tau_\delta < \infty) = 1.$$

For any time t, define the cumulative cost as $J(t) \coloneqq \sum_{k=1}^{t} C_k$. For any fixed δ, let Π_δ denote the set of all δ-PAC best arm identification policies. The goal of this work is to find $\pi \in \Pi_\delta$, such that $\pi = \arg\min_{\pi \in \Pi_\delta} \mathbb{E}_{\boldsymbol{\mu} \times \boldsymbol{c}}[J(\tau_\delta)]$. We use boldface $\boldsymbol{x}$ to denote vectors and instances, and calligraphy $\mathcal{X}$ to denote sets. We use the subscript $\mathbb{P}_{\boldsymbol{\mu} \times \boldsymbol{c}}$, $\mathbb{E}_{\boldsymbol{\mu} \times \boldsymbol{c}}$ to denote the probability measure and expectation with respect to a specific instance $(\boldsymbol{\mu}, \boldsymbol{c})$.

3 Lower Bound

We first characterize the theoretical limits of this cost minimization problem. Denote by $\mathcal{M}$ a set of exponential bandit models such that each bandit model $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$ in $\mathcal{M}$ has a unique best arm $a^*(\boldsymbol{\mu})$. Let $\Sigma_K = \left\{\boldsymbol{w} \in \mathbb{R}_+^k : w_1 + \cdots + w_K = 1\right\}$ to be the set of probability distributions on $\mathcal{A}$, then we present the following theorem which characterizes the fundamental lower bound.

Theorem 1. *Let $\delta \in (0, 1)$. For any δ-PAC algorithm and any bandit model $\boldsymbol{\mu} \in \mathcal{M}$, we have:*

$$\mathbb{E}_{\boldsymbol{\mu} \times \boldsymbol{c}}[J(\tau_\delta)] \geq T^*(\boldsymbol{\mu}) \log \frac{1}{\delta} + o\left(\log \frac{1}{\delta}\right).$$

where $T^*(\boldsymbol{\mu})$ is the instance dependent constant satisfying:

$$T^*(\boldsymbol{\mu})^{-1} = \sup_{\boldsymbol{w} \in \Sigma_K} \inf_{\boldsymbol{\lambda} \in \{a^*(\boldsymbol{\lambda}) \neq a^*(\boldsymbol{\mu})\}} \sum_a \frac{w_a}{c_a} d(\mu_a, \lambda_a).$$

The proof of Theorem 1 is deferred to the appendix but primarily relies on the "transportation" lemma proposed in Kaufmann et al. (2016), which characterizes the theoretical hardness to distinguish the bandit model $\boldsymbol{\mu}$ from any other models $\boldsymbol{\lambda}$ where $a^*(\boldsymbol{\lambda}) \neq a^*(\boldsymbol{\mu})$. Theorem 1 suggests that $\mathcal{O}(\log(1/\delta))$ cumulative cost is inevitable to identify the optimal arm, and it also characterizes the asymptotic lower bound constant $T^*(\boldsymbol{\mu})$.

Instance Dependent Constant $T^*(\boldsymbol{\mu})$: The instance dependent constant $T^*(\boldsymbol{\mu})$ obtained in our Theorem. 1 is different from classic best arm identification lower bounds, e.g., $T^*(\boldsymbol{\mu})$ in Theorem 1 of Garivier & Kaufmann (2016). Even though it captures the hardness of this instance in terms of the cumulative cost, $T^*(\boldsymbol{\mu})$ is still a vague notion in the sense that the relationship between the theoretical cumulative cost $J(\tau_\delta)$ and model parameters, $\boldsymbol{\mu}$, $\boldsymbol{c}$, is still unclear. To better understand this mysterious constant $T^*(\boldsymbol{\mu})$, we present Theorem. 1 in the simple case of 2 armed Gaussian bandits with unit variance, where $T^*(\boldsymbol{\mu})$ has a closed-form expression.

Algorithm 1: **C**ost-adapted **T**rack **A**nd **S**top (CTAS)

Input: confidence δ; $\alpha \geq 1$; sufficiently large B; oracle function $\mathsf{ComputeProportions}(\boldsymbol{\mu}, \boldsymbol{c})$.
pull each arm $a \in \mathcal{A}$ once as initialization;
for $t \geq K+1$ **do**
  forced exploration set $\mathcal{U}_t = \{a \mid N_a(t) < \sqrt{t}\}$;
  $\boldsymbol{w}^* = \mathsf{ComputeProportions}(\widehat{\boldsymbol{\mu}}(t), \widehat{\boldsymbol{c}}(t))$; // compute optimal proportion
  if $\mathcal{U}_t \neq \emptyset$ **then** // Sampling Rule
    pull the least-pulled arm: $a_t \in \operatorname*{argmin}_{a \in \mathcal{A}} N_a(t)$
  else
    $a_t \in \operatorname*{argmax}_{a \in \mathcal{A}} J(t) w_a^* - \widehat{c}_a(t) N_a(t)$; // pull the arm with largest deficit
  if $Z(t) > \log\left(\frac{Bt^\alpha}{\delta}\right)$ **then** // Stopping Rule
    break;
return $\widehat{a} = \operatorname*{argmax}_{a \in \mathcal{A}} \widehat{\mu}_a(t)$; // Decision Rule

Corollary 1. *Let $\delta \in (0,1)$. For any δ-PAC algorithm and any 2-armed Gaussian bandits with reward expectations $\{\mu_1, \mu_2\}$ and unit variance such that $\mu_1 > \mu_2$, we have:*

$$\mathbb{E}_{\boldsymbol{\mu} \times \boldsymbol{c}}\left[J(\tau_\delta)\right] \geq \frac{2\left(\sqrt{c_1} + \sqrt{c_2}\right)^2}{(\mu_1 - \mu_2)^2} \log \frac{1}{\delta} + o\left(\log \frac{1}{\delta}\right).$$

It is noticeable that the dependence on cost is non-trivial but somehow involves the square root $\sqrt{c_a}$ for each action. This inspires our low-complexity algorithm Chernoff Overlap (CO) based on a square-root rule. The lower bound of a slightly more general setting is provided in the appendix.

The Optimal Weight $\boldsymbol{w}^*$: Let $\boldsymbol{w}^* = \{w_a^*\}_{a \in \mathcal{A}}$ be the solution of the sup-inf problem in the definition of $T^*(\boldsymbol{\mu})$ in Theorem 1. The weight $\boldsymbol{w}^*$ is essential in designing efficient algorithms to match the lower bound, as it characterizes the optimal proportion of the total cumulative cost from pulling arm a. Concretely, any algorithm which matches the lower bound should satisfy:

$$\lim_{\delta \to 0} \frac{c_a \mathbb{E}_{\boldsymbol{\mu} \times \boldsymbol{c}}\left[N_a(\tau_\delta)\right]}{\mathbb{E}_{\boldsymbol{\mu} \times \boldsymbol{c}}\left[J(\tau_\delta)\right]} = w_a^*, \quad \forall a \in \mathcal{A} \tag{1}$$

This differs from Garivier & Kaufmann (2016), where w_a is the proportion of rounds that arm a is pulled. Like $T^*(\boldsymbol{\mu})$, there is no closed-form expression for $\boldsymbol{w}^*$ in general bandit models with $K \geq 3$. In the appendix, we show that one can compute the desired quantities such as $T^*(\boldsymbol{\mu})$ and $\boldsymbol{w}^*$ by similarly solving K continuous equations to Garivier & Kaufmann (2016). Therefore, we can readily apply iterative methods such as bisection to compute these values. We summarize this procedure in the ComputeProportions Algorithm (Algorithm. 3 in the Appendix), which will be called regularly as a sub-routine in our proposed algorithms (Algorithm. 1).

4 Asymptotically Cost Optimal Algorithm

In this section, we propose a BAI algorithm called **C**ost-aware **T**rack **A**nd **S**top (CTAS) whose cumulative cost performance asymptotically matches the lower bound in Theorem 1 both in expectation and almost surely. We discuss each of the sampling, stopping and decision rules for CTAS:

Sampling Rule: From (1), a necessary condition for the optimal algorithm is derived. Our sampling rule in Algorithm 1 strives to match the proportion of the cost of each arm to the optimal proportion $\boldsymbol{w}^*(\boldsymbol{\mu})$. First, we force the empirical proportions $\widehat{w}_a = \widehat{c}_a N_a(t)/J(t)$ to not differ too greatly from the empirically optimal weights $\boldsymbol{w}^*(\widehat{\boldsymbol{\mu}})$ using a largest-deficit-first like arm selection policy. We will

show that as the empirical mean $\widehat{\boldsymbol{\mu}} \to \boldsymbol{\mu}$, we will have $\boldsymbol{w}^*(\widehat{\boldsymbol{\mu}}) \to \boldsymbol{w}^*(\boldsymbol{\mu})$, and the empirical cost proportion will also converge and concentrate along the optimal proportion $\boldsymbol{w}^*(\boldsymbol{\mu})$.

Forced Exploration: Also present in Algorithm 1 is the forced exploration, which pulls the least-pulled arm when $\mathcal{U}(t)$ is not empty (Line 6). This ensures each arm is pulled at least $\Omega(\sqrt{t})$ times, and makes sure that our plug-in estimate of $\boldsymbol{w}^*$ is sufficiently accurate. The $\sqrt{t}$ rate of forced exploration is carefully chosen to balance the sample complexity and the convergence rate of the empirical mean. If chosen too small, the fraction of cost from different arms will concentrate along the inaccurate estimation which results in sub-optimality. If chosen too large, the forced exploration will dominate the sampling procedure, leading to an almost uniform exploration which is sub-optimal.

Stopping Rule and Decision Rule: We utilize the Generalized Likelihood Ratio statistic (Chernoff, 1959) between the observations of arm a and arm b $Z_{a,b}(t)$. For an arbitrary exponential family, $Z_{a,b}(t)$ has a closed-form expression as follows:

$$Z_{a,b}(t) = N_a(t) d\left(\widehat{\mu}_a(t), \widehat{\mu}_{a,b}(t)\right) + N_b(t) d\left(\widehat{\mu}_b(t), \widehat{\mu}_{a,b}(t)\right),$$

where $\widehat{\mu}_{a,b}(t) = \widehat{\mu}_{b,a}(t)$ is defined:

$$\widehat{\mu}_{a,b}(t) := \frac{N_a(t)}{N_a(t) + N_b(t)} \widehat{\mu}_a(t) + \frac{N_b(t)}{N_a(t) + N_b(t)} \widehat{\mu}_b(t).$$

In particular, the Chernoff statistics $Z(t) = \max_{a \in \mathcal{A}} \min_{b \in \mathcal{A}, b \neq a} Z_{a,b}(t)$ measures the distance between an instance where the current empirical best arm is indeed the best arm, and the "closest" instance where the current empirical best arm is not the true best arm, both reflected through reward observations. So, the larger $Z(t)$ is, the more confident that the empirical best arm is indeed the best arm. The proposition below ensures the δ-PAC guarantee of CTAS.

Proposition 1 (δ-PAC)**.** *Let $\delta \in (0,1)$ and $\alpha \geq 1$. There exists a constant B_α[1] such that for all $B \geq B_\alpha$ the* CTAS *algorithm in Algorithm. 1 is δ-PAC, i.e.,*

$$\mathbb{P}_{\boldsymbol{\mu} \times \boldsymbol{c}}\left(\tau_\delta < \infty, \widehat{a}_{\tau_\delta} \neq a^*\right) \leq \delta.$$

The cost bandit setting also encourages the algorithm to stop as early as possible, so the same stopping rules from traditional BAI (Garivier & Kaufmann, 2016) can be used. A more refined threshold can be found in Kaufmann & Koolen (2021). However, we will use the threshold in Algorithm 1 for the rest of the paper for simplicity. Our Proposition. 1 combines Theorem 10 and Proposition 11 from Garivier & Kaufmann (2016), and the proof will be provided in the appendix.

Asymptotic Optimality for CTAS: In Theorem. 2, we provide provable cost guarantees for the CTAS algorithm. Namely, the algorithm asymptotically achieves the lower bound in Theorem 1 in expectation as the confidence level δ decreases to 0.

Theorem 2 (Expected Upper Bound)**.** *Let $\delta \in [0,1)$ and $\alpha \in [1, e/2]$. Using Chernoff's stopping rule with $\beta(t,\delta) = \log(\mathcal{O}(t^\alpha)/\delta)$, the* CTAS *algorithm ensures:*

$$\limsup_{\delta \to 0} \frac{\mathbb{E}_{\boldsymbol{\mu} \times \boldsymbol{c}}\left[J(\tau_\delta)\right]}{\log(1/\delta)} \leq \alpha T^*(\boldsymbol{\mu}).$$

For optimality, we can simply take $\alpha = 1$ and choose $B \geq 2K$ from Proposition 1. The proof of the theorem along with a weaker almost sure cost upper bound result (Theorem. 6) will be provided in the appendix. The major difference of the expected upper bound and the weaker version is the rate of exploration, where Theorem. 2 requires $\mathcal{O}(\sqrt{t})$ forced exploration rate while the weaker version suffice with $o(t)$. We first show the empirical proportion of the cost for each arm converges to the optimal proportion (Theorem. 5), with the help of forced exploration rate. Then, the Chernoff stopping time ensures our algorithm stops early to guarantee δ-PAC and to minimize the cost.

[1] B_α satisfies $B_\alpha \geq 2K$ for $\alpha = 1$, or $\sum_{t=1}^{\infty} \frac{e^{K+1}}{K^K} \frac{\left(\log^2(B_\alpha t^\alpha) \log t\right)^K}{t^\alpha} \leq B_\alpha$ for $\alpha > 1$.

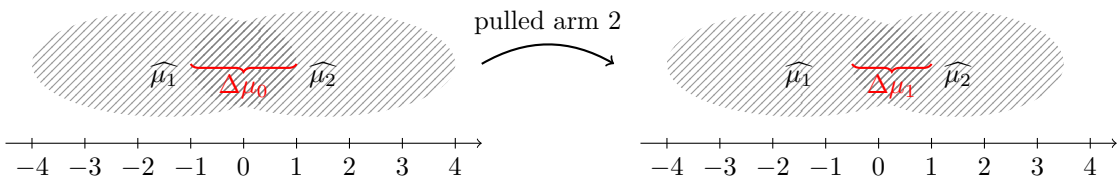


Figure 1: Change in overlap upon pulling arm 2, where the ellipsoids stand for confidence intervals. Left: wider confidence interval for μ_2. Right: reduced confidence interval upon pulling arm 2.

5 Low Complexity Algorithm

Even though CTAS achieves asymptotically optimal cost performance, this algorithm suffers from the heavy computation time of computing $\boldsymbol{w}^*$. As shown in Table 2 in Section 6, the CTAS and TAS algorithm requires much more time to compute the sampling rule at each time step. This leads to the desire for a "model-free" algorithm that does not require us to compute $\boldsymbol{w}^*$. In this section, we propose a low-complexity algorithm called Chernoff-Overlap (CO) which is summarized in Algorithm 2. CO is based on action elimination. The main idea behind these algorithms is to sample each arm uniformly and then eliminate arms that can be declared sub-optimal with high probability. However, it is easy to see that sampling uniformly would not be a good idea in the case of heterogeneous costs. This requires that the sampling rule take into account the proper ratio of information gained from pulling an arm concerning the cost of that arm.

Sampling Rule: To gain maximum information on the remaining uncertainty of reward, it is desirable to pull the arm with the largest decrease in "overlap" as shown in Fig. 1, which results in the arm with minimum pulls $N_t(a)$. However, we also need to consider the cost of arms and weigh the decrease of overlap with cost. Through analysis of the two-armed Gaussian setting, this leads to our choice of sampling rule which weighs $N_a(t)$ with $\sqrt{c_a}$, called the square-root rule.

Stopping Rule: Our stopping rule will still rely on the generalized likelihood ratio. For any time t, Let $a^*(t)$ be the empirical best arm, i.e., $a^*(t) = \arg\max_a \widehat{\mu}_a(t)$. The Chernoff statistics we adopt in CO is instead the pairwise statistic $Z_{a^*(t),a}(t)$. When it is large, the empirical reward $\widehat{\mu}_a(t)$ of arm a is significantly lower than the empirical reward of $a^*(t)$, which gives us high confidence to eliminate this arm. Naturally, we then stop when only one arm remains. The following proposition ensures that by the choice of a proper threshold, CO is δ-PAC. The proof will be in the appendix.

Proposition 2 (δ-PAC)**.** *Let $\delta \in (0,1)$ and $\alpha \geq 1$. There exists a large enough constant B[2] such that the Chernoff-Overlap algorithm in Algorithm. 2 is δ-PAC, i.e.,*

$$\mathbb{P}_{\boldsymbol{\mu}\times\boldsymbol{c}}\left(\tau_\delta < \infty, \widehat{a}_{\tau_\delta} \neq a^*\right) \leq \delta.$$

Cost Upper Bound for Chernoff-Overlap: It is difficult to relate an algorithm to the general cost lower bound in Theorem 1 without direct tracking. Therefore, we must resort to relating the cost upper bound of Chernoff-Overlap to the lower bound in cases where there is a closed-form solution. We consider the two-armed Gaussian bandits setting and show that Chernoff-Overlap is asymptotically cost-optimal for this special case, resulting in the following Theorem.

Theorem 3. *Let $\delta \in (0,1)$ and $\alpha \in (1, e/2)$. For any 2-armed Gaussian bandit model with rewards $\{\mu_1, \mu_2\}$ and costs $\{c_1, c_2\}$, under the CO algorithm in Algorithm 2 we have with probability 1:*

$$\limsup_{\delta\to 0} \frac{J(\tau_\delta)}{\log(1/\delta)} \leq \frac{2\alpha\left(\sqrt{c_1}+\sqrt{c_2}\right)^2}{(\mu_1-\mu_2)^2}.$$

The proof of Theorem. 3 will be delayed to the appendix. The key to the proof is to show that under our sampling rule balanced by $\sqrt{\widehat{c_a}(t)}$ for each arm, the empirical cost proportion $\widehat{\boldsymbol{w}}(t)$ converges

[2] B can be chosen the same as Proposition 1.

Algorithm 2: Chernoff-Overlap Algorithm

Input: confidence level δ; $\alpha \geq 1$; sufficiently large constant B
pull each arm $a \in \mathcal{A}$ once as initialization;
for $t \geq K+1$ **do**
  if $|\mathcal{R}| \leq 1$ **then** // Stopping Rule
    break;
  eliminate all arms a from $\mathcal{R}$ if $Z_{a^*(t),a}(t) > \log\left(\frac{Bt^\alpha}{\delta}\right)$, where $a^*(t) = \arg\max_a \widehat{\mu}_a(t)$;
  pull arm $a_t \in \operatorname*{argmin}_{a \in \mathcal{R}} \sqrt{c_a} N_a(t)$; // Sampling Rule
return $\widehat{a} \in \mathcal{R}$

to the optimal proportion $\widehat{\boldsymbol{w}}^*$. Then, we can apply a similar argument as the weaker version of Theorem 2 to prove the upper bound. Comparing it to Corollary. 1, we show our low-complexity algorithm is optimal in this setting. It is an important observation that in the homogeneous cost case, this algorithm reduces to a racing algorithm. It is well known that racing algorithms cannot be optimal on a general MAB model. However, we will show that it enjoys surprisingly good empirical performance over a wide range of bandit models with multiple arms in the next section. Establishing a provable suboptimality gap is an interesting future research problem.

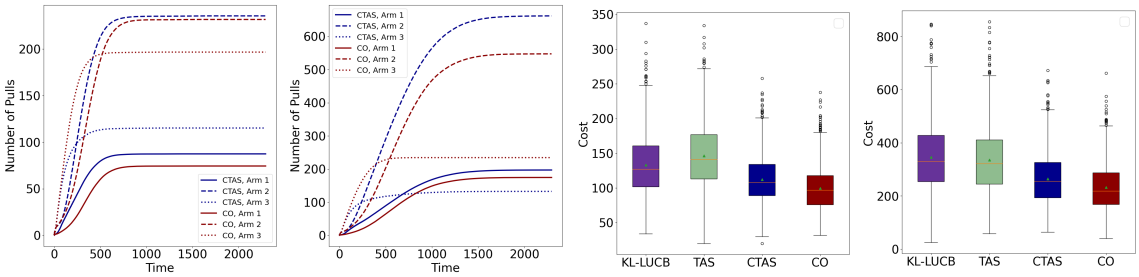
6 Numerical Experiments

As shown before, CO does not inherit the strong theoretical guarantees of CTAS. However, the main appeal of the algorithm comes from both its simplicity and the much more efficient computation time. As shown in Table 2, CO takes significantly less time to run while maintaining good performance.

	CO	CTAS	TAS	d-LUCB
Gaussian	85	1712	2410	82
Bernoulli	58	1995	2780	60
Poisson	96	3260	4633	101

Table 2: The process time (seconds) of each of the algorithms over 1000 trajectories for Gaussian, Bernoulli, and Poisson distributed rewards with $\boldsymbol{\mu} = [1.5, 1.0, 0.5]$ and $\boldsymbol{c} = [1, 0.1, 0.01]$.

Discussion: Our square-root sampling rule of CO comes from reverse-engineering the optimal proportion in the two-armed Gaussian case and then separating the multi-arm problem into pairs of two-armed problems using action elimination. However, an interesting empirical result shown in Fig. 2 is how well it generalizes to other reward distributions. This is illustrated in Figure 2(b). We see that the change in reward distribution does not drastically impact performance. From this, we can deduce that the cost factor of $\sqrt{c}$ generalizes beyond Gaussian distributions. This is partially because when the shrinkage of confidence interval overlap is small, the exponential distribution family is locally Gaussian, and therefore can be approximated by Gaussian bandits. More evidence for this cost factor is shown in Figure 2(a). Here we see that CO is approximately able to match the optimal proportions of arm pulls. The main distinction is that CO is more willing to pull the low-cost arm to eliminate it early on. This results in similar performance because the additional pulls are inexpensive relative to the other arms. Another interesting observation is CO sometimes performs better than CTAS. This is in part because of the elimination rule in CO. While the same proof as CTAS can be utilized to show that CO is δ-PAC, the theory does not utilize the full "tightness" of the CTAS stopping rule. The CO event of error lives in between the event of error for CTAS and the event bounded by theory, causing earlier stopping with less confidence. Empirically, we also had to do more exploration by a constant factor of $\sqrt{t}$ due to the added variance from random costs.



(a) Arm pull trajectories

(b) Total costs.

Figure 2: Results averaged over 1000 trajectories with fixed confidence level $\delta = 10^{-6}$. In (a), we have the average number of arm pulls at each time t. In (b) we have the statistics regarding total cost for these trajectories. This figure was generated with $\boldsymbol{\mu}_1 = [1.5, 1, .5]$ and $\boldsymbol{\mu}_2 = [.9, .6, .3]$ with $\boldsymbol{c} = [1, .1, .01]$, where $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ follow a Bernoulli and Poisson distribution respectively.

Lastly, the TAS family algorithms are very sensitive to good initial starts, meaning that the results are also obfuscated by these extraordinarily long trajectories due to insufficient exploration.

7 Conclusion

In this work, we introduced a new MAB problem: Cost-Aware Best Arm Identification. We provided a new lower bound and an asymptotically optimal cost-adapted BAI algorithm. Finally, we introduced a low-complexity algorithm with promising empirical results. As a future direction, it may be interesting to explore how this algorithm can be adapted to the regret setting in either the cost adapted setting (Sinha et al., 2021), or as an ETC algorithm for carefully chosen costs.

Acknowledgments

This work is supported in part by NSF under grants 2112471, 2134081, 2207548, 2240981, and 2331780.

References

András Antos, Varun Grover, and Csaba Szepesvári. Active learning in multi-armed bandits. In *Algorithmic Learning Theory*, pp. 287–302, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg. ISBN 978-3-540-87987-9.

Jean-Yves Audibert and Sébastien Bubeck. Best Arm Identification in Multi-Armed Bandits. In *COLT - 23th Conference on Learning Theory - 2010*, pp. 13 p., Haifa, Israel, June 2010. URL https://enpc.hal.science/hal-00654404.

Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002. URL https://www.jmlr.org/papers/volume3/auer02a/auer02a.pdf?ref=https://githubhelp.com.

Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. Bandits with knapsacks. *CoRR*, abs/1305.2545, 2013. URL http://arxiv.org/abs/1305.2545.

Antoine Barrier, Aurélien Garivier, and Tomáš Kocák. A non-asymptotic approach to best-arm identification for gaussian bandits. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera (eds.), *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 10078–10109. PMLR, 28–30 Mar 2022. URL https://proceedings.mlr.press/v151/barrier22a.html.

Robert E. Bechhofer. A sequential multiple-decision procedure for selecting the best one of several normal populations with a common unknown variance, and its use with various experimental designs. *Biometrics*, 14(3):408–429, 1958. ISSN 0006341X, 15410420. URL http://www.jstor.org/stable/2527883.

Apostolos N. Burnetas and Michael N. Katehakis. Optimal adaptive policies for sequential allocation problems. *Adv. Appl. Math.*, 17(2):122–142, jun 1996. ISSN 0196-8858. doi: 10.1006/aama.1996.0007. URL https://doi.org/10.1006/aama.1996.0007.

Olivier Cappé , Aurélien Garivier, Odalric-Ambrym Maillard, Rémi Munos, and Gilles Stoltz. Kullback–leibler upper confidence bounds for optimal sequential allocation. *The Annals of Statistics*, 41(3), jun 2013. doi: 10.1214/13-aos1119. URL https://doi.org/10.1214%2F13-aos1119.

Alexandra Carpentier and Andrea Locatelli. Tight (lower) bounds for the fixed budget best arm identification bandit problem. In Vitaly Feldman, Alexander Rakhlin, and Ohad Shamir (eds.), *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pp. 590–604, Columbia University, New York, New York, USA, 23–26 Jun 2016. PMLR. URL https://proceedings.mlr.press/v49/carpentier16.html.

Tianrui Chen, Aditya Gangrade, and Venkatesh Saligrama. Strategies for safe multi-armed bandits with logarithmic regret and risk. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 3123–3148. PMLR, 17–23 Jul 2022a. URL https://proceedings.mlr.press/v162/chen22e.html.

Tianrui Chen, Aditya Gangrade, and Venkatesh Saligrama. Doubly-optimistic play for safe linear bandits. *arXiv preprint arXiv:2209.13694*, 2022b.

Herman Chernoff. Sequential Design of Experiments. *The Annals of Mathematical Statistics*, 30 (3):755 – 770, 1959. doi: 10.1214/aoms/1177706205. URL https://doi.org/10.1214/aoms/1177706205.

Rémy Degenne, Wouter M Koolen, and Pierre Ménard. Non-asymptotic pure exploration by solving games. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019a. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/8d1de7457fa769ece8d93a13a59c8552-Paper.pdf.

Rémy Degenne, Thomas Nedelec, Clement Calauzenes, and Vianney Perchet. Bridging the gap between regret minimization and best arm identification, with application to a/b tests. In Kamalika Chaudhuri and Masashi Sugiyama (eds.), *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 1988–1996. PMLR, 16–18 Apr 2019b. URL https://proceedings.mlr.press/v89/degenne19a.html.

Wenkui Ding, Tao Qin, Xu-Dong Zhang, and Tie-Yan Liu. Multi-armed bandit with budget constraint and variable costs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 27(1): 232–238, Jun. 2013. doi: 10.1609/aaai.v27i1.8637. URL https://ojs.aaai.org/index.php/AAAI/article/view/8637.

Victor Gabillon, Mohammad Ghavamzadeh, and Alessandro Lazaric. Best arm identification: A unified approach to fixed budget and fixed confidence. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper_files/paper/2012/file/8b0d268963dd0cfb808aac48a549829f-Paper.pdf.

Aurélien Garivier, Pierre Ménard, and Gilles Stoltz. Explore first, exploit next: The true shape of regret in bandit problems. *Mathematics of Operations Research*, 44(2):377–399, 2019. doi: 10.1287/moor.2017.0928. URL https://doi.org/10.1287/moor.2017.0928.

Aurélien Garivier and Olivier Cappé. The kl-ucb algorithm for bounded stochastic bandits and beyond. In Sham M. Kakade and Ulrike von Luxburg (eds.), *Proceedings of the 24th Annual Conference on Learning Theory*, volume 19 of *Proceedings of Machine Learning Research*, pp. 359–376, Budapest, Hungary, 09–11 Jun 2011. PMLR. URL https://proceedings.mlr.press/v19/garivier11a.html.

Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In Vitaly Feldman, Alexander Rakhlin, and Ohad Shamir (eds.), *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pp. 998–1027, Columbia University, New York, New York, USA, 23–26 Jun 2016. PMLR. URL https://proceedings.mlr.press/v49/garivier16a.html.

Aurélien Garivier and Emilie Kaufmann. Nonasymptotic sequential tests for overlapping hypotheses applied to near-optimal arm identification in bandit models. *Sequential Analysis*, 40(1):61–96, 2021. doi: 10.1080/07474946.2021.1847965. URL https://doi.org/10.1080/07474946.2021.1847965.

Yunlong Hou, Vincent Y. F. Tan, and Zixin Zhong. Almost optimal variance-constrained best arm identification. *IEEE Transactions on Information Theory*, 69(4):2603–2634, 2023. doi: 10.1109/TIT.2022.3222231.

Marc Jourdan, Degenne Rémy, and Kaufmann Emilie. Dealing with unknown variances in best-arm identification. In Shipra Agrawal and Francesco Orabona (eds.), *Proceedings of The 34th International Conference on Algorithmic Learning Theory*, volume 201 of *Proceedings of Machine Learning Research*, pp. 776–849. PMLR, 20 Feb–23 Feb 2023. URL https://proceedings.mlr.press/v201/jourdan23a.html.

Shivaram Kalyanakrishnan, Ambuj Tewari, Peter Auer, and Peter Stone. Pac subset selection in stochastic multi-armed bandits. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, pp. 227–234, Madison, WI, USA, 2012. Omnipress. ISBN 9781450312851.

Kirthevasan Kandasamy, Gautam Dasarathy, Barnabas Poczos, and Jeff Schneider. The multi-fidelity multi-armed bandit. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/2ba596643cbbbc20318224181fa46b28-Paper.pdf.

Masahiro Kato and Kaito Ariu. The role of contextual information in best arm identification. *arXiv preprint arXiv:2106.14077*, 2021.

Emilie Kaufmann and Shivaram Kalyanakrishnan. Information complexity in bandit subset selection. In Shai Shalev-Shwartz and Ingo Steinwart (eds.), *Proceedings of the 26th Annual Conference on Learning Theory*, volume 30 of *Proceedings of Machine Learning Research*, pp. 228–251, Princeton, NJ, USA, 12–14 Jun 2013. PMLR. URL https://proceedings.mlr.press/v30/Kaufmann13.html.

Emilie Kaufmann and Wouter M. Koolen. Mixture martingales revisited with applications to sequential tests and confidence intervals. *Journal of Machine Learning Research*, 22(246):1–44, 2021. URL http://jmlr.org/papers/v22/18-798.html.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 17(1):1–42, 2016. URL http://jmlr.org/papers/v17/kaufman16a.html.

Pushmeet Kohli, Mahyar Salek, and Greg Stoddard. A fast bandit algorithm for recommendation to users with heterogenous tastes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 27(1):1135–1141, Jun. 2013. doi: 10.1609/aaai.v27i1.8463. URL https://ojs.aaai.org/index.php/AAAI/article/view/8463.

T.L Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985. ISSN 0196-8858. doi: https://doi.org/10.1016/0196-8858(85)90002-8. URL https://www.sciencedirect.com/science/article/pii/0196885885900028.

Tze Leung Lai. Adaptive Treatment Allocation and the Multi-Armed Bandit Problem. *The Annals of Statistics*, 15(3):1091 – 1114, 1987. doi: 10.1214/aos/1176350495. URL https://doi.org/10.1214/aos/1176350495.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Xin Liu, Bin Li, Pengyi Shi, and Lei Ying. Pond: Pessimistic-optimistic online dispatching. *arXiv preprint arXiv:2010.09995*, 2020.

Stefan Magureanu, Richard Combes, and Alexandre Proutiere. Lipschitz bandits: Regret lower bound and optimal algorithms. In Maria Florina Balcan, Vitaly Feldman, and Csaba Szepesvári (eds.), *Proceedings of The 27th Conference on Learning Theory*, volume 35 of *Proceedings of Machine Learning Research*, pp. 975–999, Barcelona, Spain, 13–15 Jun 2014. PMLR. URL https://proceedings.mlr.press/v35/magureanu14.html.

Pierre Ménard. Gradient ascent for active exploration in bandit problems. *arXiv preprint arXiv:1905.08165*, 2019.

Riccardo Poiani, Alberto Maria Metelli, and Marcello Restelli. Multi-fidelity best-arm identification. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 17857–17870. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/71c31ebf577ffdad5f4a74156daad518-Paper-Conference.pdf.

Herbert Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527 – 535, 1952. URL https://community.ams.org/journals/bull/1952-58-05/S0002-9904-1952-09620-8/S0002-9904-1952-09620-8.pdf.

Deeksha Sinha, Karthik Abinav Sankararaman, Abbas Kazerouni, and Vashist Avadhanula. Multi-armed bandits with cost subsidy. In Arindam Banerjee and Kenji Fukumizu (eds.), *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 3016–3024. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/sinha21a.html.

Williams Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 12 1933. ISSN 0006-3444. doi: 10.1093/biomet/25.3-4.285. URL https://doi.org/10.1093/biomet/25.3-4.285.

Sofía S Villar, Jack Bowden, and James Wason. Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 30(2):199, 2015. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4856206/pdf/emss-67661.pdf.

Xuchuang Wang, Qingyun Wu, Wei Chen, and John C.S. Lui. Multi-fidelity multi-armed bandits revisited. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 31570–31600. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/64602b87c31db70a3ef060f6c5d5b01d-Paper-Conference.pdf.

Zhenlin Wang, Andrew J. Wagenmaker, and Kevin Jamieson. Best arm identification with safety constraints. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera (eds.), *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 9114–9146. PMLR, 28–30 Mar 2022. URL https://proceedings.mlr.press/v151/wang22h.html.

Yingce Xia, Wenkui Ding, Xu-Dong Zhang, Nenghai Yu, and Tao Qin. Budgeted bandit problems with continuous random costs. In Geoffrey Holmes and Tie-Yan Liu (eds.), *Asian Conference on Machine Learning*, volume 45 of *Proceedings of Machine Learning Research*, pp. 317–332, Hong Kong, 20–22 Nov 2016. PMLR. URL https://proceedings.mlr.press/v45/Xia15.html.

Ali Yekkehkhany, Ebrahim Arian, Rakesh Nagi, and Ilan Shomorony. A cost–based analysis for risk–averse explore–then–commit finite–time bandits. *IISE Transactions*, 53(10):1094–1108, 2021. doi: 10.1080/24725854.2021.1882014. URL https://doi.org/10.1080/24725854.2021.1882014.

Qining Zhang and Lei Ying. Fast and regret optimal best arm identification: Fundamental limits and low-complexity algorithms. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 16729–16769. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/35fdecdf8861bc15110d48fbec3193cf-Paper-Conference.pdf.