
Balancing Performance and Costs in Best Arm Identification

Michael O. Harding
Department of Statistics
University of Wisconsin-Madison
moharding@wisc.edu

Kirthevasan Kandasamy
Department of Computer Science
University of Wisconsin-Madison
kandasamy@cs.wisc.edu

Abstract

We consider the problem of identifying the best arm in a multi-armed bandit model. Despite a wealth of literature in the traditional fixed budget and fixed confidence regimes of the best arm identification problem, it still remains a mystery to most practitioners as to how to choose an approach and corresponding budget or confidence parameter. We propose a new formalism to avoid this dilemma altogether by minimizing a risk functional which explicitly balances the performance of the recommended arm and the cost incurred by learning this arm. In this framework, a cost is incurred for each observation during the sampling phase, and upon recommending an arm, a performance penalty is incurred for identifying a suboptimal arm. The learner's goal is to minimize the sum of the penalty and cost. This new regime mirrors the priorities of many practitioners, e.g. maximizing profit in an A/B testing framework, better than classical fixed budget or confidence settings. We derive theoretical lower bounds for the risk of each of two choices for the performance penalty, the probability of misidentification and the simple regret, and propose an algorithm called DBCARE to match these lower bounds up to polylog factors on nearly all problem instances. We then demonstrate the performance of DBCARE on a number of simulated models, comparing to fixed budget and confidence algorithms to show the shortfalls of existing BAI paradigms on this problem.

1 Introduction

Best Arm Identification (BAI) in multi-armed bandits is a fundamental problem in decision-making under uncertainty. The objective is to identify the arm with the highest expected reward by adaptively drawing samples from distributions associated with each arm. BAI arises in many real-world applications. In advertising, arms represent different ads, and the aim is to find the ad which maximizes revenue generated [18]. In statistical model selection, arms represent different hyperparameter configurations, and the aim is to find the best-performing one with minimal computational resources [20].

Traditionally, BAI has been studied under two paradigms: the *fixed budget* setting [2, 8], which seeks to maximize performance—i.e. the ability of a policy to recover the optimal arm—within a given sampling budget, and the *fixed performance* (e.g., fixed confidence [33, 15]) setting, which aims to minimize the number of samples needed to meet a target performance level. While algorithms for these settings have been successfully deployed in many real-world settings [32, 35, 51, 18], these settings are not a natural fit for all use cases. For instance, while determining the best arm is desirable, a slightly suboptimal choice may be acceptable if the cost of distinguishing between top candidates is prohibitively high. On the other hand, it is often unnecessary to continue sampling until reaching some pre-specified horizon when there is already enough evidence to determine the optimal arm.

To this end, we propose a novel paradigm for BAI, in which a policy should explicitly balance performance and sampling cost on the fly, without being constrained by a fixed performance level or a pre-specified sampling budget. This framework allows policies to *adaptively* terminate according to the difficulty of the problem. The following is an example where such a framework would be natural.

Example (Advertising). Consider a firm choosing among K versions of an ad. To inform its choice, the firm may show versions to participants in a focus group (arm pull), incurring a cost c per showing. The firm wishes to choose an algorithm to maximize the expected profit, i.e. the expected revenue of the selected ad ($\hat{I}$) minus the expected cost of the sampling procedure: $\mathbb{E}[\text{revenue}_{\hat{I}}] - c \mathbb{E}[\# \text{ arm pulls}]$. Letting I^* be the ad with the highest expected revenue, then maximizing expected profit can be equivalently stated as minimizing $\mathbb{E}[\text{revenue}_{I^*} - \text{revenue}_{\hat{I}}] + c \mathbb{E}[\# \text{ arm pulls}]$. Traditional fixed budget or confidence algorithms would be a poor fit for this problem, as it is unclear how one should choose the budget or confidence level to optimize the objective.

1.1 Model

We will now formally introduce our setting. A learner has access to a MAB model $\nu = \{\nu_a\}_{a \in [K]}$, which consists of K arms, each associated with a probability distribution ν_a . Let $\mu_a = \mathbb{E}_{\nu_a}[X]$ denote the expected reward of arm a . Following common conventions in the BAI literature, we assume without loss of generality that the arms are ordered so that $\mu_1 \geq \mu_2 \geq \dots \geq \mu_K$ (the learner is unaware of this ordering). We will assume that for each arm $a \in [K]$, the distribution ν_a is σ -sub-Gaussian and that $\mu_a \in [0, B]$. The learner is aware of σ and B .

A learner interacts with the bandit model over a sequence of rounds $t = 1, 2, \dots$. On round t , the learner selects an arm $A_t \in [K]$ according to a policy π and observes an independent sample X_t drawn from ν_{A_t} . The choice of A_t may depend on the history $\{(A_s, X_s)\}_{s=1}^{t-1}$ of previous actions and observations. Upon termination, the policy recommends an arm $\hat{I} \in [K]$ as the estimated best arm.

Prior work. Traditionally, BAI has been studied under two main regimes: (1) *Fixed budget*: The learner is allowed at most $T \in \mathbb{N}$ samples and aims to minimize either the *probability of misidentification* [2] $\mathbb{P}(\mu_1 \neq \mu_{\hat{I}})$ or the *simple regret* [8] $\mathbb{E}[\mu_1 - \mu_{\hat{I}}]$, i.e. the expected gap between the optimal and selected arms. (2) *Fixed performance*: The learner must satisfy a specified performance goal while minimizing the number of samples. The most common instantiation is *fixed-confidence* BAI [6, 16], where the probability of misidentification $\mathbb{P}(\mu_1 \neq \mu_{\hat{I}})$ is at most a given goal δ .

This work. Both the fixed-budget and fixed-performance formulations neglect practical situations where one may not have a pre-specified budget or performance goal, but would like to trade-off between performance and sampling cost based on problem difficulty. Motivated by such considerations, we propose a new setting, where the goal is to minimize a risk functional that captures both a performance penalty and the cumulative sampling cost. Choosing either the probability of misidentification or the simple regret as the penalty, we study the following two risk measures:

$$\begin{aligned}\mathcal{R}_{\text{MI}}(\pi, \nu) &:= \mathbb{E}_{\nu, \pi} [\mathbb{1}(\mu_1 \neq \mu_{\hat{I}}) + c\tau] = \mathbb{P}_{\nu, \pi}(\mu_1 \neq \mu_{\hat{I}}) + c \mathbb{E}_{\nu, \pi}[\tau], \\ \mathcal{R}_{\text{SR}}(\pi, \nu) &:= \mathbb{E}_{\nu, \pi} [(\mu_1 - \mu_{\hat{I}}) + c\tau] = \mathbb{E}_{\nu, \pi}[\mu_1 - \mu_{\hat{I}}] + c \mathbb{E}_{\nu, \pi}[\tau].\end{aligned}\tag{1}$$

Here, $c > 0$ is the (known) cost required to collect a sample, relative to the performance penalty, and τ is the stopping time (total number of samples) of the policy π . Moreover, $\mathbb{P}_{\nu, \pi}$ and $\mathbb{E}_{\nu, \pi}$ denote the probability and expectation with respect to all randomness arising from the interaction between the policy π and the bandit model ν .

1.2 Summary of our contributions and results

Novel problem formalism. To the best of our knowledge, we are the first to study this risk-based formalism for BAI which trades off between performance and sampling costs. We design policies for both risk measures in (1), upper bound the risk, and provide nearly matching lower bounds.

Lower bounds. To summarize our lower bounds, let $\Delta_k = \mu_1 - \mu_k$ denote the sub-optimality gap of arm k , and let $H := \sum_{k=2}^K \Delta_k^{-2}$ be a problem complexity parameter [33, 15, 26, 21, 28, 30]. We show that the problem difficulty exhibits a phase transition depending on the magnitude of H and the smallest gap Δ_2 . Specifically, in the case of $\mathcal{R}_{\text{MI}}$, when $H \in \mathcal{O}((\sigma^2 c)^{-1})$, we show that $\mathcal{R}_{\text{MI}} \in \Omega(c\sigma^2 H \log((c\sigma^2 H)^{-1}))$, and otherwise, $\mathcal{R}_{\text{MI}} \in \Omega(1)$. In the case of $\mathcal{R}_{\text{SR}}$, when $H\Delta_2^{-1} \in \mathcal{O}((\sigma^2 c)^{-1})$, we show that $\mathcal{R}_{\text{SR}} \in \Omega(c\sigma^2 H \log(\Delta_2(c\sigma^2 H)^{-1}))$, and otherwise, $\mathcal{R}_{\text{SR}} \in \Omega(\Delta_2)$. This phase transition—absent in classical fixed-confidence or fixed-budget settings—underscores the trade-off between performance and costs inherent to our setting: probabilistically distinguishing sub-Gaussian arms scales inversely with the size of the gaps between them, so with small enough gaps it becomes optimal to simply guess the best arm without incurring the cost of sampling.

Proof ideas. Our proof employs change-of-measure arguments to lower bound the risk associated with any particular algorithm via an auxiliary function of problem parameters and the expected stopping time of the algorithm, $\mathbb{E}_{\nu,\pi}[\tau]$. Crucially, this function is convex in $\mathbb{E}_{\nu,\pi}[\tau]$, and minimizing it with respect to $\mathbb{E}_{\nu,\pi}[\tau]$ yields lower bounds on the performance of *any* algorithm while additionally revealing the phase transition behavior, via the regions where $\mathbb{E}_{\nu,\pi}[\tau] = 0$ is optimal.

Algorithm. We propose DBCARE (Dynamically Budgeted Cost-Adapted Risk-minimizing Elimination) for this setting. DBCARE maintains a subset $S \subset [K]$ of surviving arms and confidence intervals for the mean values of these arms. It takes as input a function $N^* : \mathbb{N} \rightarrow \mathbb{N}$ of the size of S , which determines the maximum number of times each arm in S may be pulled. It proceeds in epochs, where in each epoch, every surviving arm is pulled once. At the end of each epoch, DBCARE eliminates arms that can be confidently identified as suboptimal based on the confidence intervals. If any arms are eliminated, the budget for each surviving arm is updated based on N^* . If the budget of arm pulls is exhausted before there is a clear winner, i.e. only one surviving arm, it recommends the surviving arm with the highest empirical mean. However, if a clear winner emerges before the current budget, it terminates early and recommends this arm.

DBCARE combines ideas from both fixed-budget and fixed-confidence algorithms for BAI. However, unlike fixed budget algorithms, the budget is not given in advance; rather, the total number of times an arm can be pulled is determined by the function N^* which depends on the risk (1), the cost c , and the size of the current surviving set S . Similarly, unlike algorithms for fixed confidence BAI [24, 21], the confidence intervals are carefully chosen based on problem parameters, and not via a prespecified failure probability target δ . This design allows DBCARE to adapt to the problem difficulty with respect to the gaps and cost, while simultaneously ensuring control over the worst-case risk.

Upper bound. We show that the above algorithm, with carefully chosen parameters, matches the lower bounds in almost all regimes. Specifically, for $\mathcal{R}_{\text{MI}}$, our algorithm matches the lower bound up to polylog factors for all values of the complexity parameter H . For $\mathcal{R}_{\text{SR}}$, we similarly match the lower bound up to polylog factors when H is not too large. However, when $H \rightarrow \infty$, our upper bound scales as $\mathcal{O}(\log(K)(K\sigma^2c)^{1/3})$, while the lower bound is $\Omega(\Delta_2)$, leaving an additive gap.

Despite this discrepancy in the $\mathcal{R}_{\text{SR}}$ case, we make two important observations. First, we show that our algorithm is *minimax optimal*; that is, the worst-case risk over all problem instances matches the worst-case lower bound up to logarithmic factors. Second, the lower bound in the large H regime is tight and cannot be improved: a naive guessing algorithm—one that selects an arm without pulling any—achieves the lower bound on certain problem instances in this region. However, such a policy performs poorly when H is small, underscoring the value of our adaptive strategy.

Proof ideas. Our use of an elimination-style procedure allows us to guarantee that we never eliminate the optimal arm with high probability, and also identify precisely when highly suboptimal arms are guaranteed to be eliminated. Then, by choosing $N^*(|S|) \asymp \mathcal{O}((|S|c)^{-1})$ for $\mathcal{R}_{\text{MI}}$ and $N^*(|S|) \asymp \mathcal{O}(\sigma^{2/3}(|S|c)^{-2/3})$ for $\mathcal{R}_{\text{SR}}$, we ensure that DBCARE can both match the worst-case behavior of the lower bound and adapt to easier problem settings where there are relatively few good candidate arms.

Empirical evaluation. We corroborate these theoretical findings in simulations and in a real-world experiment on a drug discovery dataset. We compare to fixed budget and confidence algorithms to show the deficiencies of naive adaptations of existing BAI paradigms on this problem.

1.3 Related work

BAI. The multi-armed bandit (MAB) problem, first introduced by Thompson [45], has become a foundational framework for studying the exploration-exploitation trade-off in sequential decision-making under uncertainty. Within this framework, Best Arm Identification (BAI) focuses on identifying the arm with the highest expected reward [7, 25, 16, 10, 28, 21, 41].

BAI has primarily been studied under two paradigms: the fixed-budget and fixed-performance settings. In the fixed-budget setting, the objective is to minimize the probability of misidentification [2, 29, 30, 11, 4], or alternatively, to minimize the simple regret [7, 8, 52]. In the fixed performance setting, the majority of the literature has focused on achieving a target probability of misidentification (a.k.a fixed confidence BAI) [14, 33, 15, 21, 17, 24, 23]. To the best of our knowledge, there is no prior work on minimizing the number of pulls subject to a performance goal on the simple regret.

Our work builds on the extensive literature in this area. In particular, our algorithm draws inspiration from racing-style methods developed for fixed-confidence BAI [34, 21, 24], while our lower bounds rely on technical lemmas from Kaufmann et al. [30]. Nevertheless, the problem we study departs meaningfully from existing formulations, requiring new conceptual insights and analytical tools.

Cost of arm pulls in MAB. Several works have explored sampling costs in BAI. Xia et al. [49] and Qin et al. [39] study identifying the arm with highest reward-to-cost ratio, assuming both reward and cost are observed per sample, both in fixed-budget and fixed-confidence settings. In contrast, in our setting, once a final arm is selected, only its expected reward—not its sampling cost—remains relevant. Degenne et al. [13] and Yang et al. [50] consider minimizing the cumulative regret [40] while performing BAI, but this approach is not applicable when sampling costs are exogenous to rewards, as we consider in our setting. Kanarios et al. [27] study minimizing cumulative cost (instead of the number of pulls) in a fixed confidence setting, when the learner observes a stochastic cost on each arm pull in addition to the reward. Recent work in multi-Fidelity BAI [37, 48, 38] allows a learner to choose to incur different costs for varying magnitudes of accuracy. The last two problem settings are distinctly different from ours. Finally, some works [3, 43] address costs in the cumulative regret setting, which is also distinct from our focus on BAI.

Bayesian sequential testing in classical statistics. Arrow et al. [1] and Wald and Wolfowitz [47] study Bayesian formulations of sequential binary hypothesis testing problems (e.g., $H_1 : \mu_1 - \mu_2 = \Delta$ vs. $H_2 : \mu_1 - \mu_2 = -\Delta$), where the learner must balance the cost of incorrect decisions against the cost of continued testing. They show that the Bayes-optimal procedure for such problems is the sequential probability ratio test (SPRT) of Wald [46], with optimal thresholds determined by solving complex implicit equations that depend on the specific problem parameters. A number of works [44, 12, 5, 31] have extended this study to the more general composite hypothesis testing framework ($H_1 : \mu_1 - \mu_2 > 0$ vs. $H_2 : \mu_1 - \mu_2 \leq 0$). While there are similarities to our proposed setting, their analyses have been restricted to developing procedures that are only asymptotically Bayes-optimal and only hold in the case of exponential families.

Paper organization. The remainder of this paper is organized as follows. In §2, we study the problem in the 2-arm setting. This new formalism for BAI introduces novel intuitions which are best illustrated in the two arm setting. In §3, we present our algorithm and main results in the K -arm setting. Finally, in §4, we evaluate our methods on simulations and show that it outperforms traditional BAI methods on this problem.

2 Two-Arm Setting

To build intuition for this problem, we first study the $K = 2$ setting. Let $\mathcal{P}(\mathbb{R})$ denote all probability measures on $\mathbb{R}$, and let $G_\sigma = \{\lambda \in \mathcal{P}(\mathbb{R}) : \forall t > 0, \mathbb{P}_\lambda(X - \mathbb{E}_\lambda[X] > t) \leq \exp(-t^2/2\sigma^2)\}$ denote all σ -sub-Gaussian probability distributions. Let $\mathcal{M}$, defined below in (2), denote the class of two-armed bandit models with σ -sub-Gaussian rewards; recall that $\mu_i = \mathbb{E}_{\nu_i}[X]$. For a given gap $\Delta \geq 0$, let $\mathcal{M}_\Delta$, defined below, denote the subclass of models with sub-optimality gap Δ . We have:

$$\mathcal{M} := \{\nu = (\nu_1, \nu_2) : \nu_1, \nu_2 \in G_\sigma; \mu_1, \mu_2 \in [0, B]\}, \quad \mathcal{M}_\Delta := \{\nu \in \mathcal{M} : \mu_1 - \mu_2 = \Delta\}. \quad (2)$$

In §2.1, we begin by studying $\mathcal{R}_{\text{MI}}$ in (1), which uses the probability of misidentification as the performance criterion. In §2.2, we then consider $\mathcal{R}_{\text{SR}}$, which instead uses the simple regret. Unless otherwise stated, all results in this section will be corollaries of more general results in §3.

2.1 Probability of misidentification in the two-arm setting

Lower bound. We begin with a gap-dependent lower bound applicable to any policy on this problem.

Corollary 1.1 (Corollary of Theorem 1, Lower bound on $\mathcal{R}_{\text{MI}}$). *Fix a gap $\Delta > 0$ and the cost c per arm pull. Then, for any policy π , we have*

$$\sup_{\nu \in \mathcal{M}_\Delta} \mathcal{R}_{\text{MI}}(\pi, \nu) \geq \text{LB}_{\text{MI}}(\Delta) := \begin{cases} \frac{\sigma^2 c}{4\Delta^2} \log\left(\frac{\epsilon \Delta^2}{\sigma^2 c}\right), & \text{if } \Delta \geq \sqrt{\sigma^2 c}, \\ 1/4, & \text{if } \Delta < \sqrt{\sigma^2 c}. \end{cases} \quad (3)$$

It is instructive to compare the above result with lower bounds for fixed confidence BAI. As in the fixed confidence setting [30], we observe that for large Δ , the lower bound exhibits a familiar

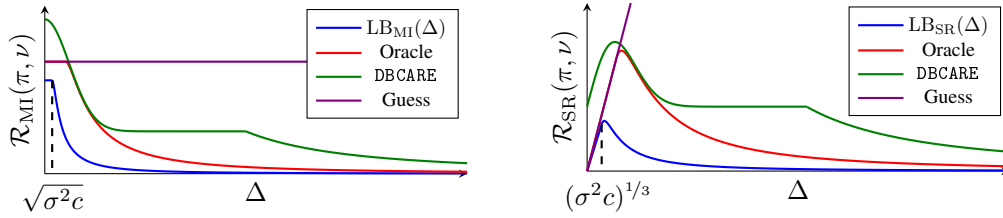


Figure 1: Illustrations of the lower and upper bounds on the risk for $\mathcal{R}_{\text{MI}}$ (on the left) and $\mathcal{R}_{\text{SR}}$ (on the right) in the 2-arm case presented throughout § 2, with the performance of the policy which guesses an arm at random without pulling at all (Guess) included as a point of reference.

dependence on $\sigma^2 \Delta^{-2}$, indicating that the problem becomes easier as the gap increases. Our bound also depends on the cost c and includes a logarithmic term in $\Delta^2 (\sigma^2 c)^{-1}$. Notably, when the gap is small, our setting departs from fixed confidence behavior: the lower bound undergoes a phase transition and saturates at a constant value of $1/4$, rather than continuing to increase with Δ^{-2} .

An oracular policy. To build intuition towards designing a policy, it is worth considering the behavior of an “oracular” policy which knows the gap Δ but does not know which of the two arms is optimal. Recall that it requires approximately $N(\Delta, \delta) \in \mathcal{O}(\sigma^2 \Delta^{-2} \log(1/\delta))$ samples to separate two sub-Gaussian distributions whose means are Δ apart [46, 22, 30] with probability at least $1 - \delta$. Hence, if we pull both arms $N(\Delta, \delta)$ times, we will incur a penalty of $\delta + \mathcal{O}(\sigma^2 c \Delta^{-2} \log(1/\delta))$. By optimally choosing $\delta \in \mathcal{O}(\sigma^2 c \Delta^{-2})$, we find that we need to pull each arm $\mathbb{N}(\Delta) \in \mathcal{O}(\sigma^2 \Delta^{-2} \log(\Delta^2 (\sigma^2 c)^{-1}))$ times. However, the above expression is non-negative only when $\Delta \geq \Omega(\sqrt{\sigma^2 c})$. Intuitively, if Δ is very small, the policy will need to incur a large cost to separate the two arms. If the policy knows Δ it is better off randomly guessing an arm instead of incurring this large cost. This intuition leads to the following policy and theoretical result. Its proof, which is straightforward, is given in Appendix C.

Proposition 1. *Let π_Δ be the policy which pulls each arm $\max \left\{ 0, \left\lceil \frac{4\sigma^2}{\Delta^2} \log \left(\frac{\Delta^2}{8\sigma^2 c} \right) \right\rceil \right\}$ times. If it pulls 0 times, it will choose an arm uniformly at random, and otherwise, outputs the empirically largest arm (breaking ties arbitrarily). Then, letting $\text{LB}_{\text{MI}}(\Delta)$ be as in (3), we have,*

$$\sup_{\nu \in \mathcal{M}_\Delta} \mathcal{R}_{\text{MI}}(\pi_\Delta, \nu) \leq 32\text{LB}_{\text{MI}}(\Delta) + 2c \in \mathcal{O}(\text{LB}_{\text{MI}}(\Delta))$$

As we see, and illustrated in Fig 1, this bound matches the lower bound up to constant factors.¹ To design a policy when Δ is unknown, we will leverage the above intuition. We will also draw inspiration from prior work on racing-style algorithms [36, 34], which have shown that sequentially pulling arms and eliminating them based on confidence intervals can match oracular policies up to logarithmic factors in the fixed confidence setting.

A policy for $\mathcal{R}_{\text{MI}}$. We will let δ be a confidence hyperparameter, aiming to output the optimal arm with probability at least $1 - \delta$. However, to avoid over-pulling when the gap Δ is too small, we also incorporate a hyperparameter N^* , which is a limit on the total amount of times we are willing to pull *each* arm. Intuitively, we know the cost grows linearly in the number of pulls, but the probability of misidentification decays exponentially, so there is a point where the trade-off between the cost of pulling and the increased precision these pulls provide no longer favors continuing to pull.

Our approach proceeds in epochs of sampling both arms once and comparing the difference between the empirical averages of the two arms against a Hoeffding confidence bound at the end of each epoch to test for separation. If the observed difference on any epoch is larger than the confidence bound, it will exit and recommend the larger arm. Otherwise, it will continue to sample each arm until reaching the N^* -th epoch, where it will return the arm with the larger empirical average even though they have not statistically separated. In the case of the 2-arm probability of misidentification setting, we use $N^* = (2ec)^{-1}$ and $\delta = c(1 + 2cN^*)^{-1}$. Here, we set N^* to be the maximum number of times the oracular policy would ever pull each arm for any Δ . The confidence parameter δ is used to control

¹Proposition 1 includes an additive penalty corresponding to the cost of two extra pulls, and a similar additive term appears in all upper bounds. This is unavoidable in general, as even as $\Delta \rightarrow \infty$, each arm must be pulled at least once to identify it. While this can be formally incorporated in the lower bound, we omit it for simplicity.

the penalty of the policy on the event that the policy's confidence interval for the gap does not contain the true gap. We have described this algorithm formally in the K -arm setting in Algorithm 1.

As the corollary below demonstrates, by careful choice of N^* and δ , we show that we can match the lower bound in Corollary 2.1 up to $\log(1/c)$ factors, for all values of Δ . Based on the relationship between algorithmic performance and lower bounds in the BAI literature, we conjecture that this logarithmic gap is largely unavoidable, and could at best be reduced to a log-log factor [28, 21, 30].

Corollary 2.1 (Corollary of Theorem 2, DBCARE under $\mathcal{R}_{\text{MI}}$). *Let π be the policy described above using $N^* = (2ec)^{-1}$ and $\delta = c(1 + 2cN^*)^{-1}$. Then, letting $\text{LB}_{\text{MI}}(\Delta)$ be as in (3),*

$$\sup_{\nu \in \mathcal{M}_\Delta} \mathcal{R}_{\text{MI}}(\pi, \nu) \leq 128 \log \left(\frac{e+1}{(ec)^2} \right) \text{LB}_{\text{MI}}(\Delta) + 3c \in \mathcal{O}(\log(c^{-1})\text{LB}_{\text{MI}}(\Delta)).$$

This bound and its comparison to the lower bound are illustrated in Fig 1. As we can see in Fig 1, by our choice of N^* , our policy actually performs within a constant factor of the lower bound for small Δ , and the $\log(1/c)$ factor is incurred mostly in the “moderate” Δ regime. After the sharp transition at the midpoint of the plot in Fig 1, representing the point at which our algorithm is guaranteed to output the optimal arm before reaching N^* epochs with high probability, we can also see that the comparison to the lower bound quickly improves until we again reach a constant factor mismatch.

2.2 Simple regret in the two-arm setting

Lower bound. We again begin by presenting a lower bound on this problem.

Corollary 3.1 (Corollary of Theorem 3, Lower bound on $\mathcal{R}_{\text{SR}}$). *Fix a gap $\Delta > 0$ and the cost c per arm pull. Then, for any policy π ,*

$$\sup_{\nu \in \mathcal{M}_\Delta} \mathcal{R}_{\text{SR}}(\pi, \nu) \geq \text{LB}_{\text{SR}}(\Delta) = \begin{cases} \frac{\sigma^2 c}{4\Delta^2} \log \left(\frac{e\Delta^3}{\sigma^2 c} \right), & \text{if } \Delta \geq (\sigma^2 c)^{1/3} \\ \Delta/4, & \text{if } \Delta < (\sigma^2 c)^{1/3} \end{cases} \quad (4)$$

Additionally, taking the worst-case over all Δ , we have, for any policy π ,

$$\sup_{\nu \in \mathcal{M}} \mathcal{R}_{\text{SR}}(\pi, \nu) \geq \text{LB}_{\text{SR}}^* = \frac{3}{8} \left(\frac{\sigma^2 c}{e} \right)^{1/3} \quad (5)$$

As in Corollary 1.1, we observe a phase transition in the lower bound: it is $\Delta/4$ when the gap is small, and scales as $\Omega(\Delta^{-2})$ when the gap is large. For what follows, we also state the minimax (worst-case) value of this lower bound as a function of Δ . As we see, this minimax lower bound decreases as the arm-pull cost c decreases. In contrast, for $\mathcal{R}_{\text{MI}}$, the minimax lower bound is $1/4$, and even a naive policy that guesses an arm without any pulls incurs a penalty of only $1/2$. However, for $\mathcal{R}_{\text{SR}}$, even achieving the minimax lower bound requires a well-designed policy.

An oracular policy. To design such a policy, let us again consider the behavior of an oracular policy which knows Δ . The motivation behind the chosen number of samples is the same as before, but when pulling the arms $N(\Delta, \delta)$ times, we now incur a penalty of $\delta\Delta + \mathcal{O}(\sigma^2 c \Delta^{-2} \log(1/\delta))$. Because of this change, we now wish to use $\delta \in \mathcal{O}(\sigma^2 c \Delta^{-3})$, leading to the following result, mirroring that of Proposition 1. Its proof, which is straightforward, is given in Appendix C.

Proposition 2. *Let π^* be the policy which pulls each arm $\max \left\{ 0, \left\lceil \frac{4\sigma^2}{\Delta^2} \log \left(\frac{\Delta^3}{8\sigma^2 c} \right) \right\rceil \right\}$ times. If it pulls them 0 times, it will choose an arm uniformly at random, and otherwise, outputs the empirically largest arm (breaking ties arbitrarily). Then, letting $\text{LB}_{\text{SR}}(\Delta)$ be as in (4) and LB_{SR}^* as in (5),*

$$\sup_{\nu \in \mathcal{M}_\Delta} \mathcal{R}_{\text{SR}}(\pi^*, \nu) \leq 32\text{LB}_{\text{SR}}(\Delta) + 2c \in \mathcal{O}(\text{LB}_{\text{SR}}(\Delta)), \quad \sup_{\nu \in \mathcal{M}} \text{LB}_{\text{SR}}(\pi^*, \nu) \leq 8\text{LB}_{\text{SR}}^* + 2c$$

A policy for $\mathcal{R}_{\text{SR}}$. Our policy will proceed exactly as before, performing rounds of equal sampling until either we reach a prespecified number of epochs or we are able to identify the optimal arm with high probability. Like the oracular policy, though, the change in risk requires updating our hyperparameters N^* and δ to ensure that our algorithm still performs well in this setting. We again motivate our choice of N^* via the behavior of the oracular policy, choosing $N^* = (3/2e)(\sigma/c)^{2/3}$. We also still use δ as a tool to control the worst-case penalty when our confidence interval does not contain the true gap, and thus we set $\delta = c(B + 2cN^*)^{-1}$.

Corollary 4.1 (Corollary of Theorem 4, DBCARE under $\mathcal{R}_{\text{SR}}$). *Let π be the policy described above using $N^* = (3/2e)(\sigma/c)^{2/3}$ and $\delta = c(B + 2cN^*)^{-1}$. Then, letting $\text{LB}_{\text{SR}}(\Delta)$ be as in (4), when $\Delta \geq (\sigma^2 c)^{1/3}$, we have,*

$$\sup_{\nu \in \mathcal{M}_\Delta} \mathcal{R}_{\text{SR}}(\pi, \nu) \leq 128 \log \left(\frac{3B\sigma^{4/3}}{c^{5/3}} \right) \text{LB}_{\text{SR}}(\Delta) + 3c \in \mathcal{O}(\log(B\sigma c^{-1})\text{LB}_{\text{SR}}(\Delta))$$

When $\Delta < (\sigma^2 c)^{1/3}$, we instead have,

$$\sup_{\nu \in \mathcal{M}_\Delta} \mathcal{R}_{\text{SR}}(\pi, \nu) \leq 4\text{LB}_{\text{SR}}(\Delta) + 2(\sigma^2 c)^{1/3} + 3c \in \mathcal{O}(\text{LB}_{\text{SR}}(\Delta) + \text{poly}(\sigma, c))$$

Finally, letting LB_{SR}^* be as in (5), taking the worst case over all Δ , we have,

$$\sup_{\nu \in \mathcal{M}} \mathcal{R}_{\text{SR}}(\pi, \nu) \leq 9\text{LB}_{\text{SR}}^* + 3c \in \mathcal{O}(\text{LB}_{\text{SR}}^*)$$

Here we see, when $\Delta \geq (\sigma^2 c)^{1/3}$, these results closely mirror that of Corollary 2.1, though the log-factor now additionally scales with $B\sigma^2$. As illustrated in Fig 1, this log-factor primarily plays a role in the moderate Δ regime like in the case of $\mathcal{R}_{\text{MI}}$. Our bound and Fig 1 also further highlight the inherent difficulty of designing a simultaneously minimax- and instance-optimal policy for $\mathcal{R}_{\text{SR}}$, as it is impossible to match the lower bound as $\Delta \rightarrow 0$ without performing fewer pulls even as the problem becomes more difficult. Illustrating why the instance-based lower bound cannot be improved in this regime, however, is the policy which guesses an arm without any pulls in purple in Fig 1.

3 K-arm Setting

We now generalize our results to the K -arm setting. We begin by adapting the notation formalities for K arms. We now let $\mathcal{M}$, defined in (6), denote the class of K -armed bandit models with σ -sub-Gaussian rewards. Further, for a bandit model $\nu \in \mathcal{M}$, assuming WLOG that we have $\mu_1 \geq \mu_2 \geq \dots \geq \mu_K$, we define the complexity measure $\mathcal{H}(\nu) := \sum_{k=2}^K \Delta_k^{-2}$, where $\Delta_k = \mu_1 - \mu_k$ is the k -th largest suboptimality gap. For a given complexity $H > 0$, let $\mathcal{M}_H$, defined below, denote the subclass of models having complexity at most H . Thus, we define:

$$\mathcal{M} = \{\nu = (\nu_a)_{a=1}^K : \nu_a \in G_\sigma, \mu_a \in [0, B] \forall a \in [K]\}, \quad \mathcal{M}_H = \{\nu \in \mathcal{M} : \mathcal{H}(\nu) \leq H\} \quad (6)$$

As we will see, while our hardness results extend naturally from two to K arms, extending the intuitions for the algorithm design requires a more careful design of the budget parameter N^* .

3.1 Probability of misidentification in the K-arm setting

Lower bound. We now present the general K -arm lower bound result for $\mathcal{R}_{\text{MI}}$.

Theorem 1. *Fix a complexity $H > 0$ and a cost per arm pull $c > 0$. Then, for any policy π ,*

$$\sup_{\nu \in \mathcal{M}_H} \mathcal{R}_{\text{MI}}(\pi, \nu) \geq \text{LB}_{\text{MI}}(H) = \begin{cases} \frac{\sigma^2 c H}{4} \log \left(\frac{e}{\sigma^2 c H} \right), & \text{if } H \leq (\sigma^2 c)^{-1} \\ 1/4, & \text{if } H > (\sigma^2 c)^{-1} \end{cases} \quad (7)$$

Comparing this result to its Corollary 1.1 in the 2-arm setting, we observe the same phase transition, now in terms of the complexity, H . Using the definition of H , we note that it still occurs when $\Delta_k \asymp \mathcal{O}((\sigma^2 c)^{-1})$, and it provides the same intuition: when at least some of the gaps are sufficiently close to zero (or if there are very many arms), the cost of separating them outweighs the decrease in the probability of misidentification, and it becomes optimal to guess the best arm without pulling.

A policy for $\mathcal{R}_{\text{MI}}$. We present our proposed algorithm, DBCARE, in its full K -arm generality in Algorithm 1. To account for there now being K arms, DBCARE maintains a “surviving set” S of arms that have not yet been determined to be sub-optimal, and performs rounds of equal sampling of all arms in S . At the end of each round, it compares the difference between the current largest empirical average in S and each other arm in S , and eliminates them based on Hoeffding confidence intervals. This continues until either there is only one arm remaining, or the remaining arms have reached their maximum per-arm budget, at which point the arm with the largest empirical average is returned.

In moving from the two arm to K -arm regimes, we once again encounter the issue of balancing performance and costs when selecting our per-arm budget. On one hand, if we naively replace the division by 2 in N^* in Corollary 2.1 with a division by K , then we will fall short on performance when there are many highly sub-optimal arms. However, if we keep the same budget for each arm from the 2-arm setting, we will perform too many total pulls when there are many near-optimal arms.

To this end, we allow the per-arm budgets to *adapt* to the problem complexity by letting N^* increase as $|S|$ decreases. This allows DBCARE to dedicate additional resources to separating the remaining arms as some are determined to be sub-optimal, but prevents the total possible number of pulls from scaling too quickly in K . Inspired by the 2-arm setting, we let $N^*(k) = (k\epsilon c)^{-1}$. Further, we still use the confidence δ to control the worst-case penalty when the confidence intervals do not contain the true gap, so we set $\delta = c(1 + 2c \log(K)N^*(2))^{-1}$. The following theorem summarizes the key properties of DBCARE when applied to $\mathcal{R}_{\text{MI}}$.

Theorem 2. *Let π_{DBCARE} be the policy defined in Algorithm 1 using $N^*(k) = (k\epsilon c)^{-1}$ and $\delta = c(1 + 2c \log(K)N^*(2))^{-1}$. Then, letting $\text{LB}_{\text{MI}}(H)$ be as in (7), we have,*

$$\sup_{\nu \in \mathcal{M}_H} \mathcal{R}_{\text{MI}}(\pi_{\text{DBCARE}}, \nu) \leq 760 \log(K) \log\left(\frac{K \log(K)}{\epsilon c^2}\right) \text{LB}_{\text{MI}}(H) + (K + 1)c,$$

which is $\mathcal{O}(\text{polylog}(K, c^{-1})\text{LB}_{\text{MI}}(H))$.

As in the 2-arm case, we see in Theorem 2 that our policy is still able to achieve performance within a polylogarithmic factor of the lower bound, with the additional $\log(K)$ factor being due to the worst-case impact of our adaptive budget updating.

Algorithm 1 Dynamically Budgeted Cost-Adapted Risk-minimizing Elimination

Require: Dynamic budget function N^* , Confidence δ

```

1: Initialization:  $\hat{\mu}_k(0) = 0 \forall k \in [K], e_0 = 0, t = 0, n = 0, S = [K]$ 
2: while  $n \leq N^*(|S|)$  AND  $|S| > 1$  do
3:    $n \leftarrow n + 1$ 
4:   for  $k \in S$  do
5:      $t \leftarrow t + 1$ 
6:      $A_t \leftarrow k$ , Observe  $X_t \sim \nu_{A_t}$ 
7:   end for
8:    $\hat{\mu}_k(n) \leftarrow \frac{1}{n} \sum_{s=1}^t \mathbb{1}_{\{k\}}(A_s) X_s$ , for  $k \in S$ 
9:    $e_n \leftarrow \sqrt{4\sigma^2 n^{-1} \log(Kn\delta^{-1})}$ .
10:   $S \leftarrow S \setminus \left\{ k \in S : \max_{\ell \in S} \hat{\mu}_\ell(n) - \hat{\mu}_k(n) > e_n \right\}$ .
11: end while
12: return  $\arg\max_{a \in S} \hat{\mu}_a(n)$  (breaking ties randomly)
```

3.2 Simple regret in the K -arm setting

Lower bound. We now present our second lower bound, for $\mathcal{R}_{\text{SR}}$ in the general K -arm setting.

Theorem 3. *Fix a complexity $H > 0$, a smallest gap $\Delta_2 \geq 0$, and a cost per arm pull $c > 0$. Then, for any policy π , we have,*

$$\sup_{\nu \in \mathcal{M}_H} \mathcal{R}_{\text{SR}}(\pi, \nu) \geq \text{LB}_{\text{SR}}(H) = \begin{cases} \frac{c\sigma^2 H}{4} \log\left(\frac{\epsilon \Delta_2}{\sigma^2 c H}\right), & \text{if } H \Delta_2^{-1} \leq (\sigma^2 c)^{-1} \\ \Delta_2/4, & \text{if } H \Delta_2^{-1} > (\sigma^2 c)^{-1} \end{cases} \quad (8)$$

Additionally, taking the worst case over all problem instances, we have, for any policy π ,

$$\sup_{\nu \in \mathcal{M}} \mathcal{R}_{\text{SR}}(\pi, \nu) \geq \text{LB}_{\text{SR}}^* = \frac{3}{8} \left(\frac{(K-1)\sigma^2 c}{e} \right)^{1/3} \quad (9)$$

Looking at the bound presented in (8), we see that the phase transition in this lower bound now jointly involves the total problem complexity and the smallest gap.

A policy for $\mathcal{R}_{\text{SR}}$. Following the same intuition as in the probability of misidentification case, we again wish to allow N^* to adapt to the problem complexity and increase as the surviving set of arms shrinks. Observing the minimax lower bound presented in 9, though, we see that the maximum problem difficulty scales with $K^{1/3}$, unlike the constant scaling in the case of $\mathcal{R}_{\text{MI}}$. Thus, we wish for N^* to scale with $K^{-2/3}$ instead of K^{-1} , and so we choose $N^*(k) = (3/2e)\sigma^{2/3}((k-1)c)^{-2/3}$. Then, controlling for the worst-case performance again, we choose $\delta = c(B + eK^{1/3}\log(K)N^*(2))^{-1}$.

Theorem 4. Let π_{DBCARE} be the policy defined in Algorithm 1 using $N^*(k) = (3/2e)\sigma^{2/3}((k-1)c)^{-2/3}$ and $\delta = c(B + eK^{1/3}\log(K)N^*(2))^{-1}$. Then, letting LB_{SR} be as in (8), when $H\Delta_2^{-1} \leq (\sigma^2c)^{-1}$,

$$\sup_{\nu \in \mathcal{M}_H} \mathcal{R}_{\text{SR}}(\pi_{\text{DBCARE}}, \nu) \leq 550 \log(K) \log\left(\frac{K \log(K) B \sigma^{4/3}}{c^{5/3}}\right) \text{LB}_{\text{SR}}(H) + (K+1)c,$$

which is $\in \mathcal{O}(\text{polylog}(B, K, \sigma, c^{-1})\text{LB}_{\text{SR}}(H))$. When $H\Delta_2^{-1} > (\sigma^2c)^{-1}$, we instead have,

$$\text{LB}_{\text{SR}}(H) + 4 \log(K)(K\sigma^2c)^{1/3} + (K+1)c \in \mathcal{O}(\text{LB}_{\text{SR}}(H) + \log(K)\text{poly}(K, \sigma, c))$$

Finally, letting LB_{SR}^* be as in (9), we have,

$$\sup_{\nu \in \mathcal{M}} \mathcal{R}_{\text{SR}}(\pi_{\text{DBCARE}}, \nu) \leq 20 \log(K) \text{LB}_{\text{SR}}^* + (K+1)c \in \mathcal{O}(\log(K) \text{LB}_{\text{SR}}^*)$$

In Theorem 4, we see similar performance of DBCARE compared to the lower bound as in the two-arm case: we are able to achieve performance within polylogarithmic factors when the complexity is relatively low, and we incur an additive logarithmic and polynomial factor in K , σ , and c when the complexity is prohibitively high. Observing our comparison to the minimax bound, we see that our policy is still minimax-optimal, being only a logarithmic factor in K beyond the lower bound.

4 Numerical Experiments

Simulation Studies. We now empirically compare our method against traditional fixed budget and fixed confidence methods to demonstrate the ability of DBCARE to perform well across all problem instances. We study the performance across a range of suboptimality gaps Δ for Gaussian and Bernoulli rewards in the two-arm setting using the cost $c = 10^{-4}$. In the Gaussian setting, the arms have variance $\sigma^2 = 1$ with means $\pm\Delta/2$, for $\Delta \in [0.05, 2]$; for Bernoulli arms, the means are $0.5 \pm \Delta/2$, for $\Delta \in [0.01, 0.95]$. Results are averaged across 10^5 runs each with different random seeds. We compare to Sequential Halving for fixed budget and elimination procedures using the optimized stopping rules of [30] for fixed confidence. We use budgets $T = 10$ and $T = 500$ and confidences of $\delta = 0.1$ and $\delta = 0.01$ for comparison against relatively low and high confidence/budget choices. We also include the oracular policies of § 2 to provide a baseline of good performance. As we can see in Fig 2, the fixed budget and confidence algorithms necessarily have some region of gaps where they perform sub-optimally: for the small budget, it is moderate Δ values, for the large budget, it is large Δ values, and for both confidences, it is small Δ values. On the contrary, our proposed algorithm exhibits uniformly good performance across all Δ values, which is preferable when Δ is unknown. In Appendix E, we provide further simulations in the K -arm setting.

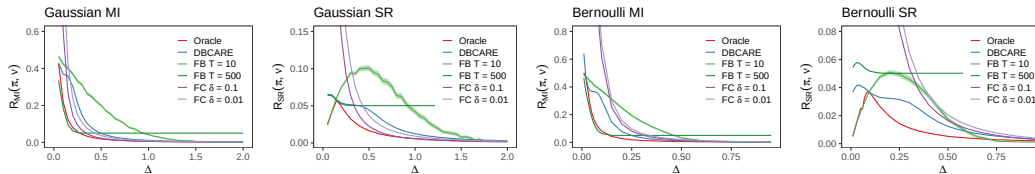


Figure 2: Comparisons between the oracular policy, DBCARE, and fixed budget and confidence algorithms for $\mathcal{R}_{\text{MI}}$ and $\mathcal{R}_{\text{SR}}$. Y-axes are adjusted per setting to highlight problem-specific behavior. Confidence regions represent empirical average risk ± 2 SE.

Drug Discovery Experiment. To demonstrate the efficacy of our approach on a problem in practice, we present the results of a real data experiment on a drug discovery dataset. For this experiment, we take the results from Table 2 of Genovese et al. [19] on the efficacy of the drug secukinumab in patients

with rheumatoid arthritis. They report outcomes for 237 patients assigned to one of 5 treatment groups (arms) and report the drug efficacy according to the American College of Rheumatology criteria ACR20, ACR50, and ACR70. We consider this data under 2 settings: 1.) a binary efficacy outcome, being whether a patient achieved at least ACR20 (1) or not (0), as this was the primary goal of [19]; and 2.) a “leveled” efficacy outcome, where no improvement results in an outcome of 0, and ACR20, ACR50, and ACR70 are outcomes of 0.2, 0.5, and 0.7, respectively, approximating a continuous efficacy metric. We treat the proportions of patients reported in each category in Table 2 of [19] as population proportions, and evaluate DBCARE, Sequential Halving, and an elimination procedure with confidence bounds $\sqrt{4\sigma^2 n^{-1} \log(Kn\delta^{-1})}$ on 10^4 runs in each setting, each with different random seeds. For the binary outcome setting, the means were $\mu = (0.537, 0.469, 0.465, 0.360, 0.340)$, and for the leveled outcome setting, the means were $\mu = (0.230, 0.227, 0.200, 0.196, 0.102)$, each presented in decreasing order (order was randomized during data generation). Because K and the means are fixed in this setting, we choose to evaluate our performance across a range of values for $c \in [10^{-3}, 10^{-5}]$. In Fig 3, we can see that no other method uniformly outperforms DBCARE across all choices of c , again highlighting the ability of our method to adapt to the problem setting at hand.

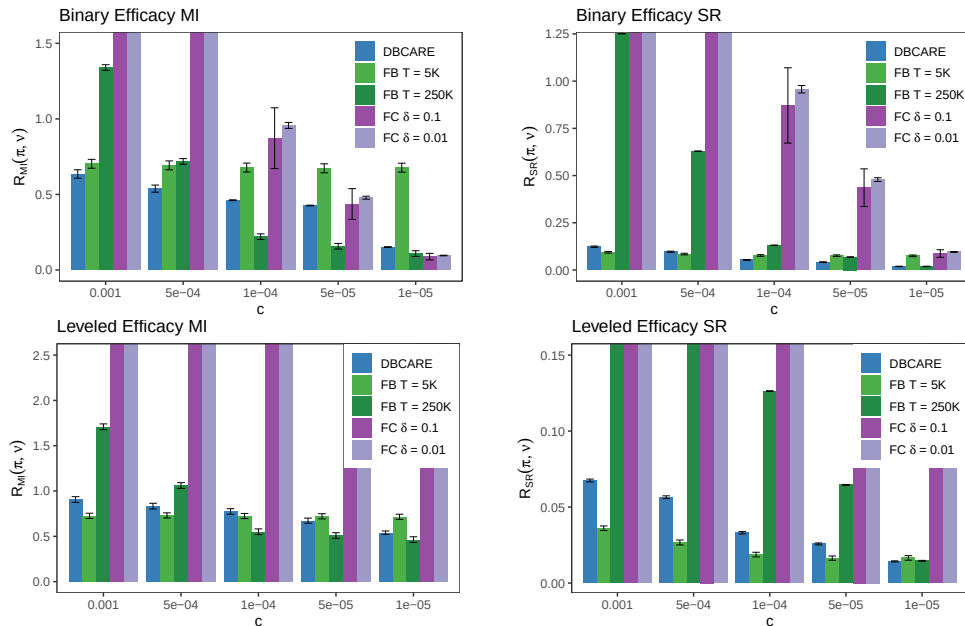


Figure 3: Comparisons between DBCARE and fixed budget and confidence algorithms for $\mathcal{R}_{\text{MI}}$ and $\mathcal{R}_{\text{SR}}$ on a drug discovery dataset. Y-axes are adjusted per setting to highlight problem-specific behavior. Error bars represent empirical average risk ± 2 SE.

5 Conclusion

We propose a novel framework for studying best arm identification. In many practical settings, the traditional fixed budget and confidence regimes do not nicely align with the objectives of practitioners. To fill this gap, our setting explicitly balances sampling costs and performance on the fly by minimizing a risk functional. We prove hardness results for this problem and provide an algorithm, DBCARE, which achieves near-optimal performance on nearly all problem instances.

Future directions. We believe our lower bound analysis for simple regret in the K -arm setting can be improved. Though our bounds are tight when suboptimality gaps are similar, we believe the bounds can be tighter when they are different. We also conjecture that the additive gap we observe in the simple regret setting is unavoidable for algorithms which achieve the minimax risk.

Acknowledgments and Disclosure of Funding

This work was supported in part by NSF Awards IIS-2441796 and DMS-2023239.

References

- [1] K. J. Arrow, D. Blackwell, and M. A. Girshick. Bayes and Minimax Solutions of Sequential Decision Problems. *Econometrica*, 17(3/4):213–244, 1949.
- [2] J.-Y. Audibert, S. Bubeck, and R. Munos. Best arm identification in multi-armed bandits. In *23rd Annual Conference on Learning Theory*, pages 41–53. OmniPress, 2010.
- [3] A. Badanidiyuru, R. Kleinberg, and A. Slivkins. Bandits with Knapsacks. *J. ACM*, 65(3), 2018.
- [4] A. Barrier, A. Garivier, and G. Stoltz. On Best-Arm Identification with a Fixed Budget in Non-Parametric Multi-Armed Bandits. In *Proceedings of The 34th International Conference on Algorithmic Learning Theory*, pages 136–181. PMLR, 2023.
- [5] J. A. Bather and A. M. Walker. Bayes procedures for deciding the sign of a normal mean. *Mathematical Proceedings of the Cambridge Philosophical Society*, 58(4):599–620, 1962.
- [6] R. 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.
- [7] S. Bubeck, R. Munos, and G. Stoltz. Pure exploration in multi-armed bandits problems. In *International Conference on Algorithmic Learning Theory*, pages 23–37. Springer, 2009.
- [8] S. Bubeck, R. Munos, and G. Stoltz. Pure exploration in finitely-armed and continuous-armed bandits. *Theoretical Computer Science*, 412(19):1832–1852, 2011.
- [9] S. Bubeck, V. Perchet, and P. Rigollet. Bounded regret in stochastic multi-armed bandits. In *Proceedings of the 26th Annual Conference on Learning Theory*, pages 122–134. PMLR, 2013.
- [10] S. Bubeck, T. Wang, and N. Viswanathan. Multiple Identifications in Multi-Armed Bandits. In *Proceedings of the 30th International Conference on Machine Learning*, pages 258–265. PMLR, 2013.
- [11] A. Carpentier and A. Locatelli. Tight (Lower) Bounds for the Fixed Budget Best Arm Identification Bandit Problem. In *29th Annual Conference on Learning Theory*, pages 590–604. PMLR, 2016.
- [12] H. Chernoff. Sequential tests for the mean of a normal distribution iv (discrete case). *The Annals of Mathematical Statistics*, 36(1):55–68, 1965.
- [13] R. Degenne, T. Nedelec, C. Calauzenes, and V. Perchet. Bridging the gap between regret minimization and best arm identification, with application to A/B tests. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, pages 1988–1996. PMLR, 2019.
- [14] E. Even-Dar, S. Mannor, and Y. Mansour. PAC bounds for multi-armed bandit and markov decision processes. In *Computational Learning Theory*, pages 255–270. Springer Berlin Heidelberg, 2002.
- [15] 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*, 7(39):1079–1105, 2006.
- [16] V. Gabillon, M. Ghavamzadeh, and A. Lazaric. Best arm identification: A unified approach to fixed budget and fixed confidence. In *Advances in Neural Information Processing Systems*, volume 25, pages 3212–3220. Curran Associates, Inc., 2012.
- [17] A. Garivier and E. Kaufmann. Optimal Best Arm Identification with Fixed Confidence. In *29th Annual Conference on Learning Theory*, pages 998–1027. PMLR, 2016.
- [18] T. Geng, X. Lin, H. S. Nair, J. Hao, B. Xiang, and S. Fan. Comparison lift: Bandit-based experimentation system for online advertising. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 15117–15126, 2021.

- [19] M. C. Genovese, P. Durez, H. B. Richards, J. Supronik, E. Dokoupilova, V. Mazurov, J. A. Aelion, S.-H. Lee, C. E. Coddington, H. Kellner, T. Ikawa, S. Hugot, and S. Mpofu. Efficacy and safety of secukinumab in patients with rheumatoid arthritis: A phase II, dose-finding, double-blind, randomised, placebo controlled study. *Annals of the Rheumatic Diseases*, 72(6): 863–869, 2013.
- [20] K. Jamieson and A. Talwalkar. Non-stochastic Best Arm Identification and Hyperparameter Optimization. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pages 240–248, 2016.
- [21] K. Jamieson, M. Malloy, R. Nowak, and S. Bubeck. lil'ucb : An optimal exploration algorithm for multi-armed bandits. In *27th Annual Conference on Learning Theory*, volume 35, pages 423–439. PMLR, PMLR, 2014.
- [22] C. Jennison, I. M. Johnstone, and B. W. Turnbull. Asymptotically optimal procedures for sequential adaptive selection of the best of several normal means. In *Statistical Decision Theory and Related Topics III*, pages 55–86. Academic Press, 1982.
- [23] M. Jourdan, R. Degenne, D. Baudry, R. de Heide, and E. Kaufmann. Top Two Algorithms Revisited. In *Advances in Neural Information Processing Systems*, volume 35, pages 26791–26803. Curran Associates, Inc., 2022.
- [24] K.-S. Jun, K. Jamieson, R. Nowak, and X. Zhu. Top arm identification in multi-armed bandits with batch arm pulls. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51, pages 139–148. PMLR, PMLR, 2016.
- [25] S. Kalyanakrishnan and P. Stone. Efficient selection of multiple bandit arms: Theory and practice. In *Proceedings of the 27th International Conference on Machine Learning*, pages 511–518. PMLR, 2010.
- [26] S. Kalyanakrishnan, A. Tewari, P. Auer, and P. Stone. Pac subset selection in stochastic multi-armed bandits. In *Proceedings of the 29th International Conference on Machine Learning*, volume 12, page 227–234. Omnipress, 2012.
- [27] K. Kanarios, Q. Zhang, and L. Ying. Cost Aware Best Arm Identification. *Reinforcement Learning Journal*, 4:1533–1545, 2024.
- [28] Z. Karnin, T. Koren, and O. Somekh. Almost Optimal Exploration in Multi-Armed Bandits. In *Proceedings of the 30th International Conference on Machine Learning*, pages 1238–1246. PMLR, 2013.
- [29] E. Kaufmann, R. Korda, and R. Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In *Algorithmic Learning Theory (ALT)*, pages 199–213. Springer Berlin Heidelberg, 2012.
- [30] E. Kaufmann, O. Cappé, and A. Garivier. On the Complexity of Best-Arm Identification in Multi-Armed Bandit Models. *Journal of Machine Learning Research*, 17(1):1–42, 2016.
- [31] T. L. Lai. On optimal stopping problems in sequential hypothesis testing. *Statistica Sinica*, 7(1):33–51, 1997.
- [32] L. Li, K. Jamieson, G. DeSalvo, A. Rostamizadeh, and A. Talwalkar. Hyperband: A novel bandit-based approach to hyperparameter optimization. *Journal of Machine Learning Research*, 18(185):1–52, 2018.
- [33] S. Mannor and J. N. Tsitsiklis. The sample complexity of exploration in the multi-armed bandit problem. *Journal of Machine Learning Research*, 5(Jun):623–648, 2004.
- [34] O. Maron and A. W. Moore. The Racing Algorithm: Model Selection for Lazy Learners. *Artificial Intelligence Review*, 11(1):193–225, 1997.
- [35] U. Misra, R. Liaw, L. Dunlap, R. Bhardwaj, K. Kandasamy, J. E. Gonzalez, I. Stoica, and A. Tumanov. Rubberband: cloud-based hyperparameter tuning. In *Proceedings of the Sixteenth European Conference on Computer Systems*, pages 327–342. Association for Computing Machinery, 2021.

- [36] E. Paulson. A sequential procedure for selecting the population with the largest mean from k normal populations. *The Annals of Mathematical Statistics*, 35(1):174–180, Mar. 1964.
- [37] R. Poiani, A. M. Metelli, and M. Restelli. Multi-Fidelity Best-Arm Identification. In *Advances in Neural Information Processing Systems*, volume 35, pages 17857–17870. Curran Associates, Inc., 2022.
- [38] R. Poiani, R. Degenne, E. Kaufmann, A. M. Metelli, and M. Restelli. Optimal Multi-Fidelity Best-Arm Identification. In *Advances in Neural Information Processing Systems*, volume 37, pages 121882–121927. Curran Associates, Inc., 2024.
- [39] Z. Qin, X. Gan, J. Liu, H. Wu, H. Jin, and L. Fu. Exploring Best Arm with Top Reward-Cost Ratio in Stochastic Bandits. In *IEEE INFOCOM 2020 - IEEE Conference on Computer Communications*, pages 159–168, 2020.
- [40] H. Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535, 1952.
- [41] D. Russo. Simple bayesian algorithms for best arm identification. In *29th Annual Conference on Learning Theory*, volume 49, pages 1417–1418. PMLR, 2016.
- [42] D. Siegmund. *Sequential Analysis*. Springer New York, 1985.
- [43] D. Sinha, K. A. Sankararaman, A. Kazerouni, and V. Avadhanula. Multi-Armed Bandits with Cost Subsidy. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, pages 3016–3024. PMLR, 2021.
- [44] M. Sobel. An essentially complete class of decision functions for certain standard sequential problems. *The Annals of Mathematical Statistics*, pages 319–337, 1953.
- [45] W. R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.
- [46] A. Wald. Sequential Tests of Statistical Hypotheses. *The Annals of Mathematical Statistics*, 16(2):117–186, 1945.
- [47] A. Wald and J. Wolfowitz. Bayes Solutions of Sequential Decision Problems. *The Annals of Mathematical Statistics*, 21(1):82–99, 1950.
- [48] X. Wang, Q. Wu, W. Chen, and J. C. S. Lui. Multi-Fidelity Multi-Armed Bandits Revisited. In *Advances in Neural Information Processing Systems*, volume 36, pages 31570–31600. Curran Associates, Inc., 2023.
- [49] Y. Xia, T. Qin, N. Yu, and T.-Y. Liu. Best action selection in a stochastic environment. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, page 758–766. International Foundation for Autonomous Agents and Multiagent Systems, 2016.
- [50] J. Yang, V. Y. F. Tan, and T. Jin. Best Arm Identification with Minimal Regret. *arXiv preprint arXiv:2409.18909*, 2024.
- [51] J. Zhang, L. Jain, Y. Guo, J. Chen, K. Zhou, S. Suresh, A. Wagenmaker, S. Sievert, T. T. Rogers, K. G. Jamieson, et al. Humor in ai: Massive scale crowd-sourced preferences and benchmarks for cartoon captioning. In *Advances in Neural Information Processing Systems*, volume 37, pages 125264–125286. Curran Associates, Inc., 2024.
- [52] Y. Zhao, C. Stephens, C. Szepesvari, and K.-S. Jun. Revisiting Simple Regret: Fast Rates for Returning a Good Arm. In *Proceedings of the 40th International Conference on Machine Learning*, pages 42110–42158. PMLR, 2023.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: The claims made regarding the K -arm lower bounds are formally stated in § 3 and proved in Appendix B. Our algorithm and theorems regarding its comparison to the lower bounds are presented in § 3 and proven in Appendix D. Our simulation studies and real data experiments are presented in § 4 and Appendix E.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: As we compare the performance of our proposed algorithm to our stated lower bounds in § 2 and § 3, we note where our algorithm falls short and point out what improvements we believe are possible. In § 5, we also discuss where we believe our lower bound analysis is loose and can be improved. When presenting empirical results in § 4 and Appendix E, we explicitly state all of our problem setting choices.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Each theoretical result in § 2 and § 3 state all relevant assumptions, and proofs are provided in Appendices B, C, and D. Necessary lemmas from external works are included in Appendix A and are properly cited.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Our proposed algorithm is clearly written in pseudocode in Algorithm 1, which should be sufficient for implementation. All of our numerical experiments in § 4 and Appendix E have their settings clearly stated, which should be sufficient for replication.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

- (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: We do not provide access to the data and code.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: All of our numerical experiments in § 4 and Appendix E have their settings clearly stated and parameters justified for their relevance to the problem.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: All experimental plots in § 4 and Appendix E include ± 2 SE error bars/regions.

Guidelines:

- The answer NA means that the paper does not include experiments.

- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [\[Yes\]](#)

Justification: Compute resources declarations accompany the additional experiments in Appendix [E](#).

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [\[Yes\]](#)

Justification:

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [\[No\]](#)

Justification: Like any machine learning algorithm, our method could be used by bad actors for explicitly nefarious purposes, but we do not consider potential negative downstream effects of our work. It is our belief that our work does not enable greedy / negligent / nefarious behavior any more than already existing methods in the field.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The real-data experiments in Appendix E are based on published scientific results and are properly cited.

Guidelines:

- The answer NA means that the paper does not use existing assets.

- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.

A Results from Prior Works

Lemma 1 (Lemma 18 of [30]). *Let ν and ν' be two bandit models, and let τ be any stopping time with respect to $\mathcal{F}_t$, where $\mathcal{F}_t = \sigma(A_1, X_1, \dots, A_t, X_t)$ is the sigma-algebra generated by all bandit interactions. For every event $\mathcal{E} \in \mathcal{F}_\tau$ (i.e., $\mathcal{E}$ such that $\mathcal{E} \cap \{\tau = t\} \in \mathcal{F}_t$),*

$$\mathbb{P}_{\nu'}(\mathcal{E}) = \mathbb{E}_\nu[\mathbb{1}_{\mathcal{E}} \exp(-L_\tau)],$$

where,

$$L_\tau = \sum_{a=1}^K \sum_{s=1}^{N_a(\tau)} \log \left(\frac{f_a(Y_{a,s})}{f'_a(Y_{a,s})} \right),$$

where $Y_{a,s}$ is the s -th i.i.d. observation of the a -th arm and f_a and f'_a are the distribution functions of the a -th arm under ν and ν' , respectively.

Lemma 2 (Lemma 4 of [9]). *Let ρ_0, ρ_1 be two probability distributions supported on some set $\mathcal{X}$, with $\rho_1 \ll \rho_0$. Then, for any measurable function $\phi : \mathcal{X} \rightarrow \{0, 1\}$, one has*

$$\mathbb{P}_{X \sim \rho_0}(\phi(X) = 1) + \mathbb{P}_{X \sim \rho_1}(\phi(X) = 0) \geq \frac{1}{2} \exp(-\text{KL}(\rho_0 \parallel \rho_1)).$$

B Proof of Lower Bounds

We begin with a Lemma that will be central to all of our lower bounds, which builds upon the work of [30] to achieve a bound with the form of their Lemma 15 which admits a random stopping time.

Lemma 3. *Let ν and ν' be two K -arm bandit models. Let π be any policy with associated stopping time τ such that $\mathbb{P}(\tau < \infty) = 1$ which outputs an arm $\hat{I} \in [K]$ at time τ . Then for any $a \in [K]$,*

$$\mathbb{P}_{\nu, \pi}(\hat{I} \neq a) + \mathbb{P}_{\nu', \pi}(\hat{I} = a) \geq \frac{1}{2} \exp \left(- \sum_{a=1}^K \mathbb{E}_{\nu, \pi}[N_a(\tau)] \text{KL}(\nu_a \parallel \nu'_a) \right),$$

where $\mathbb{P}_{\nu, \pi}$ is the probability with respect to all randomness incurred by π interacting with the bandit model ν and $N_a(t) = \sum_{s=1}^t \mathbb{1}_{\{a\}}(A_s)$ is the number of times arm a has been pulled up to round t .

Proof. Fix a policy π . For ease of notation, we suppress the π in the probabilities and expectations throughout the proof. Now, we begin by using Lemma 1 to prove that the distributions of $\hat{I}$ under each bandit model are absolutely continuous with one another. Consider an event $\mathcal{E} \in \mathcal{F}_\tau$ such that $\mathbb{P}_{\nu'}(\mathcal{E}) = 0$. Then, because $e^{-x} > 0$ for all $x \in \mathbb{R}$, we must have $\mathbb{P}_\nu(\mathbb{1}_{\mathcal{E}} \exp(-L_\tau) = 0) = 1$ by Lemma 1. Further, by our supposition that $\mathbb{P}_\nu(\tau < \infty) = 1$, we must have $\mathbb{P}_\nu(\exp(-L_\tau) > 0) = 1$, and so we have $\mathbb{P}_{\nu'}(\mathcal{E}) = 0 \implies \mathbb{P}_\nu(\mathcal{E}) = 0$, and we achieve the reverse implication by symmetry. Now, consider the fact that we necessarily have $\{\hat{I} = a\} \in \mathcal{F}_\tau$ for all $a \in [K]$ by construction, and so if we denote by $\mathcal{L}(\hat{I})$ and $\mathcal{L}'(\hat{I})$ the distributions of $\hat{I}$ under ν and ν' , respectively, then clearly $\mathcal{L}'(\hat{I}) \ll \mathcal{L}(\hat{I})$. Thus, we can apply Lemma 2 to show,

$$\mathbb{P}_\nu(\hat{I} \neq a) + \mathbb{P}_{\nu'}(\hat{I} = a) \geq \frac{1}{2} \exp \left(- \text{KL}(\mathcal{L}(\hat{I}) \parallel \mathcal{L}'(\hat{I})) \right)$$

To conclude the proof, we need only upper bound $\text{KL}(\mathcal{L}(\hat{I}) \parallel \mathcal{L}'(\hat{I}))$ by $\mathbb{E}_\nu[L_\tau]$, which we know is equal to $\sum_{a=1}^K \mathbb{E}_\nu[N_a(\tau)] \text{KL}(\nu_a \parallel \nu'_a)$ by an application of Wald's Lemma [42]. By applying the conditional Jensen inequality to the statement of Lemma 1 and rearranging the terms, we know for any $\mathcal{E} \in \mathcal{F}_\tau$, we have $\mathbb{E}_\nu[L_\tau \mid \mathcal{E}] \geq \log \frac{\mathbb{P}_\nu(\mathcal{E})}{\mathbb{P}_{\nu'}(\mathcal{E})}$. Thus, letting $\mathcal{I} = \{k \in [K] : \mathbb{P}_\nu(\hat{I} = k) \neq 0\}$, we can write,

$$\begin{aligned} \mathbb{E}_\nu[L_\tau] &= \sum_{k \in \mathcal{I}} \mathbb{E}_\nu[L_\tau \mid \hat{I} = k] \mathbb{P}_\nu(\hat{I} = k) \\ &\geq \sum_{k \in \mathcal{I}} \log \left(\frac{\mathbb{P}_\nu(\hat{I} = k)}{\mathbb{P}_{\nu' = k}(\hat{I} = k)} \right) \mathbb{P}_\nu(\hat{I} = k) \\ &= \text{KL}(\mathcal{L}(\hat{I}) \parallel \mathcal{L}'(\hat{I})), \end{aligned}$$

which concludes the proof. $\square$

We now employ Lemma 3 to prove Theorems 1 and 3 and their associated corollaries.

Proof of Theorem 1. Fix $H > 0, \sigma^2 > 0, c > 0$, and a policy π . Let ν be a Gaussian K -arm bandit model with means $\mu_1 > \mu_2 \geq \dots \geq \mu_K$ and common variance σ^2 satisfying $\mathcal{H}(\nu) = H$. Then, it is easy to show by contradiction that there must exist some arm $a \in \{2, \dots, K\}$ such that $\mathbb{E}_{\pi, \nu}[N_a(\tau)] \leq \frac{\mathbb{E}_{\pi, \nu}[\tau]}{\Delta_a^2 H(\nu)}$. Let ν' be an alternative model with Gaussian arms having the same common variance σ^2 , where $\mu_k(\nu) = \mu_k(\nu')$ for all $k \neq a$ and $\mu_a(\nu') = \mu_a(\nu) + 2\Delta_a$ so that arm a is now the optimal arm. Clearly $\mathcal{H}(\nu') \leq \mathcal{H}(\nu)$, and so we have $\nu, \nu' \in \mathcal{M}_H$. Then, we can apply Lemma 3 to show,

$$\begin{aligned} \sup_{\nu \in \mathcal{M}_H} \mathcal{R}_{\text{MI}}(\pi, \nu) &\geq \max \{ \mathcal{R}_{\text{MI}}(\pi, \nu), \mathcal{R}_{\text{MI}}(\pi, \nu') \} \\ &\geq \frac{1}{2} (\mathcal{R}_{\text{MI}}(\pi, \nu) + \mathcal{R}_{\text{MI}}(\pi, \nu')) \\ &= \frac{1}{2} (\mathbb{P}_{\nu, \pi}(\hat{I} \neq 1) + \mathbb{P}_{\nu', \pi}(\hat{I} \neq a)) + \frac{c}{2} (\mathbb{E}_{\nu, \pi}[\tau] + \mathbb{E}_{\nu', \pi}[\tau]) \\ &\geq \frac{1}{2} (\mathbb{P}_{\nu, \pi}(\hat{I} \neq 1) + \mathbb{P}_{\nu', \pi}(\hat{I} = 1)) + \frac{c}{2} (\mathbb{E}_{\nu, \pi}[\tau] + \mathbb{E}_{\nu', \pi}[\tau]) \\ &\geq \frac{1}{4} \exp \left(- \sum_{k=1}^K \mathbb{E}_{\nu, \pi}[N_k(\tau)] \text{KL}(\nu_k \parallel \nu'_k) \right) + \frac{c}{2} (\mathbb{E}_{\nu, \pi}[\tau] + \mathbb{E}_{\nu', \pi}[\tau]) \\ &\geq \frac{1}{4} \exp \left(- \frac{2 \mathbb{E}_{\nu, \pi}[\tau]}{\sigma^2 H} \right) + \frac{c}{2} (\mathbb{E}_{\nu, \pi}[\tau] + \mathbb{E}_{\nu', \pi}[\tau]) \end{aligned} \quad (10)$$

Here, we recognize the fact that (10) is convex in $\mathbb{E}_{\nu, \pi}[\tau]$, and thus we provide a π -free lower bound by minimizing over $\mathbb{E}_{\nu, \pi}[\tau], \mathbb{E}_{\nu', \pi}[\tau] \geq 0$, which is achieved by $\mathbb{E}_{\nu, \pi}[\tau] = \max\{0, \frac{\sigma^2 H}{2} \log(1/\sigma^2 cH)\}$ and $\mathbb{E}_{\nu', \pi}[\tau] = 0$. Plugging in these values completes the proof. $\square$

Proof of Theorem 3. This proof proceeds nearly identically to the proof above. Again fix $H > 0, \sigma^2 > 0, c > 0$ and any policy π . Then, let ν and ν' be the same Gaussian K -arm bandit models as in the previous proof, with optimal arms 1 and $a \in \{2, \dots, K\}$, respectively. For notational clarity, let the suboptimality gaps $\Delta_2, \dots, \Delta_K$ be with respect to ν and let $\Delta_1 \equiv 0$, so that $\mu_a(\nu') - \mu_k(\nu') = \Delta_a + \Delta_k$ for $k \neq a$. Then, again using Lemma 3, we have,

$$\begin{aligned} \sup_{\nu \in \mathcal{M}_H} \mathcal{R}_{\text{SR}}(\pi, \nu) &\geq \max \{ \mathcal{R}_{\text{SR}}(\pi, \nu), \mathcal{R}_{\text{SR}}(\pi, \nu') \} \\ &\geq \frac{1}{2} (\mathcal{R}_{\text{SR}}(\pi, \nu) + \mathcal{R}_{\text{SR}}(\pi, \nu')) \\ &= \frac{1}{2} (\mathbb{E}_{\nu, \pi}[\mu_1 - \mu_{\hat{I}}] + \mathbb{E}_{\nu', \pi}[\mu_a - \mu_{\hat{I}}]) + \frac{c}{2} (\mathbb{E}_{\nu, \pi}[\tau] + \mathbb{E}_{\nu', \pi}[\tau]) \\ &= \frac{1}{2} \left(\sum_{i=2}^K \Delta_i \mathbb{P}_{\nu, \pi}(\hat{I} = i) + \sum_{j \neq a} (\Delta_a + \Delta_j) \mathbb{P}_{\nu', \pi}(\hat{I} = j) \right) \\ &\quad + \frac{c}{2} (\mathbb{E}_{\nu, \pi}[\tau] + \mathbb{E}_{\nu', \pi}[\tau]) \\ &\geq \frac{\Delta_2}{2} (\mathbb{P}_{\nu, \pi}(\hat{I} \neq 1) + \mathbb{P}_{\nu', \pi}(\hat{I} \neq a)) + \frac{c}{2} (\mathbb{E}_{\nu, \pi}[\tau] + \mathbb{E}_{\nu', \pi}[\tau]) \\ &\geq \frac{\Delta_2}{2} (\mathbb{P}_{\nu, \pi}(\hat{I} \neq 1) + \mathbb{P}_{\nu', \pi}(\hat{I} = 1)) + \frac{c}{2} (\mathbb{E}_{\nu, \pi}[\tau] + \mathbb{E}_{\nu', \pi}[\tau]) \\ &\geq \frac{\Delta_2}{4} \exp \left(- \sum_{k=1}^K \mathbb{E}_{\nu, \pi}[N_k(\tau)] \text{KL}(\nu_k \parallel \nu'_k) \right) + \frac{c}{2} (\mathbb{E}_{\nu, \pi}[\tau] + \mathbb{E}_{\nu', \pi}[\tau]) \\ &\geq \frac{\Delta_2}{4} \exp \left(- \frac{2 \mathbb{E}_{\nu, \pi}[\tau]}{\sigma^2 H} \right) + \frac{c}{2} (\mathbb{E}_{\nu, \pi}[\tau] + \mathbb{E}_{\nu', \pi}[\tau]) \end{aligned} \quad (11)$$

Once again, (11) is convex in $\mathbb{E}_{\nu, \pi}[\tau]$, so we can further lower bound (11) by setting $\mathbb{E}_{\nu, \pi}[\tau] = \max\{0, \frac{\sigma^2 H}{2} \log(\Delta_2(\sigma^2 cH)^{-1})\}$ and $\mathbb{E}_{\nu', \pi}[\tau] = 0$, completing the proof of (8).

To prove (9), we can consider a specific set of means satisfying this problem instance. Consider the instance where $\mu_1 = \Delta$ and $\mu_2 = \dots = \mu_K = 0$, so that $\Delta_2 = \dots = \Delta_K = \Delta$ for some $\Delta > 0$ that we will specify later. With these means, we have $H = (K-1)\Delta^{-2}$. Then, using (8), we can write,

$$\begin{aligned} \sup_{\nu \in \mathcal{M}_{(K-1)\Delta^{-2}}} \mathcal{R}_{\text{SR}}(\pi, \nu) &\geq \text{LB}_{\text{SR}}((K-1)\Delta^{-2}) \\ &= \begin{cases} \frac{(K-1)\sigma^2 c}{4\Delta^2} \log\left(\frac{e\Delta^3}{(K-1)\sigma^2 c}\right), & \text{if } \Delta \geq ((K-1)\sigma^2 c)^{1/3} \\ \Delta/4, & \text{if } \Delta < ((K-1)\sigma^2 c)^{1/3} \end{cases} \end{aligned}$$

We can then find the Δ which maximizes this function, which occurs at $\Delta^* = (\sqrt{e}(K-1)\sigma^2 c)^{1/3}$, which gives,

$$\sup_{\nu \in \mathcal{M}} \mathcal{R}_{\text{SR}}(\pi, \nu) \geq \sup_{\nu \in \mathcal{M}_{(K-1)(\Delta^*)^{-2}}} \mathcal{R}_{\text{SR}}(\pi, \nu) \geq \text{LB}_{\text{SR}}((K-1)(\Delta^*)^{-2}) = \frac{3}{8} \left(\frac{(K-1)\sigma^2 c}{e} \right)^{1/3}$$

□

Proof of Corollaries 1.1 and 3.1. Recall that, when $K = 2$, $H = \Delta^{-2}$ and $\Delta_2 = \Delta$. The conclusions then follow directly from Theorems 1 and 3. □

C Oracular Policy Proofs

Proof of Proposition 1. Fix a gap $\Delta > 0$. Because samples from each arm are i.i.d. σ -sub-Gaussian, by equally sampling the arms, we have i.i.d. $\sqrt{2}\sigma$ -sub-Gaussian observations of the gap Δ . By a Hoeffding confidence bound, if π_T pulls each arm a fixed number of times $\lceil T \rceil$ and outputs the empirically largest arm, then we have

$$\mathcal{R}_{\text{MI}}(\pi_T, \nu) = \mathbb{P}(\hat{\Delta}_{\lceil T \rceil} < 0) + 2c \lceil T \rceil \leq \exp\left(-\frac{T\Delta^2}{4\sigma^2}\right) + 2c(T+1)$$

Plugging in the proposed number of pulls, we get $\mathcal{R}_{\text{MI}}(\pi_\Delta, \nu) \leq \frac{8\sigma^2 c}{\Delta^2} \log\left(\frac{e\Delta^2}{8\sigma^2 c}\right) + 2c$ when $\Delta \geq \sqrt{8\sigma^2 c}$, and exactly $\mathcal{R}_{\text{MI}}(\pi_\Delta, \nu) = 1/2$ otherwise, as then the policy guesses the optimal arm uniformly at random. Multiplying (3) by 32 and adding $2c$ then clearly upper bounds this quantity. □

Proof of Proposition 2. This proof proceeds nearly identically to the previous. Again fix a gap $\Delta > 0$, and consider that we can write,

$$\mathcal{R}_{\text{SR}}(\pi_T, \nu) = \Delta \mathbb{P}(\hat{\Delta}_{\lceil T \rceil} < 0) + 2c \lceil T \rceil \leq \Delta \exp\left(-\frac{T\Delta^2}{4\sigma^2}\right) + 2c(T+1)$$

Then, plugging in the proposed number of pulls, we get $\mathcal{R}_{\text{SR}}(\pi^*, \nu) \leq \frac{8\sigma^2 c}{\Delta^2} \log\left(\frac{e\Delta^3}{8\sigma^2 c}\right) + 2c$ when $\Delta \geq (8\sigma^2 c)^{1/3}$ and exactly $\mathcal{R}_{\text{SR}}(\pi^*, \nu) = \Delta/2$ otherwise, as then the policy guesses the optimal arm uniformly at random. Multiplying (4) by 32 and adding $2c$ then clearly upper bounds this quantity. Further, maximizing this upper bound in terms of Δ (occurring at $\Delta = (8\sqrt{e}\sigma^2 c)^{1/3}$) yields,

$$\sup_{\nu \in \mathcal{M}} \mathcal{R}_{\text{SR}}(\pi^*, \nu) = \sup_{\Delta} \sup_{\nu \in \mathcal{M}_\Delta} \mathcal{R}_{\text{SR}}(\pi^*, \nu) \leq 3 \left(\frac{\sigma^2 c}{e} \right)^{1/3} + 2c = 8\text{LB}_{\text{SR}}^* + 2c$$

□

D Upper Bounds for DBCARE

We begin by presenting a number of technical lemmas allowing us to control the behavior of DBCARE and prove our desired upper bounds on its performance.

Lemma 4 (Bound on total number of pulls). *For any bandit instance ν , using $N^*(k) = (k\epsilon c)^{-1}$ and $N^*(k) = (3/2e)\sigma^{2/3}((k-1)c)^{-2/3}$, DBCARE satisfies,*

$$\mathbb{E}_{\nu,\pi}[\tau] \leq \frac{2\log(K)}{\epsilon c}, \quad \mathbb{E}_{\nu,\pi}[\tau] \leq \frac{3\log(K)(K\sigma^2)^{1/3}}{2c^{2/3}},$$

respectively.

Proof. Let $\hat{k}$ denote the index of the k -th arm eliminated by the algorithm. Then by construction, $\mathbb{E}_{\nu,\pi}[N_{\hat{k}}(\tau)] \leq N^*(K-k+1)$. Further, $\mathbb{E}_{\nu,\pi}[N_{\hat{1}}(\tau)] \leq N^*(2)$. Thus,

$$\mathbb{E}_{\nu,\pi}[\tau] = \sum_{a=1}^K \mathbb{E}_{\nu,\pi}[N_a(\tau)] \leq N^*(2) + \sum_{k=2}^K N^*(k)$$

Then, apply the fact that $1/2 + \sum_{k=2}^K k^{-1} \leq 2\log(K)$ and $1 + \sum_{k=2}^K (k-1)^{-2/3} \leq eK^{1/3}\log(K)$ to prove the statements. $\square$

Lemma 5 (Elimination behavior). *Consider a bandit instance ν satisfying, WLOG, $\mu_1 \geq \mu_2 \geq \dots \geq \mu_K$. Let $n(t)$ be the epoch associated with time t . Define the good event,*

$$G = \bigcap_{n(t) \leq n(\tau)} \bigcap_{k \in S \setminus \{1\}} \{\Delta_k \in (\hat{\mu}_1(n(t)) - \hat{\mu}_k(n(t)) - e_{n(t)}, \hat{\mu}_1(n(t)) - \hat{\mu}_k(n(t)) + e_{n(t)})\}.$$

Then,

1. $\mathbb{P}_{\nu,\pi}(G^c) \leq \delta$
2. On G , $1 \in S \forall n(t) \leq n(\tau)$ (i.e. the optimal arm is never eliminated)
3. On G , if $\Delta_k > \sqrt{\frac{16\sigma^2 \log(KN^*(k)/\delta)}{N^*(k)}}$ for all $k \geq \ell \in \{2, \dots, K\}$ and N^* decreasing in k ,

$$N_k(\tau) \leq \frac{16\sigma^2 \log(KN^*(k)/\delta)}{\Delta_k^2} < N^*(k) \forall k \geq \ell$$

Proof.

Part 1. Letting $Y_{a,s}$ denote the s -th i.i.d. observation from arm a , by assumption, $Y_{1,s} - Y_{k,s}$ are $\sqrt{2}\sigma$ -sub-Gaussian random variables with mean Δ_k . Thus, $\sqrt{\frac{4\sigma^2 \log(n/\delta)}{n}}$ is a δ -correct confidence interval width for Δ_k after n observations using $\hat{\Delta}_{k,n} = \hat{\mu}_1(n) - \hat{\mu}_k(n)$ as the point estimate [21, 30]. Replacing δ by δ/K and taking a union bound across all $k \in S \setminus \{1\}$ then proves 1. **Part 2.** Consider that on G , $\hat{\mu}_k(n) - \hat{\mu}_1(n) \leq e_n - \Delta_k \leq e_n$ for all $k \neq 1$, which proves 2. **Part 3.** We begin with arm K . By the supposition, on G , there exists $n < N^*(K)$ such that $\hat{\mu}_1(n) - \hat{\mu}_K(n) - e_n \geq \Delta_K - 2e_n > 0$, and thus $K \notin S$ for all $m > n$. Further, we can upper bound the n at which this is true by $\frac{16\sigma^2 \log(KN^*(K)/\delta)}{\Delta_K^2}$ by construction of e_n , and this quantity less than $N^*(K)$ by the supposition. Then, because $K \notin S$ at time $N^*(K)$, if N^* is decreasing in k , the algorithm will not be forced to terminate at time $N^*(K)$ by number of epochs, only if all arms other than 1 have already been eliminated, under which the statement would hold anyway. We can then use the same construction for each $k = K-1, \dots, \ell$, proving 3. $\square$

Lemma 6 (Bound on probability of misidentification on the good event). *For any bandit instance ν satisfying $\mu_1 \geq \mu_2 \geq \dots \geq \mu_K$, and N^* decreasing in k , if $M \in \{2, \dots, K\}$ is the smallest value such that for each $k = M+1, \dots, K$, $\Delta_k > \sqrt{\frac{16\sigma^2 \log(KN^*(k)/\delta)}{N^*(k)}}$ (if no Δ_k satisfy this, $M = K$), then $\mathbb{P}_{\nu,\pi}(\{\hat{I} = j\} \cap G) = 0$ if $j > M$ and $\mathbb{P}_{\nu,\pi}(\{\hat{I} = j\} \cap G) \leq \exp(-\frac{N^*(M)\Delta_j^2}{4\sigma^2})$ otherwise.*

Proof. We begin with the case $j > M$. By Lemma 5, arm 1 is never eliminated by the algorithm and arms $j, j+1, \dots, K$ are eliminated before round $N^*(j)$. Then, on G , DBCARE only terminates

because either $S = \{1\}$ or $n(\tau) = N^*(|S|) \geq N^*(j)$, making $\mathbb{P}_{\nu,\pi}(\{\hat{I} = j\} \cap G) = 0$. Now consider $j \leq M$. Then,

$$\begin{aligned}\mathbb{P}_{\nu,\pi}(\{\hat{I} = j\} \cap G) &= \mathbb{P}_{\nu,\pi}(\{\hat{I} = j\} \cap G \cap \{j \in S \text{ at } \tau\}) + \mathbb{P}_{\nu,\pi}(\{\hat{I} = j\} \cap G \cap \{j \notin S \text{ at } \tau\}) \\ &= \mathbb{P}_{\nu,\pi}(\{\hat{I} = j\} \cap G \cap \{j \in S \text{ at } \tau\}) \\ &= \mathbb{P}_{\nu,\pi}(\{\hat{\mu}_j(n(\tau)) > \hat{\mu}_1(n(\tau))\} \cap G \cap \{j \in S \text{ at } \tau\}) \\ &\leq \mathbb{P}_{\nu,\pi}(\{\hat{\mu}_j(N^*(M)) > \hat{\mu}_1(N^*(M))\} \cap G \cap \{j \in S \text{ at } \tau\}) \\ &= \mathbb{P}_{\nu,\pi} \left(\left\{ \frac{1}{N^*(M)} \sum_{n=1}^{N^*(M)} (Y_{j,n} - Y_{1,n}) > 0 \right\} \cap G \cap \{j \in S \text{ at } \tau\} \right) \\ &\leq \mathbb{P}_{\nu,\pi} \left(\frac{1}{N^*(M)} \sum_{n=1}^{N^*(M)} (Y_{j,n} - Y_{1,n}) > 0 \right) \\ &\leq \exp \left(-\frac{N^*(M)\Delta_j^2}{4\sigma^2} \right)\end{aligned}$$

□

Lemma 7 (Bound on simple regret on the good event). *For any bandit instance ν satisfying $\mu_1 \geq \mu_2 \geq \dots \geq \mu_K$, and N^* decreasing in k , if $M \in \{2, \dots, K\}$ is the smallest value such that for each $k = M + 1, \dots, K$, $\Delta_k > \sqrt{\frac{16\sigma^2 \log(KN^*(k)/\delta)}{N^*(k)}}$ (if no Δ_k satisfy this, $M = K$), then,*

$$\mathbb{E}_{\nu,\pi}[(\mu_1 - \mu_{\hat{I}}) \mathbb{1}_G] \leq \sqrt{\frac{4\sigma^2}{\sqrt{e}N^*(M)}}$$

Proof. We begin by relating the simple regret with the probability of misidentification by applying Lemma 6,

$$\begin{aligned}\mathbb{E}_{\nu,\pi}[(\mu_1 - \mu_{\hat{I}}) \mathbb{1}_G] &= \sum_{i=2}^K \Delta_i \mathbb{P}_{\nu,\pi}(\{\hat{I} = i\} \cap G) \\ &\leq \sum_{k=2}^M \Delta_k \mathbb{P}_{\nu,\pi} \left(\bigcap_{\ell=1}^{k-1} \{\hat{\mu}_k(n(\tau)) > \hat{\mu}_\ell(n(\tau))\} \cap G \right)\end{aligned}$$

Now, consider that for $k > \ell \geq 2$, $\mathbb{P}_{\nu,\pi}(\{\hat{\mu}_k(n(\tau)) > \hat{\mu}_\ell(n(\tau))\} \cap G)$ is maximized when $\mu_k = \mu_\ell$ and is equal to $1/2$ when this is the case. Thus, again applying Lemma 6, we can write,

$$\begin{aligned}\mathbb{E}_{\nu,\pi}[(\mu_1 - \mu_{\hat{I}}) \mathbb{1}_G] &\leq \Delta_2 \sum_{k=2}^M \frac{\mathbb{P}_{\nu,\pi}(\{\hat{\mu}_2(n(\tau)) > \hat{\mu}_1(n(\tau))\} \cap G)}{2^{k-1}} \\ &\leq 2\Delta_2 \mathbb{P}_{\nu,\pi}(\{\hat{\mu}_2(n(\tau)) > \hat{\mu}_1(n(\tau))\} \cap G) \\ &\leq 2\Delta_2 \exp \left(-\frac{N^*(M)\Delta_2^2}{4\sigma^2} \right)\end{aligned}$$

Maximizing in terms of Δ_2 then proves the statement. □

With this collection of technical lemmas providing control on the behavior of DBCARE, we are ready to prove Theorems 2 and 4.

Proof of Theorem 2. We break this proof into two cases. First, consider problems of complexity $H \leq (\sigma^2 c)^{-1}$ with $\mu_1 \geq \mu_2 \geq \dots \geq \mu_K$. Further, let $M \in \{1, \dots, K\}$ be the smallest value such that for each $k = M + 1, \dots, K$, $\Delta_k > \sqrt{16ek\sigma^2 c \log(KN^*(k)/\delta)}$ (if no Δ_k satisfy this, $M = K$). Then, by the definition of H , we can write

$$\text{LB}_{\text{MI}}(H) = \frac{\sigma^2 c H}{4} \log \left(\frac{e}{\sigma^2 c H} \right) \geq \frac{M-1}{64eM \log(KN^*(M)/\delta)} + \frac{\sigma^2 c}{4} \sum_{k=M+1}^K \frac{1}{\Delta_k^2} \quad (12)$$

Now, if $M \geq 2$, we apply Lemmas 4 and 5 to show the following:

$$\begin{aligned} \mathbb{P}_{\nu,\pi}(\{\hat{I} \neq 1\} \cap G) + c \mathbb{E}_{\nu,\pi}[\tau \mathbb{1}_G] &= \mathbb{P}_{\nu,\pi}(\{\hat{I} \neq 1\} \cap G) + c \sum_{k=1}^K \mathbb{E}_{\nu,\pi}[N_k(\tau) \mathbb{1}_G] \\ &\leq 1 + \frac{2 \log(M)}{e} + 16\sigma^2 c \sum_{k=M+1}^K \frac{\log(KN^*(k)/\delta)}{\Delta_k^2} \end{aligned} \quad (13)$$

If, in fact, $M = 1$, then combining the results of Lemmas 4, 5, and 6, we can write

$$\mathbb{P}_{\nu,\pi}(\{\hat{I} \neq 1\} \cap G) + c \mathbb{E}_{\nu,\pi}[\tau \mathbb{1}_G] \leq 16\sigma^2 c \sum_{k=M+1}^K \frac{\log(KN^*(k)/\delta)}{\Delta_k^2} \quad (14)$$

Then, multiplying (12) by $760 \log(K) \log(K \log(K)/ec^2)$ and adding Kc (to account for non-integer pulls) then upper bounds both (13) and (14). Now, consider the case where $H > (\sigma^2 c)^{-1}$. Then $\text{LB}_{\text{MI}}(H) = 1/4$. Directly applying Lemma 4 gives us for all H ,

$$\mathbb{P}_{\nu,\pi}(\{\hat{I} \neq 1\} \cap G) + c \mathbb{E}_{\nu,\pi}[\tau \mathbb{1}_G] \leq 1 + \frac{2 \log(K)}{e} \leq 760 \log(K) \log\left(\frac{K \log(K)}{ec^2}\right) \left(\frac{1}{4}\right)$$

Finally, consider that by our choice of δ , using Lemmas 4 and 5, regardless of the value of H , we have

$$\mathbb{P}_{\nu,\pi}(G^c) \left(\mathbb{P}_{\nu,\pi}(\hat{I} \neq 1 \mid G^c) + c \mathbb{E}_{\nu,\pi}[\tau \mid G^c] \right) \leq \delta \left(1 + \frac{2 \log(K)}{e} \right) \leq c$$

This then proves the statement. $\square$

Proof of Theorem 4. This proof largely mirrors that of 2. Again, first consider problems satisfying $H\Delta_2^{-1} \leq (\sigma^2 c)^{-1}$ with $\mu_1 \geq \mu_2 \geq \dots \geq \mu_K$, and let $M \in \{1, \dots, K\}$ be the smallest value such that for each $k = M + 1, \dots, K$, $\Delta_k > \sqrt{(32e/3)((k-1)\sigma^2 c)^{2/3} \log(KN^*(k)/\delta)}$ (if no Δ_k satisfy this, $M = K$). Then,

$$\text{LB}_{\text{SR}}(H) \geq \frac{3(M-1)^{1/3}(\sigma^2 c)^{1/3}}{128e \log(KN^*(M)\delta^{-1})} + \frac{\sigma^2 c}{4} \sum_{k=M+1}^K \frac{1}{\Delta_k^2} \quad (15)$$

If $M \geq 2$, we apply Lemmas 4, 5, and 7 to show

$$\begin{aligned} \mathbb{E}_{\nu,\pi}[(\mu_1 - \mu_{\hat{I}}) \mathbb{1}_G] + c \mathbb{E}_{\nu,\pi}[\tau \mathbb{1}_G] &\leq \sqrt{\frac{8\sqrt{e}}{3}}((M-1)\sigma^2 c)^{1/3} + \frac{3 \log(M)}{2}(M\sigma^2 c)^{1/3} \\ &\quad + 16\sigma^2 c \sum_{k=2}^K \frac{\log(KN^*(k)\delta^{-1})}{\Delta_k^2} \end{aligned} \quad (16)$$

Additionally, if $M = 1$, then,

$$\mathbb{E}_{\nu,\pi}[(\mu_1 - \mu_{\hat{I}}) \mathbb{1}_G] + c \mathbb{E}_{\nu,\pi}[\tau \mathbb{1}_G] \leq 32\sigma^2 c \sum_{k=2}^K \frac{\log(KN^*(k)\delta^{-1})}{\Delta_k^2} \quad (17)$$

Then, multiplying (15) by $575 \log(K) \log(K \log(K) B \sigma^{5/3} c^{-4/3})$ upper bounds both (16) and (17). Now, for the case where $H\Delta_2^{-1} > (\sigma^2 c)^{-1}$ and for the worst-case comparison, we apply Lemmas 4 and 7 to show for all H ,

$$\mathbb{E}_{\nu,\pi}[(\mu_1 - \mu_{\hat{I}}) \mathbb{1}_G] + c \mathbb{E}_{\nu,\pi}[\tau \mathbb{1}_G] \leq \sqrt{\frac{8\sqrt{e}}{3}}((K-1)\sigma^2 c)^{1/3} + \frac{3 \log(K)}{2}(K\sigma^2 c)^{1/3} \quad (18)$$

We then have (18) upper bounded by $4 \log(K)(K\sigma^2 c)^{1/3}$, and $\text{LB}_{\text{SR}}(H) \geq 0$. We can also upper bound (18) by $20 \log(K) \text{LB}_{\text{SR}}^*$. Finally, we never incur more than an additional Kc risk due to integer pulls, and by choice of δ ,

$$\mathbb{P}_{\nu,\pi}(G^c) (\mathbb{E}_{\nu,\pi}(\mu_1 - \mu_{\hat{I}} \mid G^c) + c \mathbb{E}_{\nu,\pi}[\tau \mid G^c]) \leq \delta \left(B + \frac{3c \log(K) \sigma^{2/3}}{ec^{2/3}} \right) \leq c$$

which proves all statements. $\square$

Now, despite our 2-arm results being corollaries of their more general K -arm counterparts, we are able to provide tighter constants in Corollaries 2.1 and 4.1 by utilizing some more precise techniques that are not generally applicable in the K -arm case. For both cases, we apply Lemma 5 in the 2-arm case to identify a Δ^* such that, for all $\Delta > \Delta^*$, the algorithm is guaranteed to identify the optimal arm before reaching N^* samples per arm on the good event G . We then show that we simply need to find a multiplier which makes the lower bound larger than the upper bound at Δ^* , and this multiplier will work for all other Δ .

Proof of Corollary 2.1. We begin by using Lemma 5 to identify $\Delta^* = \sqrt{32e\sigma^2c \log\left(\frac{e+1}{(ec)^2}\right)}$, which, combined with Lemma 6, allows us to write,

$$\sup_{\nu \in \mathcal{M}_\Delta} \mathcal{R}_{\text{MI}}(\pi, \nu) \leq \text{UB}_{\text{MI}}(\Delta) := \begin{cases} \exp\left(-\frac{\Delta^2}{8e\sigma^2c}\right) + \frac{1}{e} + 3c, & \text{if } \Delta \leq \Delta^* \\ \frac{32\sigma^2c \log\left(\frac{e+1}{(ec)^2}\right)}{\Delta^2} + 3c, & \text{if } \Delta > \Delta^* \end{cases} \quad (19)$$

where the additive $3c$ term is to account for integer pulls for each of the 2 arms and an additional c bound for the expected risk on G^c . Clearly, for any $a \geq 128 \log\left(\frac{e+1}{(ec)^2}\right)$, (19) is upper bounded by $a\text{LB}_{\text{MI}}(\Delta) + 3c$ for all $\Delta > \Delta^*$. We then divide our analysis for the remaining Δ into two cases: when $\Delta \leq \sqrt{e\sigma^2c}$ and otherwise. First, when $\Delta \leq \sqrt{e\sigma^2c}$,

$$\text{UB}_{\text{MI}}(\Delta) \leq \frac{e+1}{e} + 3c, \quad \text{LB}_{\text{MI}}(\Delta) \geq \frac{1}{2e},$$

and so $\text{UB}_{\text{MI}}(\Delta) \leq 8\text{LB}_{\text{MI}}(\Delta)$ for $\Delta \leq \sqrt{e\sigma^2c}$. Finally, we must consider $\Delta \in (\sqrt{e\sigma^2c}, \Delta^*]$. We begin by comparing (19) and (3) at Δ^* , then we prove that this is sufficient. This gives us,

$$\text{UB}_{\text{MI}}(\Delta^*) = \left(\frac{(ec)^2}{e+1}\right)^4 + \frac{1}{e} + 3c, \quad \text{LB}_{\text{MI}}(\Delta^*) = \frac{\log\left(32e^2 \log\left(\frac{e+1}{(ec)^2}\right)\right)}{128e \log\left(\frac{e+1}{(ec)^2}\right)}$$

Supposing $c < 1/4$,² we can see that $\text{UB}_{\text{MI}}(\Delta^*) \leq 128 \log\left(\frac{e+1}{(ec)^2}\right) \text{LB}_{\text{MI}}(\Delta^*)$. Finally, we conclude that this is sufficient to prove the statement by showing that $128 \log\left(\frac{e+1}{(ec)^2}\right) \text{LB}_{\text{MI}}(\Delta) - \text{UB}_{\text{MI}}(\Delta)$ is decreasing for $\Delta \in (\sqrt{e\sigma^2c}, \Delta^*]$. We show this here:

$$\begin{aligned} \frac{\partial}{\partial \Delta} a\text{LB}_{\text{MI}}(\Delta) - \text{UB}_{\text{MI}}(\Delta) &= -\frac{a\sigma^2c}{2\Delta^3} \log\left(\frac{\Delta^2}{\sigma^2c}\right) + \frac{\Delta}{4e\sigma^2c} \exp\left(-\frac{\Delta^3}{8e\sigma^2c}\right) \\ &\leq -\frac{a\sigma^2c}{2\Delta^3} + \frac{\Delta}{4e\sigma^2c} \left(\frac{8e\sigma^2c}{\Delta^2}\right)^2 \\ &= -\frac{a\sigma^2c}{2\Delta^3} + \frac{16e\sigma^2c}{\Delta^3}, \end{aligned}$$

which is < 0 when $a > 32e$, which is true for $a = 128 \log\left(\frac{e+1}{(ec)^2}\right)$. Thus, we have proven $\forall \Delta$,

$$128 \log\left(\frac{e+1}{(ec)^2}\right) \text{LB}_{\text{MI}}(\Delta) \geq \text{UB}_{\text{MI}}(\Delta) \geq \sup_{\nu \in \mathcal{M}_\Delta} \mathcal{R}_{\text{MI}}(\pi, \nu)$$

□

Proof of Corollary 4.1. We follow the same general proof strategy as in the previous proof. We again apply Lemma 5 to identify $\Delta^* = (\sigma^2c)^{1/3} \sqrt{(8e)/3 \log(2N^*/\delta)}$ and combine it with Lemma 6 to write,

$$\sup_{\nu \in \mathcal{M}_\Delta} \mathcal{R}_{\text{SR}}(\pi, \nu) \leq \text{UB}_{\text{SR}}(\Delta) := \begin{cases} \Delta \exp\left(-\frac{3\Delta^2}{8e(\sigma^2c)^{2/3}}\right) + \frac{3}{e}(\sigma^2c)^{1/3} + 3c, & \text{if } \Delta \leq \Delta^* \\ \frac{32\sigma^2c \log(2N^*/\delta)}{\Delta^2} + 3c, & \text{if } \Delta > \Delta^* \end{cases} \quad (20)$$

²Previously, we have not put any restriction on the value of c , but we have implicitly assumed $c \ll 1$ by the construction of our problem setting. Consider that, under $\mathcal{R}_{\text{MI}}$, if $c \geq 1/4$, one will perform uniformly best on all instances by simply guessing the optimal arm uniformly at random. We do not explicitly account for this behavior in our algorithm construction for simplicity, but it is unrealistic to let $c \geq 1/4$ in practical settings.

First, when $\Delta < (\sigma^2 c)^{1/3}$, clearly $\text{UB}_{\text{SR}}(\Delta) \leq 4\text{LB}_{\text{MI}}(\Delta) + 2(\sigma^2 c)^{1/3} + 3c$ by (20). Then, noting that $\frac{3B\sigma^{4/3}}{c^{5/3}} \geq \frac{2N^*}{\delta}$, we can clearly see that $\text{UB}_{\text{SR}}(\Delta) \leq 128 \log\left(\frac{3B\sigma^{4/3}}{c^{5/3}}\right) \text{LB}_{\text{SR}}(\Delta) + 3c$ for $\Delta > \Delta^*$. To prove this same bound for $\Delta \in [(\sigma^2 c)^{1/3}, \Delta^*]$, we follow the same technique as in the previous proof. First, when $\Delta \in [(\sigma^2 c)^{1/3}, (\sqrt{e}\sigma^2 c)^{1/3}]$,

$$\text{UB}_{\text{SR}}(\Delta) \leq (\sigma^2 c)^{1/3} \left[\exp\left(\frac{1}{6} - \frac{3}{8e^{2/3}}\right) + \frac{3}{e} \right] + 3c, \quad \text{LB}_{\text{SR}}(\Delta) \geq \frac{(\sigma^2 c)^{1/3}}{4},$$

and thus $\text{UB}_{\text{SR}}(\Delta) \leq 9\text{LB}_{\text{SR}}(\Delta) + 3c$ for $\Delta \in [(\sigma^2 c)^{1/3}, (\sqrt{e}\sigma^2 c)^{1/3}]$. To prove the bound for $\Delta \in ((\sqrt{e}\sigma^2 c)^{1/3}, \Delta^*]$, we again compare the two at Δ^* and then show that the difference between the functions is decreasing in this range of Δ , and thus this is sufficient. At Δ^* , we have,

$$\begin{aligned} \text{UB}_{\text{SR}}(\Delta^*) &= \frac{(\sigma^2 c)^{1/3} \sqrt{\frac{32}{3e} \log(2N^*/\delta)}}{(2N^*/\delta)^4} + \frac{3}{e}(\sigma^2 c)^{1/3} + 3c \leq \frac{5}{e}(\sigma^2 c)^{1/3} + 3c \\ \text{LB}_{\text{SR}}(\Delta^*) &= \frac{3(\sigma^2 c)^{1/3}}{128e \log(2N^*/\delta)} \log\left(\frac{32e^{5/2}}{3^{3/2}} \log^{3/2}(2N^*/\delta)\right) \geq \frac{9(\sigma^2 c)^{1/3}}{128e \log(2N^*/\delta)} \end{aligned}$$

Thus, we have $\text{UB}_{\text{SR}}(\Delta^*) \leq 128 \log\left(\frac{3B\sigma^{4/3}}{c^{5/3}}\right) \text{LB}_{\text{SR}}(\Delta^*) + 3c$. We then conclude this portion of the proof by showing $128 \log\left(\frac{3B\sigma^{4/3}}{c^{5/3}}\right) \text{LB}_{\text{SR}}(\Delta) - \text{UB}_{\text{SR}}(\Delta)$ is decreasing for $\Delta \in ((\sqrt{e}\sigma^2 c)^{1/3}, \Delta^*]$. We show this here:

$$\frac{\partial}{\partial \Delta} a\text{LB}_{\text{SR}}(\Delta) - \text{UB}_{\text{SR}}(\Delta) = -\frac{a\sigma^2 c}{4\Delta^3} \log\left(\frac{\Delta^6}{e(\sigma^2 c)^2}\right) - \exp\left(-\frac{3\Delta^2}{8e(\sigma^2 c)^{2/3}}\right) \left(1 - \frac{3\Delta^2}{4e(\sigma^2 c)^{2/3}}\right)$$

This is < 0 for any $a \geq 0$ when $\Delta \in ((\sqrt{e}\sigma^2 c)^{1/3}, \sqrt{4e/3}(\sigma^2 c)^{1/3})$. When $\Delta \in [\sqrt{4e/3}(\sigma^2 c)^{1/3}, \Delta^*]$,

$$\begin{aligned} \frac{\partial}{\partial \Delta} a\text{LB}_{\text{SR}}(\Delta) - \text{UB}_{\text{SR}}(\Delta) &\leq -\frac{a\sigma^2 c}{2\Delta^3} + \frac{3\Delta^2}{4e(\sigma^2 c)^{2/3}} \left(\frac{8e(\sigma^2 c)^{2/3}}{3\Delta^2}\right)^{5/2} \\ &= -\frac{a\sigma^2 c}{2\Delta^3} + \frac{\sigma^2 c}{4\Delta^3} \left(8^{5/2} \left(\frac{e}{3}\right)^{3/2}\right), \end{aligned}$$

which is < 0 for any $a > 78$, and in particular, $a = 128 \log\left(\frac{3B\sigma^{4/3}}{c^{5/3}}\right)$. Now, all that is left to prove is the worst-case comparison with LB_{SR}^* . We can show this simply by considering that (20) is maximized at $\Delta = \sqrt{4e/3}(\sigma^2 c)^{1/3}$, where it takes value $(\sqrt{4/3} + 3/e)(\sigma^2 c)^{1/3} + 3c$, which is clearly upper bounded by $9\text{LB}_{\text{SR}}^* + 3c$. $\square$

E Additional Experiments

Here we provide additional K -arm experiments. All experiments were performed using a 3.7GHz AMD Ryzen 9 5900X 12-Core processor with 24 GB of memory. Total runtime across all experiments took approximately 7.5 hours, and safeguards were employed to prevent the fixed confidence algorithms from continuing to sample after already severely underperforming the other methods when the sub-optimality gaps were particularly small ($10/c$ total samples allowed).

K-arm simulations. We now include a number of additional K -arm experiments to demonstrate that our algorithm continues to perform well compared to traditional fixed budget and confidence methods when we move beyond the simple 2-arm case. For all of our K -arm experiments, we choose to use Gaussian arms with $\sigma^2 = 1$ for simplicity. We begin with the “1-sparse” setting, where $\mu_1 = \Delta$ and $\mu_k = 0$ for all $k \neq 1$, resulting in $H = (K - 1)\Delta^{-2}$, for $\Delta \in [0.05, 2]$ for the probability of misidentification performance penalty and $\Delta \in [0.05, 3]$ for the simple regret performance penalty. We additionally vary K among 8, 16, and 32. For these experiments, we average across 10^4 runs each with different random seeds. As in § 4, we compare to Sequential Halving [28] for fixed budget and we use an elimination, or “racing,” procedure for fixed confidence, with confidence bounds $\sqrt{4\sigma^2 n^{-1} \log(Kn\delta^{-1})}$. To extend to the K -arm case, our “low” budget is now $5K$, and our “high” budget is $250K$, which align with our choices of 10 and 500 in the 2-arm case. We still use $\delta = 0.1$

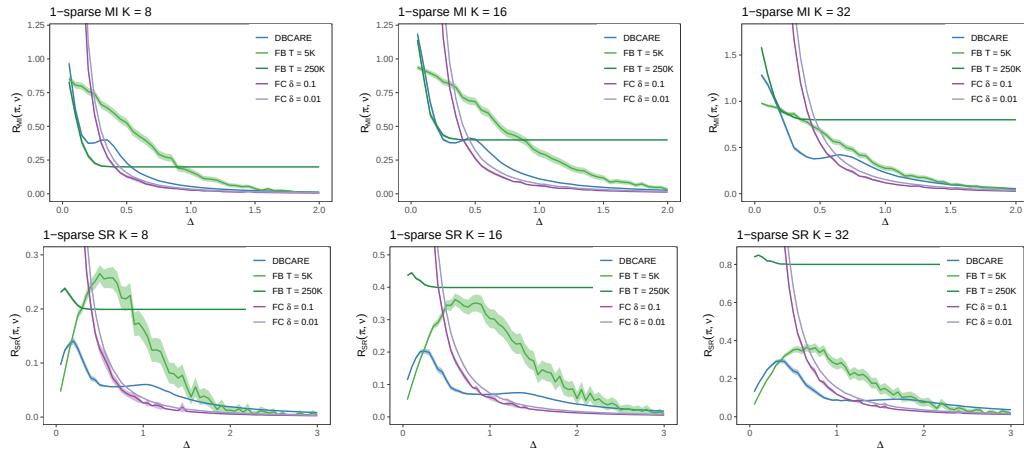


Figure 4: Comparisons between DBCARE and fixed budget and confidence algorithms for $\mathcal{R}_{\text{MI}}$ and $\mathcal{R}_{\text{SR}}$ in the K -arm 1-sparse setting. Y -axes are adjusted per setting to highlight problem-specific behavior. Confidence regions represent empirical average risk ± 2 SE.

and $\delta = 0.01$ for our confidences. As we can see in Fig 4, in the 1-sparse setting, DBCARE still enjoys uniformly good performance across the full range of Δ , while the fixed budget and confidence approaches have some region where they perform sub-optimally.

To explore the performance of DBCARE and fixed confidence and budget approaches across a variety of problem structures, we additionally considered the “linear decay” setting, where we set $\mu_1 = \Delta_2$ and $\mu_k = -\Delta_2(\frac{k-2}{K-2})$ for $k \neq 1$ so that the suboptimality gaps linearly increase from Δ_2 to $2\Delta_2$. This results in $H \approx 0.5K\Delta_2^{-2}$. We again let $\Delta_2 \in [0.05, 2]$ for $\mathcal{R}_{\text{MI}}$ and $\Delta_2 \in [0.05, 3]$ for $\mathcal{R}_{\text{SR}}$, average across 10^4 runs each with a different random seed, and vary K among 8, 16, and 32. As we can see in Fig 5, this setting provides similar results to the 1-sparse and 2-arm settings, with DBCARE performing well across the range of Δ_2 values, while the other methods generally perform sub-optimally for some Δ_2 values.

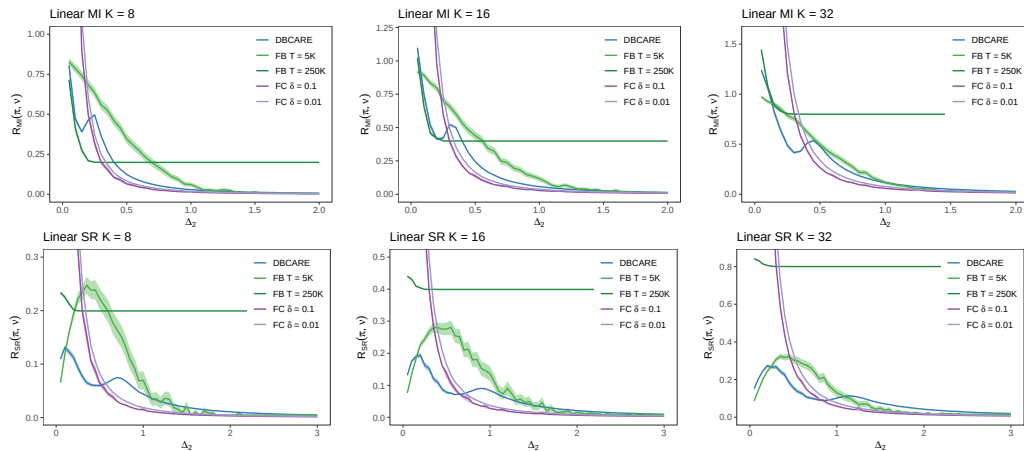


Figure 5: Comparisons between DBCARE and fixed budget and confidence algorithms for $\mathcal{R}_{\text{MI}}$ and $\mathcal{R}_{\text{SR}}$ in the K -arm linear decay setting. Y -axes are adjusted per setting to highlight problem-specific behavior. Confidence regions represent empirical average risk ± 2 SE.