Revisiting Simple Regret Minimization in Multi-Armed Bandits

Yao Zhao
University of Arizona
yaoz@arizona.edu

Connor Stephens
University of Alberta
cjs@ualberta.ca

Csaba Szepesvári
Amii, University of Alberta, DeepMind
szepesva@ualberta.ca

Kwang-Sung Jun
University of Arizona
kjun@cs.arizona.edu

Abstract

Simple regret is a natural and parameter-free performance criterion for identifying a good arm in multi-armed bandits yet is less popular than the probability of missing the best arm or an ϵ-good arm, perhaps due to lack of easy ways to characterize it. In this paper, we achieve improved simple regret upper bounds for both data-rich ($T \geq n$) and data-poor regime ($T \leq n$) where n is the number of arms and T is the number of samples. At its heart is an improved analysis of the well-known Sequential Halving (SH) algorithm that bounds the probability of returning an arm whose mean reward is not within ϵ from the best (i.e., not ϵ-good) for *any* choice of $\epsilon > 0$, although ϵ is not an input to SH. We show that this directly implies an optimal simple regret bound of $\mathcal{O}(\sqrt{n/T})$. Furthermore, our upper bound gets smaller as a function of the number of ϵ-good arms. This results in an accelerated rate for the (ϵ, δ)-PAC criterion, which closes the gap between the upper and lower bounds in prior art. For the more challenging data-poor regime, we propose Bracketing SH (BSH) that enjoys the same improvement even without sampling each arm at least once. Our empirical study shows that BSH outperforms existing methods on real-world tasks.

1 Introduction

We consider the pure exploration problem in multi-armed bandits. In this problem, given n arms, the learner sequentially chooses an arm $I_t \in [n] := \{1, \dots, n\}$ at time t to observe a reward $r_t = \mu_{I_t} + \eta_t$ where μ_i is the mean reward of arm i and η_t is stochastic σ^2-sub-Gaussian noise. Without loss of generality, we assume that $1 \geq \mu_1 \geq \cdots \geq \mu_K \geq 0$. After T time steps, the learner is required to output an arm J_T that is estimated to have the largest mean reward. Such a problem was formalized by Even-Dar et al. [11] and Bubeck et al. [6], but similar problems were considered much earlier [5, 10]. Recently, pure exploration has found many applications such as efficient crowdsourcing for cartoon caption contests [21], hyperparameter optimization [19], and accelerating time complexity of clustering algorithms [4].

Among many performance criteria, *simple regret*, proposed by Bubeck et al. [6], is a natural measure of the quality of estimated best arm J_T:

$$\mathsf{SReg}_T := \mathbb{E}\left[\mu_1 - \mu_{J_T}\right] . \tag{1}$$

Simple regret is an unverifiable performance guarantee, meaning that the algorithm need not verify its performance to a prescribed level. This is in stark contrast to the fixed confidence setting Even-Dar et al. [11] that requires the algorithm to *verify* the quality of the estimated best arm to the prescribed target accuracy ϵ and confidence level δ. While certain applications do need such a verification, there are many applications where the sampling budget is limited such as cartoon caption contest [14]

Preprint. Under review.

where algorithms with verifiable guarantee may not be meaningful. Furthermore, algorithms with verifiable guarantees are mostly vacuous for the data-poor regime of $T \leq n$ as it requires each arm to be pulled at least once Katz-Samuels and Jamieson [16, Section C.1]. Another natural criterion is measuring the probability of misidentifying the best or an ϵ-good arm as in the fixed budget setting. While it has unverifiable guarantees, the bound is inherently subject to a given error level ϵ.[1] In contrast, simple regret integrates out ϵ in the expectation. We refer to Section 2 for a detailed comparison with other criteria.

Despite being attractive, the analysis of simple regret has been elusive. Existing studies focus on either achieving minimax simple regret bound [17] or characterization of how simple regret is different from the cumulative regret [6]. For example, it is not known whether simple algorithms like Sequential Halving (SH) [15] enjoy the optimal simple regret or not.

In this paper, we revisit simple regret and make two main contributions. First, we provide a novel and tight analysis of Sequential Halving (SH) that is one of the state-of-the-art algorithms for pure exploration. Let $\Delta_i = \mu_1 - \mu_i$. Our analysis result is flexible in that it bounds the probability that SH outputs an ϵ-good arm for *any choice* of $\epsilon > 0$ (note ϵ is *not* an input to SH); i.e.,

$$\mathbb{P}(\mu_{J_T} < \mu_1 - \epsilon) \leq \min_{(m,\epsilon'):\Delta_m + \epsilon' \leq \epsilon} \exp\left(-\tilde{\Theta}\left(m\frac{(\epsilon')^2 T}{n}\right)\right). \tag{2}$$

for large enough T (see Theorem 1 for detail). Interestingly, this bound shows that the error probability is small when many arms have small gaps. In an instance where there are m arms that are ϵ-good, the bound above translates to the sample complexity of $\mathcal{O}\left(\frac{n}{m\epsilon^2}\log\left(\frac{1}{\delta}\right)\right)$, which shaves off a factor of $\log(1/\delta)$ from Chaudhuri and Kalyanakrishnan [9] and settles down the gap between upper and lower bounds [9]. Interestingly, the issue of having $\log^2(1/\delta)$ not $\log(1/\delta)$ for bounds adaptive to the number of good arms also appears in Aziz et al. [3] and Katz-Samuels and Jamieson [16, Section E]. We summarize the sample complexities of various algorithms in Table 1.

A remarkable aspect of the bound like (2) is that we can directly derive a simple regret bound. That is, using the identity $\mathbb{E}[X] = \int_0^\infty \mathbb{P}(X > \epsilon)\,\mathrm{d}\epsilon$ for nonnegative random variable X, the bound (2) with $(m = 1, \epsilon' = \epsilon)$ implies the minimax simple regret bound of $\tilde{\mathcal{O}}(\sqrt{n/T})$. We emphasize that this is possible since ϵ is only an analysis parameter and not an input to the algorithm. This is in contrast to algorithms that require either the target error ϵ or failure probability δ in the guarantee to be an input to the algorithm (e.g., Katz-Samuels and Jamieson [16], Even-Dar et al. [11]) for which such a trick is not available, and the only ways to obtain simple regret bounds that we are aware of requires setting ϵ or δ in certain ways; details found in Section 2. We summarize our novel analysis of SH and its anytime version called Doubling SH (DSH) in section 3.

As a second contribution, we propose an improved pure exploration algorithm for the data-poor regime (i.e., $T \leq n$). This is motivated by the fact that the guarantee 2 becomes vacuous in the data-poor regime (i.e., $T \leq n$) as the tight budget does not allow us to pull every arm once. Inspired by Katz-Samuels and Jamieson [16], we propose Bracketing SH (BSH) that progressively subsamples a larger set of arms (i.e., brackets) for which we invoke SH to find a good arm. Our analysis shows that BSH enjoys a bound that is comparable to (2) but even for $T \leq n$. Compared to BUCB [16], the state-of-the-art algorithm for the data-poor regime, BSH achieves two main improvements: (i) BSH is parameter-free and does not take any input whereas BSH requires the target error rate δ and the noise level σ^2 as input and (ii) BSH has a high probability sample complexity bound as opposed to the expected sample complexity bound of BUCB. In terms of the sample complexity, BSH tends to show a favorable bound for moderately large ϵ or large enough n. In the linear instance with $\Delta_i = (i/n)$, for example, the sample complexity bound of BSH is $\frac{1}{\epsilon^3}$ and that of BUCB is $\frac{n}{\epsilon}$, which implies that BSH is favorable as n gets larger while ϵ is fixed, though BUCB is favorable when ϵ is small enough. We present BSH and the analysis in Section 4 and perform an empirical study with the real-world task of evaluating cartoon captions in Section 5, which shows the superiority of BSH over BUCB. Finally, we conclude our paper with exciting future research directions in Section 6.

[1] In fact, there are no explicit algorithms in the literature for guaranteeing the identification of an ϵ-good arm in the fixed budget setting, to our knowledge. However, one can easily turn Median Elimination [11] into a fixed budget version by adjusting the confidence level δ so it terminates within the budget T.

	Linear	Equal gap	Data-poor	Anytime	Parameter-free
ME [11]	$\frac{n}{\epsilon^2}\log\left(\frac{1}{\delta}\right)$	$\frac{n}{\epsilon^2}\log\left(\frac{1}{\delta}\right)$	✗	✗	✗
$\mathcal{P}_3$ [9]	$\frac{1}{\epsilon^2}\left(\frac{1}{\epsilon}\log\left(\frac{1}{\delta}\right)+\log^2\left(\frac{1}{\delta}\right)\right)$	$\frac{1}{\epsilon^2}\left(\frac{n}{m}\log\left(\frac{1}{\delta}\right)+\log^2\left(\frac{1}{\delta}\right)\right)$	✓	✗	✗
BUCB† [16]	$\frac{n}{\epsilon}\log\left(\frac{1}{\delta}\right)$	$\frac{n}{m\epsilon^2}\log\left(\frac{1}{\delta}\right)$	✓	✓	✗
DSH (ours)	$\frac{1}{\epsilon^3}\log\left(\frac{1}{\delta}\right)\vee\frac{n}{\epsilon^2}$	$\frac{n}{m\epsilon^2}\log\left(\frac{1}{\delta}\right)\vee\frac{n}{\epsilon^2}$	✗	✓	✓
BSH (ours)	$\frac{1}{\epsilon^3}\log\left(\frac{1}{\delta}\right)$	$\frac{n}{m\epsilon^2}\log\left(\frac{1}{\delta}\right)$	✓	✓	✓

Table 1: Comparison of sample complexity for identifying an ϵ-good arm; we omit the logarithmic dependence on n and m. The linear instance has the gap $\Delta_i = i/n, \forall i \geq 2$ and the equal gap instance has $m-1$ arms with the gap $\epsilon/2$ and $n-m$ arms with the gap $\frac{3}{2}\epsilon$. The dagger symbol means that the sample complexity is in expectation rather than high probability. One can also derive simple regret bounds for these instances, which is deferred to the appendix due to space constraints.

2 Problem Setup and Preliminaries

In the pure exploration problem, we are given n arms with mean rewards $\{\mu_i\}_{i\in[n]}$. We assume $1 \geq \mu_1 \geq \cdots \geq \mu_n \geq 0$, which is for ease of exposition and algorithms are assumed to have access to arbitrarily permuted indices. At each time step t, the learner chooses an arm $I_t \in [K]$ and receives a reward $r_t = \mu_{I_t} + \eta_t$ where η_t is σ^2-sub-Gaussian. For simplicity, however, we present our results with $\sigma^2 = 1$ throughout.

Performance Criteria. For the *fixed budget* setting, the learner is given a budget $T \geq 1$ as input and is required to output an estimated best arm J_T at the end of T-th time step. For the *anytime* setting, the learner is required to do so for every $T \geq 1$. The simple regret of a learner is then defined as the gap between the mean reward of arm J_T and that of the best arm as stated in (1).

Another natural performance criterion we mainly focus in theoretical development is the *ϵ-error probability*, which measures the probability of not returning an ϵ-good arm:

$$\mathbb{P}(\mu_1 - \mu_{J_T} > \epsilon) \,.$$

This criterion is in fact closely related to simple regret. To see this, suppose we have an algorithm does not take ϵ as input yet bounds the ϵ-error probability as a function of $f(\epsilon)$ for *any* $\epsilon > 0$ in which case we say that the algorithm enjoys a *uniform ϵ-error probability bound*. Then, one can use the identity $\mathbb{E}[X] = \int_0^\infty \mathbb{P}(X > \epsilon)\,\mathrm{d}\epsilon$ for a nonnegative random variable X to obtain

$$\mathsf{SReg}_T \leq \int_0^\infty f(\epsilon)\,\mathrm{d}\epsilon \,.$$

The ϵ-error probability is closely related to the (ϵ,δ)-PAC guarantee, which allows the algorithm to use as many samples T as possible until it can claim an ϵ-good arm J_T with a verification that the claim is true with probability at least $1-\delta$: $\mathbb{P}(\mu_1 - \mu_{J_T} > \epsilon) \leq \delta$. The main difference is that this requires verification, which is not suited for the data-poor regime as mentioned in Section 1.

A more general setting is the (m,ϵ)-*identification problem*, where the goal is to identify one of the arms i is (m,ϵ)-good, i.e., $\mu_i \geq \mu_m - \epsilon$. This is a special case of the (k,m,n)-problem defined in Chaudhuri and Kalyanakrishnan [9] for $k=1$, which was originally developed for the fixed confidence setting where m and ϵ are input to the algorithm. In fact, finding an (m,ϵ)-good arm is equivalent to finding a $(\Delta_m+\epsilon)$-good arm. The only difference from $(1,m,n)$-problem is that we do not require the algorithm to verify correctness and that we do not provide m and ϵ as input to the algorithm.

Sample Complexity. When comparing bounds for simple regret or ϵ-error probability, it is often easier to consider the sample complexity, which measures the number of samples required to achieve the target performance level. For example, the simple regret sample complexity of an algorithm is the smallest time step τ such that $\mathsf{SReg}_T \geq \epsilon$ for every $T \geq \tau$. Similarly, the (ϵ,δ)-PAC sample complexity of an algorithm is the smallest time step τ such that $\mathbb{P}(\mu_1 - \mu_{J_T} > \epsilon) \leq \delta$ for every $T \geq \tau$. Note that the latter is slightly different from (ϵ,δ)-unverifiable sample complexity introduced in Katz-Samuels and Jamieson [16].

Sample complexity in pure exploration mainly comes with two tastes: the worst-case sample complexity that depends only on a target error ϵ or the instance-dependent sample complexity that depends directly on $\mu_1,\cdots,\mu_n$. The main result of our paper is closer to the worst-case sample complexity but it depends on the $\mu_1,\cdots,\mu_n$ via the number of ϵ-good arms denoted by $m(\epsilon)$. A fully instance-

Algorithm 1: Sequential Halving (SH)

Input: budget: T, arms: $[n]$
Initialize: $S_1 = [n]$
for $\ell = 1, \dots, \lceil \log_2 n \rceil$ **do**
  Sample each arm $i \in S_\ell$ for T_ℓ times where $T_\ell = \left\lfloor \frac{T}{|S_\ell| \lceil \log_2 n \rceil} \right\rfloor$.
  Let $S_{\ell+1}$ be the set of $\lceil S_\ell / 2 \rceil$ arms in S_ℓ with the largest empirical rewards.
Set J_T as the only arm in $S_{\lceil \log_2 n \rceil}$.
Output: J_T

dependent sample complexity that enjoys a similar accelerated rate as a function of $m(\epsilon)$ is beyond the scope of our work and is left as future work.

As tight sample complexity bounds for ϵ-good arm or simple regret often take complicated form, we use bandit instances for comparison. The first one is the *polynomial* instance with $\Delta_i = \left(\frac{i}{n}\right)^\alpha$ where $\alpha > 0$ adjust the fraction of good-arms. The other one is the *equal gap* instance that such that

$$\mu_1 = (3/2)\epsilon, \mu_i = \epsilon, \forall i \in \{2, \dots, m\}, \mu_i = 0, \forall i \geq m+1 \text{ for some } m \geq 2, \epsilon > 0\,. \tag{3}$$

Notations. We define $\Delta_i := \mu_1 - \mu_i$ and $\Delta_{i,j} := \mu_i - \mu_j$ for $i < j$. Denote $\mathrm{Top}_m(\epsilon) = \{i : \mu_i \geq \mu_m - \epsilon\}$ and $\mathrm{Bot}_m(\epsilon) = \{i : \mu_i < \mu_m - \epsilon\}$. We define shortcuts $\mathrm{Top}_m := \mathrm{Top}_m(0)$ and $\mathrm{Bot}_m := \mathrm{Bot}_m(0)$. Throughout the paper, "const" is a universal and positive constant, which may have different values for different expressions.

3 Improved Analyses of Sequential Halving

Sequential Halving (SH) [15] is one of the state-of-the-art algorithms for the best-arm identification. Specifically, it is known to achieve a nearly optimal instance-dependent bound on the $(\epsilon = 0)$-error probability, with details in the summary of Table 2 in the appendix. However, the analysis of ϵ-error probability for $\epsilon > 0$ for SH is not known, to our knowledge. It is not clear if SH enjoys a low (ϵ, δ)-PAC sample complexity like $\frac{n}{\epsilon^2}\log(1/\delta)$ of Median Elimination (ME) [11] or not because the sampling scheme of SH is different from ME.

In this section, we provide a positive answer via a powerful analysis of the ϵ-error probability incurred by SH, which takes a generic form that can be specialized to become a minimax optimal unverifiable (ϵ, δ)-PAC sample complexity [20] and minimax optimal simple regret bound [18, Exercise 33.1]. To describe SH, it consists of $\lceil \log_2(n) \rceil$ stages. In each stage ℓ, the algorithm performs a uniform sampling over the surviving arms S_ℓ followed by eliminating the bottom half of S_ℓ w.r.t. empirical means and sets $S_{\ell+1}$ as the resulting arm set. We present SH in Algorithm 1.

We first show in Theorem 1 the ϵ-error probability bound for SH and discuss its implications.

Theorem 1. *For any $\epsilon \in (0,1)$, the error probability of SH for identifying an ϵ-good arm satisfies*

$$\mathbb{P}\left(\mu_{J_T} < \mu_1 - \epsilon\right) \leq \min_{\{(m',\epsilon'):\Delta_{m'}+\epsilon' \leq \epsilon\}} \log_2 n \cdot \exp\left(-\mathrm{const} \cdot m' \cdot \left(\frac{\epsilon'^2 T}{4n \log_2^2(2m') \log_2 n} - \ln(4e)\right)\right).$$

We stress that, unlike all existing work we are aware of, Theorem 1 bounds the ϵ-error probability for an algorithm that does not require ϵ as input. That is to say, ϵ is a parameter for analysis only. The next corollary is a rephrased version that is easier to compare with existing work.

Corollary 1.1. *Suppose we run SH. Then, for any $m \leq n$ and any $\epsilon \in (0,1)$,*

$$\mathbb{P}\left(\mu_{J_T} < \mu_m - \epsilon\right) \leq \log_2 n \cdot \exp\left(-\mathrm{const} \cdot m \cdot \left(\frac{\epsilon^2 T}{4n \log_2^2(2m) \log_2 n} - \ln(4e)\right)\right).$$

Thus, there exist an absolute constant $c_1 > 0$ s.t. if $T \geq c_1 \frac{(\ln(4e) + \ln\ln n) n \log_2^2(2m) \log_2 n}{\epsilon^2} = \tilde{\Theta}\left(\frac{n}{\epsilon^2}\right)$,

$$\mathbb{P}\left(\mu_{J_T} < \mu_m - \epsilon\right) \leq \exp\left(-\tilde{\Theta}\left(m \frac{\epsilon^2 T}{n}\right)\right).$$

Algorithm 2: Doubling Sequential Halving (DSH)

Input arms: $[n]$

Initialize $T_1 = \lceil n \log_2 n \rceil$, $\hat{j} = j$ for some arbitrarily chosen $j \in [n]$, $\hat{\mu}^* = -\infty$. Define the time blocks $\mathcal{T}_1 = \{1, \dots, T_1\}$ and $\mathcal{T}_k = \{T_1(2^{k-1} - 1) + 1, \dots, T_1(2^k - 1)\}$, $\forall k \geq 2$, which satisfies $|\mathcal{T}_k| = T_1 2^{k-1}$. $k = 1$. An instance of SH denoted by $\mathcal{S}_1$ with budget $|\mathcal{T}_1|$.

for $t = 1, 2, \dots$ **do**

  Pull arm I_t according to the recommendation from $\mathcal{S}_k$

  Receive reward R_t and send it to $\mathcal{S}_k$

  if $t = \max \mathcal{T}_k$ **then**

    Set $(\hat{j}, \hat{\mu}^*)$ as the output arm and its empirical mean from the last stage of $\mathcal{S}_k$

    $k \leftarrow k + 1$

    Initialize a new instance of SH denoted by $\mathcal{S}_k$ with budget $|\mathcal{T}_k|$

  Output: $J_t = \hat{j}$ and $M_t = \hat{\mu}^*$.

Note the requirement of $T \geq \tilde{\Theta}\left(\frac{n}{\epsilon^2}\right)$ is not a weakness of our result. One can easily see that if $\Delta_i = \epsilon, \forall i \geq 2$, the standard result of SH [15] becomes vacuous for $T = o(\frac{n}{\epsilon^2})$.

Implications for $m = 1$. The worst-case upper bound of Corollary 1.1 for $m = 1$ corresponds to the sample complexity of $\tilde{\mathcal{O}}\left(\frac{n}{\epsilon^2} \log \frac{1}{\delta}\right)$ where, hereafter, $\tilde{\mathcal{O}}$ hides logarithmic factors except for $1/\delta$. This is the same as the classic lower bound result of the (ϵ, δ)-PAC problem Mannor and Tsitsiklis [20] up to logarithmic factors. While numerous algorithms achieve the matching upper bound including Even-Dar et al. [11] and Hassidim et al. [13], our result in Theorem 1 indicates a tighter bound when here are many good arms. Note that one can adjust T_ℓ to achieve the exact minimax optimality up to a constant factor, which we show in Appendix B.6 due to space constraints.

We can now apply the the trick explained in Section 2 that converts a uniform ϵ-error probability bound into a simple regret bound.

Corollary 1.2. *SH satisfies* $\mathsf{SReg}_T \leq \tilde{\mathcal{O}}\left(\sqrt{\frac{n}{T}}\right)$.

Note that our bound for is minimax optimal up to logarithmic factors [18, Exercise 33.1]. To our knowledge, the only algorithm that we are aware of that achieves the minimax simple regret is MOSS [2]. However, as MOSS is designed to minimize cumulative regret, it cannot enjoy low instance-dependent bounds for pure exploration tasks [7]. While uniform sampling can also achieve a near-optimal simple regret, one can show that uniform sampling can be arbitrarily worse than SH w.r.t. the $(\epsilon = 0)$-error probability as its sample assignments over the arms are inherently non-adaptive.

Remark 1. *Note that one can alter T_ℓ in SH to achieve the optimal (ϵ, δ)-PAC sample complexity and simple regret without extra logarithmic factors. However, it is not clear whether this modified version enjoys near-optimal instance-dependent sample complexity. We discuss details in the appendix for space constraints.*

Implications for $m > 1$. We now show that SH enjoys the optimal worst-case guarantee for the (m, ϵ)-identification problem up to logarithmic factors. Our result is especially attractive since the target number of good arms m is not an input to SH.

Our upper bound of Corollary 1.1 corresponds to a sample complexity of $\tilde{\mathcal{O}}\left(\frac{n}{m\epsilon^2} \log\left(\frac{1}{\delta}\right)\right)$, which is matches the lower bound in Chaudhuri and Kalyanakrishnan [8], thereby closing the gap between the upper and lower bounds from previous work; the algorithms therein have suboptimal upper bounds that scale with $\log^2\left(\frac{1}{\delta}\right)$. Furthermore, these algorithms all require (m, ϵ, δ) as input, which is natural for obtaining verifiable sample complexities but becomes a limitation in other settings. In contrast, in our result of Corollary 1.1, m, ϵ are freely chosen to measure the performance of the algorithm. In other words, the parameters for measuring the performance are not necessary to be used for executing the algorithm anymore.

Anytime version of SH. To support anytime simple regret minimization, we combine SH with the doubling trick [14], which we call Doubling SH (DSH). Specifically, DSH repeatedly runs SH. The k-th Sequential Halving has a doubled budget input compared with the $(k-1)$-th Sequential Halving. Before it finishes k-th Sequential Halving, it always returns the output of $(k-1)$-th Sequential

Halving. The pseudocode can be found in Algorithm 2. The following theorem shows that DSH enjoys the same guarantee as SH up to a constant factor.

Theorem 2. *The error probability of DSH satisfies, for any* $\epsilon > 0$*,*

$$\mathbb{P}\big(\mu_{J_t} < \mu_1 - \epsilon\big) \le \min_{\{(m',\epsilon'):\Delta_{m'}+\epsilon'\le\epsilon\}} \log_2 n \cdot \exp\left(-\text{const}\cdot m'\left(\frac{\epsilon'^2 t}{16 n \log_2^2(2m')\log_2 n} - \ln(4e)\right)\right).$$

Proof. This is a direct consequence of the fact that, at time t, the latest finished SH was run with a budget of at least $t/4$. The full proof can be found in Appendix C.2. □

4 Simple Regret in the Data-Poor Regime

Despite the optimality of SH shown in the last section, it requires at least $\tilde{\Theta}\big(n/\epsilon^2\big)$ samples. This requirement is unacceptable when the size of the instance is very large, especially for the situation where n is large, or even infinity, while the fraction of the ϵ-good arms is kept constant. This setup is commonly referred to as the data-poor regime where one would like to identify a good arm even before the algorithm samples every arm once.

To cope with the data-poor regime, we take inspiration from BUCB [16] and propose Bracketing SH (BSH) that enjoys a similar guarantee as DSH but enjoys nonvacuous sample complexity in the data-poor regime. To understand the bracketing trick, suppose that we uniformly sample a subset of size n/m $(0 < m \le n)$ from the entire arm set $[n]$. With constant probability, this subset includes at least one of the top m arms. By applying any pure exploration algorithm to this subset, we expect to find the best arm within the subset with the sample complexity that scales with n/m rather than n. Such a subset is called a *bracket*. Note that there is a natural trade-off here with the size of the bracket. If the size is large, then we are likely to include many good arms, but the sample complexity of identifying a good arm becomes large. On the other hand, if the size is small, then we are not likely to include any good arm, although the sample complexity of identifying a good arm within the bracket is small. As we show later, one can precisely work out such a tradeoff mathematically and find the best size of the bracket. The challenge is, however, that the best bracket size typically requires knowledge of the number of good arms.

To avoid requiring such knowledge, BSH adopts the bracketing technique of Katz-Samuels and Jamieson [16] that progressively creates a larger and larger bracket size as the time step t gets larger while invoking a base algorithm for each bracket in a parallel manner. Specifically, BSH uses DSH as the base algorithm. At each time step, BSH takes in the estimated best arms and their empirical means from all the brackets and outputs the one with the largest empirical mean. We summarize BSH in Algorithm 3 where the operator $U\big([n],k\big)$ samples k with replacement from $[n]$ (uniformly at random). If $k \ge n$, $U([n],k)$ returns $[n]$. To define terminology, we say a new bracket is *opened* when a new bracket is sampled. The bracket-opening schedule is set such that the B-th bracket is opened at time step $(B-1)\cdot 2^{B-1}$. The arm pulls between round $(B-1)\cdot 2^{B-1}$ and $B\cdot 2^B$ are equally allocated to the opened brackets (total B of them). We say an arm *represents* bracket A_B at round t if it is the output returned by the DSH on bracket A_B at round t.

Let us first show the properties of the bracketing technique. The bracketing design ensures the diversity of the size of opened brackets. Specifically, the smallest bracket has 2 arms and the largest bracket has $\Theta(t)$ arms. What is more, it also ensures at anytime all the opened brackets have received an (order-wise) equal amount of sampling budget.

We now present the ϵ-error probability bound of BSH.

Theorem 3. *Let* $m(\epsilon/2) = \Big|\big\{i \in [n] : \mu_i \ge \mu_1 - \epsilon/2\big\}\Big|$*. Then, for any* $\epsilon \in (0,1)$*, the* ϵ*-error probability of BSH satisfies*

$$\mathbb{P}\big(\mu_{J_t} < \mu_1 - \epsilon\big) \le \exp\left(-\tilde{\Theta}\left(\min\left\{\max_{i\in[m(\epsilon/2)]} \frac{i\Delta^2_{i,m(\epsilon/2)+1}}{n}, \epsilon^2\right\}\frac{t}{\ln t}\right)\right),$$

This implies the (ϵ,δ)*-PAC sample complexity of* $\tilde{\mathcal{O}}\left(\max\left\{\min_{i\in[m(\epsilon/2)]} \frac{n}{i\Delta^2_{i,m(\epsilon/2)+1}}, \frac{1}{\epsilon^2}\right\}\log\big(\frac{1}{\delta}\big)\right)$.

Algorithm 3: Bracketing SH (BSH)

Input: arms: $[n]$
Initialize: $t = 1$, $B = 0$, $b_1 = 1$. Define $(I(\mathcal{D}), J(\mathcal{D}), M(\mathcal{D}))$ to be the arm to be pulled next, the current estimated best arm, and its empirical mean from an algorithm $\mathcal{D}$, respectively.
for $t = 1, 2, 3 \dots$ **do**
  if $t \geq B2^B$ **then**
    $B \leftarrow B + 1$
    Sample a bracket $A_B = U\big([n], n \vee 2^B\big)$
    Initialize a new instance of DSH, denoted by $\mathcal{D}_B$, with the bracket A_B.
  Pull arm $I_t = I(\mathcal{D}_{b_t})$
  Receive a reward R_t from the environment and send R_t to $\mathcal{D}_{b_t}$
  Output $J_t = J(\mathcal{D}_{\hat{b}})$ where $\hat{b} = \arg\max_{b \in [B]} M(\mathcal{D}_b)$
  $b_{t+1} = \begin{cases} b_t + 1 & \text{if } b_t < B \\ 1 & \text{otherwise} \end{cases}$

The key difference between Corollary 1.1 and Theorem 3 is the latter does not require the number of samples to be at least $\tilde{\Theta}\big(n/\epsilon^2\big)$. We provide a sketch of the proof to explain the intuition behind this result and defer the complete proof to the appendix.

Proof. Let $t' = t/\ln(t)$. Due to the doubling nature of the bracketing scheme, it is not hard to see that each bracket receives an order-wise equal amount of sampling budget $\Theta(t')$ (details in the appendix C.2). Our analysis is centered around finding the ideal bracket whose representative arm is expected to be of the highest quality, which we called the *best bracket*. Let L be the size of the best bracket, which will be specified later. On the one hand, if the best bracket is too lager, the per arm sampling budget will be too small to guarantee a meaningful output. On the other hand, if the best bracket is too small, it may not include a sufficient number of good arms from the subsampling. The core idea of this proof is to well balance this trade-off. Let $i \in [m(\epsilon/2)]$ be a free parameter to be tuned later. Define the following events:

- E_1: the number of arms with mean reward at least μ_i included in the best bracket is at least j where j will be specified later.
- E_2: an $\epsilon/2$-good arm represents the best bracket.

Then the error probability of Bracketing SH can be expressed as follows,

$$\begin{aligned}\mathbb{P}\big(\mu_{J_t} < \mu_1 - \epsilon\big) &= \mathbb{P}\big(\mu_{J_t} < \mu_1 - \epsilon, E_1^c\big) + \mathbb{P}\big(\mu_{J_t} < \mu_1 - \epsilon, E_1, E_2^c\big) + \mathbb{P}\big(\mu_{J_t} < \mu_1 - \epsilon, E_1, E_2\big) \\ &\leq \mathbb{P}\big(E_1^c\big) + \mathbb{P}\big(E_1, E_2^c\big) + \mathbb{P}\big(\mu_{J_t} < \mu_1 - \epsilon, E_2\big). \end{aligned} \tag{4}$$

For the proof sketch, we focus on the first two terms as they represent the trade-off we explain before. In the full proof, we show that the third term is bounded by $\exp(-\tilde{\Theta}(\epsilon^2 t'))$. Thus, it remains to bound both the first and the second term by $\exp(-\tilde{\Theta}(\max_{i \in [m(\epsilon/2)]} \frac{i\Delta^2_{i,m(\epsilon/2)+1} t'}{n}))$.

For the first term, event E_1^c can be regarded as the failure of achieving at least j successes in L trials of Bernoulli experiments with parameter i/n. Define $\mathsf{KL}(p, q)$ to be the Kullback–Leibler divergence between two Bernoulli random variables with mean p and q, By the additive Chernoff bound for Bernoulli distribution,

$$\mathbb{P}\big(E_1^c\big) \leq \exp\left(-L \cdot \mathsf{KL}\left(\frac{j}{L}, \frac{i}{n}\right)\right) \overset{(a_1)}{\leq} \exp\left(-\frac{Ln}{2i}\left(\frac{j}{L} - \frac{i}{n}\right)^2\right),$$

where (a_1) is due to $\mathsf{KL}(p, q) \geq \frac{(p-q)^2}{2\max\{p,q\}}$ and with the foresight that our choice of j, L will satisfy $j/L \leq i/n$.

For the second term, denote by $\mu_{(j)}$ the mean reward of the j-th best arm within the best bracket. By Corollary 1.1,

$$\mathbb{P}\big(E_1, E_2^c\big) \leq \mathbb{P}\big(E_2^c \mid E_1\big) \leq \mathbb{P}\big(\mu_{J_t} < \mu_{(j)} - \Delta_{i,m(\epsilon/2)+1} \big| E_1\big) \leq \exp\left(-\tilde{\Theta}\left(j \frac{\Delta^2_{i,m(\epsilon/2)+1} t'}{L}\right)\right).$$

The best bracket is expected to achieve the balance between the two terms of the following summation on the right-hand side,

$$\mathbb{P}\left(E_1^c\right) + \mathbb{P}\left(E_1, E_2^c\right) \leq \exp\left(-\tilde{\Theta}\left(j\frac{\Delta^2_{i,m(\epsilon/2)+1}t'}{L}\right)\right) + \exp\left(-\frac{Ln}{2i}\left(\frac{j}{L} - \frac{i}{n}\right)^2\right).$$

By taking $L = \Delta^2_{i,m(\epsilon/2)+1}t', j = \frac{i\Delta^2_{i,m(\epsilon/2)+1}t'}{2n}$, we have,

$$\mathbb{P}\left(E_1^c\right) \leq \exp\left(-\tilde{\Theta}\left(\frac{i\Delta^2_{i,m(\epsilon/2)+1}t'}{n}\right)\right) \quad \text{and} \quad \mathbb{P}\left(E_1, E_2^c\right) \leq \exp\left(-\tilde{\Theta}\left(\frac{i\Delta^2_{i,m(\epsilon/2)+1}t'}{n}\right)\right).$$

We have the final result by considering all possible i because $i \in [m(\epsilon/2)]$ is a free parameter. □

To show the effectiveness of Theorem 3, we now discuss the implication of our result and compare it with BUCB by Katz-Samuels and Jamieson [16] using their performance measure called unverifiable (ϵ, δ) sample complexity.

Definition 1 (unverifiable sample complexity)**.** *Katz-Samuels and Jamieson [16]. For an algorithm π and an instance ρ. Let $\tau_{\epsilon,\delta}$ be a stopping time such that*

$$\mathbb{P}\left(\forall t \geq \tau_{\epsilon,\delta} : \mu_{J_t} > \mu_1 - \epsilon\right) \geq 1 - \delta.$$

Then $\tau_{\epsilon,\delta}$ is called (ϵ, δ)-unverifiable sample complexity of the algorithm with respect to ρ.

The unverifiable sample complexity indicates how many samples are sufficient for an algorithm to begin to output an ϵ-good arm with high probability. Compared with the verifiable sample complexity, it does not require the algorithm to verify the output is ϵ-good. More discussion can be found in Katz-Samuels and Jamieson [16]. Our sample complexity in Theorem 3 is naturally compatible with the unverifiable sample complexity. We discuss how the sample complexity w.r.t. the ϵ-error probability of an anytime algorithm can be converted to the unverifiable sample complexity in the appendix. As a notion that holds for any ϵ, it is appropriate for being interpreted as a high probability simple regret.

To make explicit comparisons, we turn to specific problem instances and show the sample complexity bounds for BUCB Katz-Samuels and Jamieson [16, Theorem 7] and BSH; the proof can be found in the appendix.

Corollary 3.1. *Consider the equal gap instance (3). The Bracketing UCB achieves an expected unverifiable sample complexity as*

$$\mathbb{E}\left[\tau_{\epsilon,\delta}\right] \leq \tilde{\mathcal{O}}\left(\frac{n}{\epsilon^2 m}\log\left(\frac{1}{\delta}\right)\right).$$

In this instance, Bracketing SH achieves an unverifiable sample complexity as, with probability $1-\delta$,

$$\tau_{\epsilon,\delta} \leq \tilde{\mathcal{O}}\left(\frac{n}{\epsilon^2 m}\log\left(\frac{1}{\delta}\right)\right).$$

In this instance, BUCB and BSH have the nearly same performance guarantees except that BUCB's sample complexity is stated as an expected sample complexity. Interestingly, Katz-Samuels and Jamieson [16, Section E] remark that their attempt to derive a high-probability bound resulted in the factor of $\ln^2(1/\delta)$, which we speculate to be an artifact of analysis.

Scaling with n/m instead of n reveals the merit of the unverifiable sample complexity that were not achieved by the algorithms designed to achieve verifiable sample complexity such as Median Elimination [11].

Corollary 3.2. *Consider the polynomial instance; i.e., $\Delta_i = \left(\frac{i}{n}\right)^\alpha$ with $\alpha > 0$. For any $\epsilon \in (0,1)$, let $\hat{\tau}_{\epsilon,\delta}$ be the upper bound of the expected unverifiable sample complexity reported in Katz-Samuels and Jamieson [16, Theorem 7] for Bracketing UCB. Then, BUCB satisfies*

$$\mathbb{E}\left[\tau_{\epsilon,\delta}\right] \leq \hat{\tau}_{\epsilon,\delta} = \tilde{\Theta}\left(\epsilon^{-\frac{2\alpha-1}{\alpha}} n \log\left(\frac{1}{\delta}\right)\right).$$

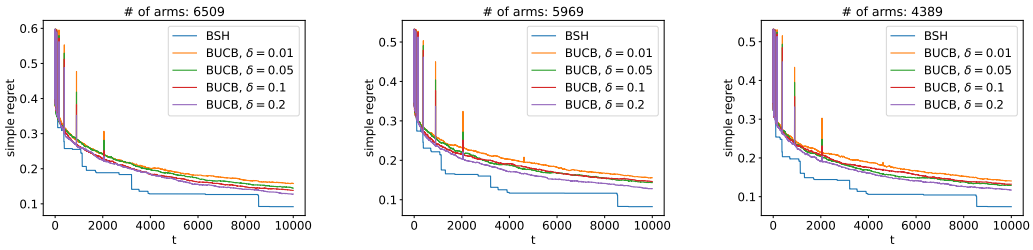

Figure 1: The simple regret comparison between Bracketing UCB and Bracketing SH for the New Yorker Cartoon Caption contest 780, 781, 782.

On the other hand, BSH satisfies, with probability $1-\delta$*,*

$$\tau_{\epsilon,\delta} \leq \tilde{\mathcal{O}}\left(\epsilon^{-\frac{2\alpha+1}{\alpha}} \log\left(\frac{1}{\delta}\right)\right).$$

The unverifiable sample complexity of Bracketing SH does not scale with the instance size n linearly, unlike BUCB. For a large scale instance instance, Bracketing SH provides a stronger guarantee over Bracketing UCB. On the other hand, when n is not too large yet ϵ is small, BUCB shows a favorable sample complexity.

5 Experiments

We test Bracketing SH on a real-world dataset called The New Yorker Cartoon Caption Contest [1]. For each cartoon, the editors of The New Yorker collect the evaluation score of n captions from the participants. Specifically, upon arrival of a participant, the algorithm sequentially shows a number of captions and receives the evaluation score of "unfunny", "somewhat funny", or "funny". Following Katz-Samuels and Jamieson [16], we use the proportion of times the score was "somewhat funny" or "funny" as the ground truth mean reward for each caption. We then set the reward distribution as the Bernoulli. We choose the contest 780, 781, and 782 that contains 6509 arms, 5969 arms, and 4389 arms, respectively. As we are especially interested in the data-poor regime, we report the performance of the algorithms up to time step 10,000, which is only about twice larger than the number of arms in our dataset. We run BSH and BUCB with various δ choices to see the best version of BUCB, which was repeated 500 times with a different random seed. We use a desktop with AMD Ryzen 5 PRO 4650GE CPU and 16GB RAM to conduct the experiment, which takes two hours to produce each plot. We summarize the results in Figure 1 where BSH is a clear winner over BUCB. Specifically, in the contest 781, at time step $4,000$, Bracketing UCB achieves simple regret about 0.18 for $\delta = 0.2$, while Bracket SH achieves simple regret about 0.11, amounting to 39% improvement. To our knowledge, the dataset we have used does not contain any personally identifiable information or offensive content.

6 Conclusion

Obtaining a uniform ϵ-error probability bound is precisely a way to characterize the distribution function of Δ_{J_T} induced by a particular algorithm, based on which we believe that a uniform ϵ-error probability bound is a fundamental quantity for any measures for identifying an ϵ-good arm.

As such, we believe our paper opens up numerous exciting open problems. First, our bound is closer to the minimax nature except for the dependence on the number of good arms $m(\epsilon)$. We like to push it towards a fully instance-dependent sample complexity bound and study instance optimality. Second, SH does not achieve an optimal (ϵ, δ)-PAC sample complexity (due to logarithmic factors). On the other hand, as we report in our appendix, adjusting its sample allocation scheme does achieve the optimal rate, but it does not seem to achieve the usual near-optimal instance-dependent sample complexity of $\tilde{\Theta}(\sum_i \Delta_i^{-2})$. We wonder if a modified SH or other algorithms can be optimal for both sample complexity measures simultaneously. Third, another practical pure exploration algorithm is track-and-stop by Garivier and Kaufmann [12], which is also parameter-free if we only take its sampling strategy and not the stopping strategy. It would be interesting to investigate if track-and-stop achieves similar guarantees as DSH.

We remark that, to our knowledge, there is no negative societal impact of our paper.

Acknowledgements

Kwang-Sung Jun and Yao Zhao was supported by the Eighteenth Mile TRIF Funding from Research, Innovation & Impact at University of Arizona.

References

[1] https://github.com/nextml/caption-contest-data.

[2] Jean-Yves Audibert, Sébastien Bubeck, et al. Minimax policies for adversarial and stochastic bandits. In *Annual Conference on Computational Learning Theory (COLT)*, volume 7, pages 1–122, 2009.

[3] Maryam Aziz, Jesse Anderton, Emilie Kaufmann, and Javed Aslam. Pure exploration in infinitely-armed bandit models with fixed-confidence. In *Algorithmic Learning Theory*, pages 3–24. PMLR, 2018.

[4] Tavor Baharav and David Tse. Ultra fast medoid identification via correlated sequential halving. *Advances in Neural Information Processing Systems (NeurIPS)*, 32, 2019.

[5] Robert E Bechhofer. A Sequential Multiple-Decision Procedure for Selecting the Best One of Several Normal Populations with a Common Unknown Variance, and Its Use with Various Experimental Designs. *Biometrics*, 14(3):408–429, 1958.

[6] Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. Pure exploration in multi-armed bandits problems. In *International conference on Algorithmic learning theory*, pages 23–37. Springer, 2009.

[7] Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. Pure exploration in finitely-armed and continuous-armed bandits. *Theoretical Computer Science*, 412(19):1832–1852, apr 2011. URL https://hal-hec.archives-ouvertes.fr/hal-00609550.

[8] Arghya Roy Chaudhuri and Shivaram Kalyanakrishnan. PAC identification of a bandit arm relative to a reward quantile. In *Thirty-First AAAI Conference on Artificial Intelligence*, pages 1777–1783, 2017.

[9] Arghya Roy Chaudhuri and Shivaram Kalyanakrishnan. PAC identification of many good arms in stochastic multi-armed bandits. In *International Conference on Machine Learning (ICML)*, pages 991–1000. PMLR, 2019.

[10] Herman Chernoff. Sequential design of experiments. *The Annals of Mathematical Statistics*, 30 (3):755–770, 1959.

[11] Eyal Even-Dar, Shie Mannor, Yishay Mansour, and Sridhar Mahadevan. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of machine learning research*, 7(6), 2006.

[12] Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *Proceedings of the Conference on Learning Theory (COLT)*, pages 998–1027, 2016.

[13] Avinatan Hassidim, Ron Kupfer, and Yaron Singer. An optimal elimination algorithm for learning a best arm. *Advances in Neural Information Processing Systems (NeurIPS)*, 33: 10788–10798, 2020.

[14] K.-S. Jun and R. Nowak. Anytime exploration for multi-armed bandits using confidence information. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 48, pages 974–982, 2016. ISBN 9781510829008.

[15] Zohar Karnin, Tomer Koren, and Oren Somekh. Almost optimal exploration in multi-armed bandits. In *International Conference on Machine Learning (ICML)*, pages 1238–1246. PMLR, 2013.

[16] Julian Katz-Samuels and Kevin Jamieson. The true sample complexity of identifying good arms. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 108, pages 1781–1791, 2020.

[17] Finnian Lattimore, Tor Lattimore, and Mark D Reid. Causal bandits: Learning good interventions via causal inference. *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

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

[19] Lisha Li, Kevin Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, and Ameet Talwalkar. Hyperband: A Novel Bandit-Based Approach to Hyperparameter Optimization. *Journal of Machine Learning Research*, 18(185):1–52, 2018.

[20] Shie Mannor and John N Tsitsiklis. The sample complexity of exploration in the multi-armed bandit problem. *Journal of Machine Learning Research*, 5(Jun):623–648, 2004.

[21] Ervin Tanczos, Robert Nowak, and Bob Mankoff. A KL-LUCB algorithm for Large-Scale Crowdsourcing. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5894–5903. 2017.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No] Implementation is straightforward and details are sufficiently provided.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] We repeated 500 times.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [N/A] The source of dataset does not mention any license.
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] The data was on a public GitHub repo.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Appendix

Table of Contents

A Related work

Pure exploration. Pure exploration has a broad position in several close related research directions including multi-armed bandits [1, 21, 6, 3, 21, 16, 14, 11, 28, 15] and reinforcement learning [13, 2, 24]. Even more, the Monte-Carlo tree search with tree depth 1 is also a pure exploration problem [25]. Even we only consider multi-armed bandits, there are various types of problems have been formulated. Our investigation focus on the most traditional K-armed bandits model. But we note here pure exploration has also been studied in the structured bandits like kernel bandits [5] and linear bandits [29, 30].

Started from [13, 26], pure exploration in K-armed bandits is studied with the celebrated (ϵ, δ)-PAC framework. [26] shows a lower bound. The upper bound of Median Elimination in [13] is proven to match with this lower bound. The recent work of [15] pushes forward and shows that it is possible to achieve $\frac{n}{2\epsilon^2}\log\frac{1}{\delta}$ asymptotically, where the constant factor of $1/2$ is the optimal one. [3] studies (ϵ, δ)-PAC for the infinitely-armed bandit model, where they model the infinitely-armed setting as the arm reservoir. Among them, [13, 15] focus on the worst-case guarantee. [3] provides more instance-dependent bounds.

Best arm identification. The best arm identification is the factually dominating one among many approaches for studying the pure exploration problem. Specifically, there are two problem setups of the best arm identification, fix budget [1, 21, 6] and fix confidence [21, 16, 14, 11, 28]. In the fixed budget setting, the algorithm takes a budget as the input, and is required to return an output before exhausting all the sample budget. [1] is the first one opened up this direction and proposes Successive Reject algorithm, which achieves the optimality up to logarithmic factors. The follow-up work of [21] proposes SH algorithm, which is also an elimination based algorithm as Successive Reject, but it eliminates empirically bad arms more aggressively. It is until [6] we know that SH is optimal up to doubly-logarithmic factors.

There are lots of fruitful results have been produced for the fixed confidence setting. Perhaps [14] is the first one that claims the meaningful optimality. They propose a non-asymptotic lower bound for the sample complexity, and also shows an algorithm called Track-and-Stop that matches the lower bound asymptotically when δ goes to 0. There still exists a computation issue for Track-and-Stop algorithm as it needs to solve an optimization problem, which requires a nested binary search.

Simple regret. Though simple regret is the first proposed concrete measure for pure exploration and many works claim simple regret as their target, perhaps only [22, 7] truly deal with simple regret directly. The reason that the results from other works do not apply to simple regret is they usually predefine a target threshold, simple regret of 0 for best arm identification or simple regret of ϵ for (ϵ, δ)-PAC identification. However simple regret itself is a parameter-free notion. In that sense, the algorithms of [22, 7] work for the simple regret minimization problem since they do not take such a predefined threshold as the input. There is another different problem setup, in which the algorithm is required to output one of the top-m arms [8, 9]. Though it also characterizes how close the output is to the best one, it still needs m as the predefined threshold.

(ϵ, δ)-Verifiable and unverifiable sample complexity. The fixed confidence best arm identification shares some similarities with the (ϵ, δ)-PAC identification. For both settings, the algorithm is required to verify the output can meet the corresponding criteria, simple regret of 0 or simple regret of ϵ, with error probability δ. The requirement of verifiability enforces the algorithm to pull each arm at least once. Therefore the (ϵ, δ)-PAC is inherently impossible for the data-poor regime, e.g. $t < n$. [22] argues (ϵ, δ)-PAC is not aligned with the practical usage. They propose the (ϵ, δ)-unverifiable sample complexity, which does not require the algorithm to verify the output is ϵ-good. Instead, the (ϵ, δ)-unverifiable sample complexity represents the number of samplings that the algorithm minimally needs such that it has the ability to output an ϵ-good arm. Note the sample complexity of a fixed budget algorithm, (or to be more accurate, the doubling trick version of the fixed budget algorithm) also captures the nature of the (ϵ, δ)-unverifiable sample complexity. In reinforcement learning setting, [12] also discuss the limitation of (ϵ, δ)-PAC and propose a notion called Uniform-PAC.

Top-k arm identification. The top-k arm identification problem requires the algorithm to return k arms with the highest mean reward, instead of only the best one. [19, 20, 10, 28]. The seminal work is [19]. Their problem formulation is a direct extension of the (ϵ, δ)-PAC identification problem. The goal is to return k arms that have mean rewards no less than $\mu_k - \epsilon$. The worst-case lower bound is shown in [20]. The first instance-dependent lower bound for the k-identification problem is given by [28, 10] almost at the same time, together with the near-optimal algorithms. The best arm identification can be regarded as a special case of the top-k arm identification problem. The similarity is that we are both interested in good arms, but we are measuring the likelihood of identifying one arm out of the top-k rather than the set of top-k arms.

B Improved Analyses of Sequential Halving

B.1 Implication of Corollary 1.1 for $m = 1$

The following theorem corresponds to the implication of $m = 1$ of Corollary 1.1. However we prove it independently here.

Theorem 4. *For any $\epsilon \in (0, 1)$, the error probability of SH for identifying an ϵ-good arm satisfies,*

$$\mathbb{P}\left(\mu_{J_T} < \mu_1 - \epsilon\right) \le 3 \log_2 n \cdot \exp\left(-\frac{\epsilon^2}{32} \frac{T}{n \log_2 n}\right).$$

Proof. To avoid redundancy and for the sake of readability, we assume n is of a power of 2. It is easy to verify the result for any n. Let $\epsilon_1 = \epsilon/4$, $T' := \frac{T}{\log_2 n}$. And define $\epsilon_{\ell+1} = \frac{3}{4} \cdot \epsilon_\ell$. For each stage ℓ, define the event G_ℓ as

$$G_\ell := \left\{ \max_{i \in S_{\ell+1}} \mu_i \ge \max_{i \in S_\ell} \mu_i - \epsilon_\ell \right\}.$$

Thus as long as $\bigcap_{\ell=1}^{\log_2 n} G_\ell$ happens, we have that the arm returned after the final stage is an ϵ-good arm, because

$$\sum_{\ell=1}^{\log_2 n} \epsilon_\ell < \sum_{\ell=1}^{\infty} \left(\frac{3}{4}\right)^{\ell-1} \cdot \epsilon_1 = \frac{\epsilon}{4} \sum_{\ell=1}^{\infty} \left(\frac{3}{4}\right)^{\ell-1} \leq \frac{\epsilon}{4} \lim_{n\to\infty} \frac{1-(3/4)^n}{1-3/4} = \epsilon.$$

Thus, by a union bound,

$$\begin{aligned}\mathbb{P}\big(\mu_{J_T} < \mu_1 - \epsilon\big) &\leq \mathbb{P}\left(\left(\bigcap_{\ell=1}^{\log_2 n} G_\ell\right)^c\right) \\ &\leq \sum_{\ell=1}^{\log_2 n} \mathbb{P}\big(G_\ell^c\big). \end{aligned} \tag{5}$$

Let a_ℓ be the best arm in S_ℓ,

$$\begin{aligned}\mathbb{P}\big(G_\ell^c\big) &= \mathbb{P}\big(G_\ell^c, \hat{\mu}_{a_\ell} < \mu_{a_\ell} - \epsilon_\ell/2\big) + \mathbb{P}\big(G_\ell^c, \hat{\mu}_{a_\ell} \geq \mu_{a_\ell} - \epsilon_\ell/2\big) \\ &\leq \mathbb{P}\big(\hat{\mu}_{a_\ell} < \mu_{a_\ell} - \epsilon_\ell/2\big) + \mathbb{P}\big(G_\ell^c \mid \hat{\mu}_{a_\ell} \geq \mu_{a_\ell} - \epsilon_\ell/2\big). \\ &\leq \exp\left(-\frac{\epsilon_\ell^2}{2}\frac{T'}{|S_\ell|}\right) + \mathbb{P}\big(G_\ell^c \mid \hat{\mu}_{a_\ell} \geq \mu_{a_\ell} - \epsilon_\ell/2\big).\end{aligned}$$

For the second term,

$$\begin{aligned}\mathbb{P}\big(G_\ell^c \mid \hat{\mu}_{a_\ell} \geq \mu_{a_\ell} - \epsilon_\ell/2\big) &\leq \mathbb{P}\Big(\big|\{i \in S_\ell \mid \hat{\mu}_i > \mu_i + \epsilon_\ell/2\}\big| \geq |S_\ell|/2\Big) \\ &\overset{(a_1)}{\leq} \frac{\mathbb{E}\Big[\big|\{i \in S_\ell \mid \hat{\mu}_i > \mu_i + \epsilon_\ell/2\}\big|\Big]}{|S_\ell|/2} \\ &\leq \frac{|S_\ell| \exp\left(-\frac{\epsilon_\ell^2}{2}\frac{T'}{|S_\ell|}\right)}{|S_\ell|/2} \\ &= 2\exp\left(-\frac{\epsilon_\ell^2}{2}\frac{T'}{|S_\ell|}\right).\end{aligned}$$

For (a_1), we use Markov's inequality. Then,

$$\mathbb{P}\big(G_\ell^c\big) \leq 3\exp\left(-\frac{\epsilon_\ell^2}{2}\frac{T'}{|S_\ell|}\right) = 3\exp\left(-\left(\frac{9}{16}\right)^{\ell-1}\frac{\epsilon^2}{32}\frac{T'}{2^{-(\ell-1)}n}\right) = 3\exp\left(-\left(\frac{9}{8}\right)^{\ell-1}\frac{\epsilon^2}{32}\frac{T'}{n}\right).$$

Taking the above into (5), we have

$$\begin{aligned}\mathbb{P}\big(\mu_{J_T} < \mu_1 - \epsilon\big) &\leq \sum_{\ell=1}^{\log_2 n} \mathbb{P}\big(G_\ell^c\big) \\ &\leq \sum_{\ell=1}^{\log_2 n} 3\exp\left(-\left(\frac{9}{8}\right)^{\ell-1}\frac{\epsilon^2}{32}\frac{T'}{n}\right) \\ &\leq 3\log_2 n \cdot \exp\left(-\frac{\epsilon^2}{32}\frac{T}{n\log_2 n}\right).\end{aligned}$$

□

B.2 Proof of Corollary 1.2

We state the following result that includes two different budget allocation schemes for SH. The appendix B.6 should be read first where we define the two budget allocation schemes:

$$\text{Option 1: } T_\ell = \left\lfloor \frac{T}{|S_\ell|\lceil \log_2 n \rceil} \right\rfloor, \quad \text{Option 2: } T_\ell = \left\lfloor \frac{T}{81n} \cdot \left(\frac{16}{9}\right)^{\ell-1} \cdot \ell \right\rfloor.$$

Corollary 1.2. *The simple regret of SH satisfies*

$$\textit{Option 1: } \mathsf{SReg}_T \le \mathcal{O}\left(\sqrt{\frac{n\log^3 n}{T}}\right), \quad \textit{Option 2: } \mathsf{SReg}_T \le \mathcal{O}\left(\sqrt{\frac{n}{T}}\right).$$

Proof. The simple regret can be calculated by considering the following integral with respect to ϵ. We use SH with Option 2. For Option 1, the same analysis holds.

$$\mathsf{SReg}_T = \int_0^\infty \mathbb{P}\big(\mu_1 - \mu_{J_T} > \epsilon\big)\,\mathrm{d}\epsilon = \int_0^\infty \exp\left(-\epsilon^2\frac{T}{n}\right)\mathrm{d}\epsilon.$$

We can borrow the result from Gaussian distribution with mean 0 and variance σ^2,

$$\int_0^\infty \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{x^2}{2\sigma^2}\right)\mathrm{d}x = \frac{1}{2}.$$

Taking $\frac{1}{2\sigma^2} = \frac{T}{n}$, we have

$$\int_0^\infty \exp\left(-\epsilon^2\frac{T}{n}\right)\mathrm{d}\epsilon = \frac{\sigma\sqrt{2\pi}}{2} = \frac{\sqrt{\pi}}{2}\sqrt{\frac{n}{T}}.$$

□

B.3 Proof of Corollary 1.1

Corollary 1.1 is in fact an equivalent result to Theorem 1, so proof of each implies the other. While Theorem 1 takes a form that is easier to instantiate the bound for specific instances, Corollary 1.1 takes a form that is easier to prove. Thus, we choose to prove Corollary 1.1 directly, leaving the proof of Theorem 1 as a consequence of Corollary 1.1.

Let us first provide an intuitive explanation. Imagine one possible 'typical' scenario where, in each stage, the set of surviving arms for the next stage happens to maintain the fraction of good arms as at least m/n. This means that, after $\Theta(\log_2 m)$ stages, we expect to have at least one good arm in the surviving arm set. The number of surviving arms at that time is around n/m. At this point, the rest of the procedure can be analyzed by the standard analysis of SH where the goal is to find the arm that is the best in the current surviving arm set. Thus, it remains to bound how likely it is to have the 'typical' event above. That is, we like to bound the failure probability of this event at each stage as tightly as possible. Bounding this failure probability is our key technical innovation (Proposition 5), which promotes the fast rate of the failure probability that improves as a function of the number of good arms. Another key technical step of the proof is a careful design of events that would lead to guaranteeing the suboptimality of the chosen arm in the last stage to be at most ϵ, which requires splitting ϵ into smaller pieces to be distributed over the stages.

Corollary 1.1. *Suppose we run SH. Then, for any $m \le n$ and any $\epsilon \in (0,1)$,*

$$\mathbb{P}\big(\mu_{J_T} < \mu_m - \epsilon\big) \le \log_2 n \cdot \exp\left(-\text{const}\cdot m\left(\frac{\epsilon^2 T}{4n\log_2^2(2m)\log_2 n} - \ln(4e)\right)\right).$$

Thus, there exist a positive constant c_1 such that for $T \ge c_1\frac{(\ln(4e)+\ln\ln n)n\log_2^2(2m)\log_2 n}{\epsilon^2} = \tilde{\Theta}\big(\frac{n}{\epsilon^2}\big)$,

$$\mathbb{P}\big(\mu_{J_T} < \mu_m - \epsilon\big) \le \exp\left(-\tilde{\Theta}\left(m\frac{\epsilon^2 T}{n}\right)\right),$$

where, $\tilde{\Theta}$ means ignoring the logarithmic factors of m, n and constants.

Proof. Let us consider the case of $m \le n/2$ first.

Let $\ell^* = \lceil\log_2 m\rceil$, $\epsilon' = \frac{\epsilon}{2\log_2 m}$, $T' := \left\lfloor\frac{T}{\log_2 n}\right\rfloor$. And g_ℓ denotes the number of $(m, \ell\cdot\epsilon')$-good arms after finishing stage $\ell \in \left[\lceil\log_2 m\rceil\right]$. For stage ℓ, we define the event G_ℓ as,

$$G_\ell := \left\{g_\ell \ge 2^{-\ell}\cdot m\right\}.$$

Specifically, we have G_{ℓ^*} to be the event that the number of $(m, \epsilon/2)$-good arms after finishing stage ℓ^* is at least 1. We define G_{ℓ^*+1} as the algorithm succeeds to return an arm in $\text{Top}_m(\epsilon)$,

$$G_{\ell^*+1} := \left\{\mu_{J_T} \geq \mu_m - \epsilon\right\}.$$

Then the event $\bigcap_{\ell=1}^{\ell^*+1} G_\ell$ is a possible path the algorithm returns an arm in $\text{Top}_m(\epsilon)$ in the end. Thus the probability of missing all of the (m, ϵ)-good arms can be upper bounded as follows. Define $F_\ell = \left(\bigcap_{i=1}^{\ell-1} G_i\right) \cap G_\ell^c$ for $\ell > 1$, $F_1 = G_1^c$. Since $\bigcap_{\ell=1}^{\ell^*+1} G_\ell \subset G_{\ell^*+1}$ implies $G_{\ell^*+1}^c = \left\{\mu_{J_T} < \mu_m - \epsilon\right\} \subset \left(\bigcap_{\ell=1}^{\ell^*+1} G_\ell\right)^c$,

$$\begin{aligned}
\mathbb{P}\left(\mu_{J_T} < \mu_m - \epsilon\right) &\leq \mathbb{P}\left(\left(\bigcap_{\ell=1}^{\ell^*+1} G_\ell\right)^c\right) \\
&= \mathbb{P}\left(\bigcup_{\ell=1}^{\ell^*+1} F_\ell\right) \\
&\overset{(a_1)}{\leq} \sum_{\ell=1}^{\ell^*+1} \mathbb{P}\left(G_\ell^c \mid \bigcap_{i=1}^{\ell-1} G_i\right) \\
&= \sum_{\ell=1}^{\ell^*} \mathbb{P}\left(G_\ell^c \mid \bigcap_{i=1}^{\ell-1} G_i\right) + \mathbb{P}\left(G_{\ell^*+1}^c \mid \bigcap_{i=1}^{\ell^*} G_i\right). \qquad (6)
\end{aligned}$$

For (a_1), we use $\mathbb{P}(A, B) \leq \mathbb{P}(A|B)$.

For the first term of (6), we apply the result of Proposition 5 to each stage of SH with the parameters therein as $k = 2^{-\ell}\cdot m$, $s = 2^{-\ell}\cdot n$, $m' = 2^{-\ell+1}\cdot m$, $\epsilon = \epsilon'$, and $T = T'$. Thus, for stage ℓ,

$$\begin{aligned}
&\mathbb{P}\left(G_\ell^c \mid G_{\ell-1}\right) \\
&\leq \binom{2^{-\ell+1}\cdot m}{2^{-\ell}\cdot m} \exp\left(-2^{-\ell}\cdot m \frac{\epsilon'^2 T'}{2\cdot 2^{-(\ell-1)}\cdot n}\right) + \binom{2^{-\ell+1}\cdot n}{2^{-\ell}\cdot (n-m)} \exp\left(-2^{-\ell}\cdot (n-m) \frac{\epsilon'^2 T'}{2\cdot 2^{-(\ell-1)}\cdot n}\right) \\
&\overset{(a_3)}{\leq} h_{\ell,1} \exp\left(-2^{-\ell}\cdot m \frac{\epsilon'^2 T'}{2\cdot 2^{-(\ell-1)}\cdot n}\right) + h_{\ell,2} \exp\left(-2^{-\ell}\cdot (n-m) \frac{\epsilon'^2 T'}{2\cdot 2^{-(\ell-1)}\cdot n}\right) \\
&\leq h_{\ell,1} \exp\left(-\frac{m}{2}\frac{\epsilon'^2 T'}{2n}\right) + h_{\ell,2} \exp\left(-\frac{n-m}{2}\frac{\epsilon'^2 T'}{2n}\right) \\
&\overset{(a_4)}{\leq} h_{\ell,1} \exp\left(-\frac{m}{2}\frac{\epsilon'^2 T'}{2n}\right) + h_{\ell,2} \exp\left(-\frac{n}{4}\frac{\epsilon'^2 T'}{2n}\right) \qquad (7) \\
&\leq \exp\left(-\frac{m}{2}\frac{\epsilon^2 T}{2n \log_2^2 m \log_2 n} + \ln(2e) 2^{-\ell}\cdot m\right) \\
&\quad + \exp\left(-\frac{n}{4}\frac{\epsilon^2 T}{2n \log_2^2 m \log_2 n} + \ln(4e) 2^{-\ell}\cdot (n-m)\right) \\
&\leq \exp\left(-m\left(\frac{\epsilon^2 T}{4n \log_2^2 m \log_2 n} - 2\ln(4e) 2^{-\ell}\right)\right) \\
&\quad + \exp\left(-\frac{n}{2}\left(\frac{\epsilon^2 T}{4n \log_2^2 m \log_2 n} - 2\ln(4e) 2^{-\ell}\right)\right) \\
&\leq 2\exp\left(-m\left(\frac{\epsilon^2 T}{4n \log_2^2 m \log_2 n} - 2\ln(4e) 2^{-\ell}\right)\right)
\end{aligned}$$

where (a_3) is $\binom{2^{-\ell+1}\cdot m}{2^{-\ell}\cdot m} \leq (2e)^{2^{-\ell}\cdot m} =: h_{\ell,1}$ (Sterling's formula $\binom{x}{y} \leq \left(\frac{ex}{y}\right)^y$) and $\binom{2^{-\ell+1}\cdot n}{2^{-\ell}\cdot (n-m)} \leq (2en/(n-m))^{2^{-\ell}\cdot (n-m)} \leq (4e)^{2^{-\ell}\cdot n} =: h_{\ell,2}$ and (a_4) is by the assumption $m \leq n/2$.

Then we have,

$$\sum_{\ell=1}^{\ell^*} \mathbb{P}\big(G_\ell^c \mid G_{\ell-1}\big) \le \text{const}\cdot \log_2 m \cdot \exp\left(-m\left(\frac{\epsilon^2 T}{4n\log_2^2 m \log_2 n} - \ln(4e)\right)\right). \tag{8}$$

For the second term of (6), $G_{\ell^*+1}^c \mid G_{\ell^*}$ indicates the events where SH fails to return an (m,ϵ)-good arm given the event that there is at least one $(m,\epsilon/2)$-good arm after finishing stage ℓ^*. Note the stages from ℓ^* to the last stage can be regarded as a whole process of SH with a budget $\big(\lceil \log_2 n \rceil - \ell^*\big)T' \ge \text{const}\cdot \log_2\big(\frac{n}{m}\big)\frac{T}{\log_2(n)}$. The initial arms are the surviving arms after finishing stage ℓ^*. Note we have $2^{-\ell^*}\cdot n = n/m$ arms surviving, denoted by S_{ℓ^*}, after finishing stage ℓ^*. Let c be the index of the true best arm in S_{ℓ^*}. Due to the event G_{ℓ^*}, we have $\mu_c \ge \mu_m - \frac{\epsilon}{2}$. Then, by applying the result of Theorem 4,

$$\begin{aligned}
\mathbb{P}\big(G_{\ell^*+1}^c \mid G_{\ell^*}\big) &= \mathbb{P}\big(\mu_{J_T} < \mu_m - \epsilon \mid G_{\ell^*}\big) \\
&\le \mathbb{P}\left(\mu_{J_T} < \mu_c - \frac{\epsilon}{2} \mid G_{\ell^*}\right) \\
&\le 3\log_2\left(\frac{n}{m}\right)\cdot \exp\left(-\frac{m}{128}\frac{\epsilon^2\big(\lceil \log_2 n\rceil - \ell^*\big)T'}{n\log_2\big(\frac{n}{m}\big)}\right) \\
&\le \log_2\left(\frac{n}{m}\right)\cdot \exp\left(-\text{const}\cdot m\frac{\epsilon^2 T}{n\log_2 n}\right).
\end{aligned} \tag{9}$$

Bringing (8),(9) into (6), we have,

$$\mathbb{P}\big(\mu_{J_T} < \mu_m - \epsilon\big) \le \log_2 n \cdot \exp\left(-\text{const}\cdot m\left(\frac{\epsilon^2 T}{4n\log_2^2(m)\log_2 n} - \ln(4e)\right)\right).$$

Note the above analysis is for $m > 1$. To incorporate the result of Theorem 4 for $m = 1$, we rewrite the final formula as

$$\mathbb{P}\big(\mu_{J_T} < \mu_m - \epsilon\big) \le \log_2 n \cdot \exp\left(-\text{const}\cdot m\left(\frac{\epsilon^2 T}{4n\log_2^2(2m)\log_2 n} - \ln(4e)\right)\right).$$

Thus, there exist a positive constant c_1 such that for $T \ge c_1\frac{(\ln(4e)+\ln\ln n)n\log_2^2(2m)\log_2 n}{\epsilon^2} = \tilde{\Theta}\big(\frac{n}{\epsilon^2}\big)$,

$$\mathbb{P}\big(\mu_{J_T} < \mu_m - \epsilon\big) \le \exp\left(-\tilde{\Theta}\left(m\frac{\epsilon^2 T}{n}\right)\right).$$

For the case of $m > n/2$, we consider

$$\mathbb{P}\big(\mu_{J_T} < \mu_m - \epsilon\big) \le \mathbb{P}\big(\mu_{J_T} < \mu_{n/2} - \epsilon\big).$$

Thus, one can repeat the same analysis as above with m replaced by $n/2$. Using $m/2 \le n/2 \le m$, the statement of this theorem holds. □

The result of Corollary 1.1 directly improves the previous works including $\mathcal{O}\Big(\frac{n}{m\epsilon^2}\log^2\big(\frac{1}{\delta}\big)\Big)$ of Chaudhuri and Kalyanakrishnan [8], and $\mathcal{O}\Big(\frac{1}{\epsilon^2}\Big(\frac{n}{m}\log\big(\frac{1}{\delta}\big) + \log^2\big(\frac{1}{\delta}\big)\Big)\Big)$ of Chaudhuri and Kalyanakrishnan [9]. They share a similar strategy that, assuming the knowledge of m, ϵ and δ, uniformly samples $\Theta\Big(\frac{n}{m}\log\big(\frac{1}{\delta}\big)\Big)$ arms first and then runs the Median-Elimination algorithm of [13] or LUCB of Kaufmann and Kalyanakrishnan [23] for this subset.

B.4 Proof of Theorem 1

Theorem 1. *For any $\epsilon \in (0,1)$, the error probability of SH for identifying an ϵ-good arm satisfies,*

$$\mathbb{P}\big(\mu_{J_T} < \mu_1 - \epsilon\big) \leq \min_{\{(m',\epsilon'):\Delta_{m'}+\epsilon'\leq\epsilon\}} \log_2 n \cdot \exp\left(-\text{const}\cdot m'\left(\frac{\epsilon'^2 T}{4n\log_2^2(2m')\log_2 n} - \ln(4e)\right)\right).$$

Proof. Note the result we proved in Corollary 1.1 holds for any $m \leq n$ and any $\epsilon \in (0,1)$. Thus $\mathbb{P}\big(\mu_{J_T} < \mu_1 - \epsilon\big)$ can be upper bounded for any $(m',\epsilon') \in \big\{(m',\epsilon') : \Delta_{m'} + \epsilon' \leq \epsilon\big\}$. The result is therefore implied. □

B.5 Analysis for the uniform sampling

The uniform sampling means the algorithm that takes a budget T as the input and equally allocates the budget to all the arms.

Proposition 5. *Suppose we run the uniform sampling on a n-armed bandit with a budget T and output the s arms with the largest empirical rewards. Let $C_T(s,m',\epsilon)$ be the (m',ϵ)-good arms of the output. Then,*

$$\mathbb{P}\Big(\big|C_T(s,m',\epsilon)\big| \leq k\Big) \leq \binom{m'}{m'-k}\exp\left(-\big(m'-k\big)\frac{\epsilon^2 T}{2n}\right) + \binom{n-\big|\text{Top}_{m'}(\epsilon)\big|}{s-k}\exp\left(-(s-k)\frac{\epsilon^2 T}{2n}\right).$$

Proof. **Step 1**
We claim that the intersection of the two conditions below is a sufficient condition for $\big|C_T(s,m',\epsilon)\big| > k$.

- $\Big|\Big\{i \in \text{Top}_{m'} : \hat{\mu}_i > \mu_i - \frac{\epsilon}{2}\Big\}\Big| \geq k+1$.
- $\Big|\Big\{i \in \text{Top}^c_{m'}(\epsilon) : \hat{\mu}_i < \mu_i + \frac{\epsilon}{2}\Big\}\Big| \geq \big|\text{Top}^c_{m'}(\epsilon)\big| - s + k + 1$.

If all arms in $\Big\{i \in \text{Top}_{m'} : \hat{\mu}_i > \mu_i - \frac{\epsilon}{2}\Big\}$ are included in the output, the claim naturally holds. Thus we focus on the situation where at least one of the arms in $\Big\{i \in \text{Top}_{m'} : \hat{\mu}_i > \mu_i - \frac{\epsilon}{2}\Big\}$ is not included in the output. We prove the claim by contradiction. Suppose $\big|C_T(s,m',\epsilon)\big| \leq k$. Then the output includes no less than $s-k$ non-(m',ϵ)-good arms. Since

$$s - k + \big|\text{Top}^c_{m'}(\epsilon)\big| - s + k + 1 = \big|\text{Top}^c_{m'}(\epsilon)\big| + 1 > \big|\text{Top}^c_{m'}(\epsilon)\big|,$$

at least one arm in $\Big\{i \in \text{Top}^c_{m'}(\epsilon) : \hat{\mu}_i < \mu_i + \frac{\epsilon}{2}\Big\}$ is included in the output. To be more specific, there must exists overlap between $\Big\{i \in \text{Top}^c_{m'}(\epsilon) : \hat{\mu}_i < \mu_i + \frac{\epsilon}{2}\Big\}$ and the set of the non-(m',ϵ)-good arms included in the output, otherwise the union of these two sets is even larger than the set of all the non-(m',ϵ)-good arms we have in the entire arm set. Additionally, all arms in $\Big\{i \in \text{Top}^c_{m'}(\epsilon) : \hat{\mu}_i < \mu_i + \frac{\epsilon}{2}\Big\}$ have less empirical rewards than any arms in $\Big\{i \in \text{Top}_{m'} : \hat{\mu}_i > \mu_i - \frac{\epsilon}{2}\Big\}$. Thus the output includes all arms in $\Big\{i \in \text{Top}_{m'} : \hat{\mu}_i > \mu_i - \frac{\epsilon}{2}\Big\}$ as well. This leads to the contradiction with our supposition that at least one of the arms in $\Big\{i \in \text{Top}_{m'} : \hat{\mu}_i > \mu_i - \frac{\epsilon}{2}\Big\}$ are not included in the output.

Step 2
We apply union bound and Hoeffding's inequality for the event that at least one of the two conditions

does not hold.

$$\begin{aligned}
\mathbb{P}\Big(\big|C_T(s,m',\epsilon)\big| \le k\Big) &\le \mathbb{P}\left(\left|\left\{i \in \mathrm{Top}_{m'} : \hat{\mu}_i > \mu_i - \frac{\epsilon}{2}\right\}\right| < k+1\right) \\
&\quad + \mathbb{P}\left(\left|\left\{i \in \mathrm{Top}^c_{m'}(\epsilon) : \hat{\mu}_i < \mu_i + \frac{\epsilon}{2}\right\}\right| < \left|\mathrm{Top}^c_{m'}(\epsilon)\right| - s + k + 1\right) \\
&= \mathbb{P}\left(\left|\left\{i \in \mathrm{Top}_{m'} : \hat{\mu}_i \le \mu_i - \frac{\epsilon}{2}\right\}\right| \ge m' - k\right) \\
&\quad + \mathbb{P}\left(\left|\left\{i \in \mathrm{Top}^c_{m'}(\epsilon) : \hat{\mu}_i \ge \mu_i + \frac{\epsilon}{2}\right\}\right| \ge s - k\right) \\
&\le \mathbb{P}\left(\exists \mathcal{A} \subset \mathrm{Top}_{m'}, \text{s.t. } |\mathcal{A}| = m' - k \text{ and } \forall i \in \mathcal{A}, \hat{\mu}_i \le \mu_i - \frac{\epsilon}{2}\right) \\
&\quad + \mathbb{P}\left(\exists \mathcal{A} \subset \mathrm{Top}^c_{m'}(\epsilon), \text{s.t. } |\mathcal{A}| = s - k \text{ and } \forall i \in \mathcal{A}, \hat{\mu}_i \ge \mu_i + \frac{\epsilon}{2}\right) \\
&\le \binom{m'}{m'-k} \exp\left(-(m'-k)\frac{\epsilon^2 T}{2n}\right) + \binom{n - \left|\mathrm{Top}_{m'}(\epsilon)\right|}{s-k} \exp\left(-(s-k)\frac{\epsilon^2 T}{2n}\right).
\end{aligned}$$

□

B.6 A minmax optimal budget allocation scheme for SH

We show that with a different budget allocation scheme $T_\ell = \left\lfloor \frac{T}{81n} \cdot \left(\frac{16}{9}\right)^{\ell-1} \cdot \ell \right\rfloor$, SH can achieve a sample complexity of $\mathcal{O}\left(\frac{n}{\epsilon^2}\log\frac{1}{\delta}\right)$ (without any logarithmic terms of n). We call this budget allocation scheme as Option 2 and the original one of Karnin et al. [21] Option 1. The price of achieving this exact minimax optimality is the potential loss of the near instance optimality (or at least the standard proof technique does not work). We summarize the comparison in Table 2. It is not clear to us whether there exists a different value of T_ℓ in SH that can achieve both the minimax and instance-dependent optimality. It is also not clear whether Option 2 can achieve a similar bound as Theorem 1. Note for the fixed budget setting, there is still a $\log\log(n)$ gap between the instance dependent upper and lower bounds.

Theorem 6. *For any $\epsilon \in (0,1)$, with Option 2, there exists an absolute constant c_1, such that the error probability of SH for identifying an ϵ-good arm satisfies,*

$$\mathbb{P}\big(\mu_{J_T} < \mu_1 - \epsilon\big) \le \exp\left(-\Theta\left(\frac{\epsilon^2 T}{n}\right)\right),$$

for $T > c_1 \cdot \frac{n}{\epsilon^2}$.

Proof. Still let $\epsilon_1 = \epsilon/4$. And define $\epsilon_{\ell+1} = \frac{3}{4}\cdot\epsilon_\ell$. For each stage ℓ, define the event G_ℓ as

$$G_\ell := \left\{\max_{i\in S_{\ell+1}} \mu_i \ge \max_{i\in S_\ell} \mu_i - \epsilon_\ell\right\}.$$

We have

$$\sum_{\ell=1}^{\log n} \epsilon_\ell < \sum_{\ell=1}^{\infty}\left(\frac{3}{4}\right)^{\ell-1}\cdot\epsilon_1 = \frac{\epsilon}{4}\sum_{\ell=1}^{\infty}\left(\frac{3}{4}\right)^{\ell-1} \le \frac{\epsilon}{4}\lim_{n\to\infty}\frac{1-(3/4)^n}{1-3/4} = \epsilon.$$

For stage ℓ, we assign budget $T_\ell \cdot n2^{-(\ell-1)} := \frac{T}{81n}\cdot\left(\frac{16}{9}\right)^{\ell-1}\cdot\ell\cdot n2^{-(\ell-1)} = \frac{T}{81}\cdot\left(\frac{8}{9}\right)^{\ell-1}\cdot\ell$. We can verify this allocation scheme does not exceed the budget as follows,

$$\sum_{\ell=1}^{\log n} T_\ell \cdot n2^{-(\ell-1)} \le \sum_{\ell=1}^{\infty}\frac{T}{81}\cdot\left(\frac{8}{9}\right)^{\ell-1}\cdot\ell \le \frac{T}{9}\lim_{\ell\to\infty}\frac{1-(8/9)^\ell}{1-8/9} = T.$$

Let a_ℓ be the best arm in S_ℓ,

$$\begin{aligned}\mathbb{P}\big(G_\ell^c\big) &= \mathbb{P}\big(G_\ell^c, \hat{\mu}_{a_\ell} < \mu_{a_\ell} - \epsilon_\ell/2\big) + \mathbb{P}\big(G_\ell^c, \hat{\mu}_{a_\ell} \geq \mu_{a_\ell} - \epsilon_\ell/2\big) \\ &\leq \mathbb{P}\big(\hat{\mu}_{a_\ell} < \mu_{a_\ell} - \epsilon_\ell/2\big) + \mathbb{P}\big(G_\ell^c \mid \hat{\mu}_{a_\ell} \geq \mu_{a_\ell} - \epsilon_\ell/2\big).\end{aligned}$$

For the first term

$$\begin{aligned}\mathbb{P}\big(\hat{\mu}_{a_\ell} < \mu_{a_\ell} - \epsilon_\ell/2\big) &\leq \exp\left(-\frac{\epsilon_\ell^2}{2}\frac{T_\ell}{|S_\ell|}\right) \\ &\leq \exp\left(-\frac{\epsilon_1^2\big(\frac{9}{16}\big)^{\ell-1}}{2}\frac{\frac{T}{81n}\cdot\big(\frac{16}{9}\big)^{\ell-1}\cdot\ell\cdot n2^{-(\ell-1)}}{n2^{-(\ell-1)}}\right) \\ &\leq \exp\left(-\frac{\epsilon^2 T}{32\cdot 81n}\ell\right).\end{aligned}$$

For the second term,

$$\begin{aligned}\mathbb{P}\big(G_\ell^c \mid \hat{\mu}_{a_\ell} \geq \mu_{a_\ell} - \epsilon_\ell/2\big) &\leq \mathbb{P}\big(|\{i \in S_\ell \mid \hat{\mu}_i > \mu_i + \epsilon_\ell/2\}| \geq |S_\ell|/2\big) \\ &\leq \frac{\mathbb{E}\big[|\{i \in S_\ell \mid \hat{\mu}_i > \mu_i + \epsilon_\ell/2\}|\big]}{|S_\ell|/2} \\ &\leq \frac{|S_\ell|\exp\big(-\frac{\epsilon^2 T}{32\cdot 81n}\ell\big)}{|S_\ell|/2} \\ &= 2\exp\left(-\frac{\epsilon^2 T}{32\cdot 81n}\ell\right).\end{aligned}$$

By contraposition,

$$\begin{aligned}\mathbb{P}\big(\mu_{J_T} < \mu_1 - \epsilon\big) &\leq \sum_{\ell=1}^{\log n}\mathbb{P}\big(G_\ell^c\big) \\ &\leq \sum_{\ell=1}^{\infty} 3\exp\left(-\frac{\epsilon^2 T}{32\cdot 81n}\ell\right).\end{aligned}$$

Let $X \coloneqq \frac{\epsilon^2 T}{32\cdot 81n} > 0$,

$$\begin{aligned}\mathbb{P}\big(\mu_{J_T} < \mu_1 - \epsilon\big) &\leq 3\sum_{\ell=1}^{\infty}\exp(-\ell X) \\ &\leq 3\exp(-X)\frac{1-\exp(-\ell X)}{1-\exp(-X)} \\ &\leq 3\exp(-X)\frac{1}{1-\exp(-X)}.\end{aligned}$$

As $\frac{1}{1-\exp(-X)}$ is a monotonically decreasing function, there exists an absolute constant c_1, such that

$$\mathbb{P}\big(\mu_{J_T} < \mu_1 - \epsilon\big) \leq \exp\left(-\Theta\left(\frac{\epsilon^2 T}{n}\right)\right),$$

holds for $T > c_1 \cdot \frac{n}{\epsilon^2}$. □

B.7 Lower bound for uniform sampling

Theorem 7. *Consider a family $\mathcal{E}(n, m, \epsilon)$ of all two-level univariate Gaussian bandit instances with n arms, of which m arms have mean $\epsilon > 0$ and all other arms have mean 0. We consider the problem in which samples are drawn from some instance $\theta \in \mathcal{E}(n, m, \epsilon)$ in an off-line, uniform fashion with B samples from each arm (i.e. with a total sampling budget of T samples, $B = T/n$ ignoring integer effects) and a subset of arms $\hat{S} \subseteq [n]$ of size s is then chosen based only on the observed samples*

	Worst-case	Instance-dependent
Option 1: $T_\ell = \left\lfloor \frac{T}{\lvert S_\ell \rvert \lceil \log_2 n \rceil} \right\rfloor$	✗ $\log_2 n \cdot \exp\left(-\frac{\epsilon^2 T}{n \log_2 n}\right)$	✓ $\log_2 n \cdot \exp\left(-\frac{T}{\log_2 n H_2}\right)$
Option 2: $T_\ell = \left\lfloor \frac{T}{81n} \cdot \left(\frac{16}{9}\right)^{\ell-1} \cdot \ell \right\rfloor$	✓ $\exp\left(-\frac{\epsilon^2 T}{n}\right)$	✗ $\log_2 n \cdot \exp\left(-\frac{T}{n \log_2 n H_2}\right)$

Table 2: The ϵ-error probability of SH (worst-case) and $(\epsilon = 0)$-error probability (instance-dependent) for Option 1 and 2 where $H_2 := \max_i \frac{i}{\Delta_i^2}$ is the problem hardness parameter. Here, we omit constant factors. Check marks for the worst-case means that it is optimal and those for the instance-dependent means that it is optimal up to $\log\log(n)$ factors in the sample complexity Mannor and Tsitsiklis [26], Carpentier and Locatelli [6].

(i.e. without knowledge of θ) in a symmetric fashion. Here 'symmetric' is taken to mean that for any subset $S \subseteq [n]$, $\mathbb{P}_\theta(\hat{S} = S) = \mathbb{P}_{\sigma(\theta)}(\hat{S} = \sigma(S))$ where $\sigma \in \Sigma_n$ is an element of the permutation group on sets of size n, and $\sigma(S)$ is taken to mean the element-wise application of σ to the elements of S.[2] *That is to say that the choice of $\hat{S}$ is independent of the specific indices of the arms chosen.*

Suppose $m \le s \wedge n/3$ and $s < \frac{6}{11}n$. Let $\tilde{M}_\theta$ be the number of false negatives (i.e. $\| \ge$) Then, there exists $\theta \in \mathcal{E}(n, m, \epsilon)$ s.t.

$$\mathbb{P}_\theta(\tilde{M}_\theta \ge \frac{1}{4}m) \ge \min\left(\frac{1}{2},\ 2\left(\frac{6}{5}\cdot\frac{n-s}{s}\right)^{\frac{3}{16}m} \exp\left(-\left(1 + 16\cdot\frac{\ln(en/m)}{\ln\left(\frac{6}{5}\cdot\frac{n-s}{s}\right)}\right) mB\epsilon^2\right)\right)$$

where $\mathbb{P}_\theta$ is the probability measure induced by sampling from instance θ.

Proof. Let $\theta = (\epsilon, \dots, \epsilon, 0, \dots, 0) \in \mathbb{R}^n$ with $\epsilon > 0$ be a vector of mean rewards for each arm where the first m coordinates are nonzero. Let $\hat{S} \in [n]$ be the output of the algorithm (recall $|\hat{S}| = s$). Let $\tilde{M}_\theta$ be the number of false negatives when taking $\hat{S}$ as the prediction for the true support $[s]$ of θ. Let $\tilde{Q}_a \subset 2^{[n]}$ be the collection of all subsets of $[n]$ of size s such that $\tilde{M}_\theta = a$. For convenience, let $\gamma = \frac{1}{8}m$. Let $k = \frac{3}{4}s$.

$$\begin{aligned}\xi := \mathbb{P}_\theta(\tilde{M}_\theta \ge m - k) &= \sum_{a=m-k}^{k} \mathbb{P}_\theta(\tilde{M}_\theta = a) \\ &\ge \sum_{a=3\gamma}^{5\gamma} \mathbb{P}_\theta(\tilde{M}_\theta = a).\end{aligned}$$

Let $\mathsf{first}(\tilde{Q}_a)$ be the first member of $\tilde{Q}_a$ in lexicographic order. For example, $\mathsf{first}(tQ_a) = [a+1 : a+s]$ where $[A : B] := \{A, A+1, \dots, B\}$. Then, by symmetry, one can see that $\mathbb{P}_\theta(\tilde{M}_\theta = a) = |\tilde{Q}_a| \mathbb{P}_\theta(\hat{S} = \mathsf{first}(\tilde{Q}_a))$.

We consider the event on which the 'empirical KL-divergence' concentrates, which can be shown to be true with probability at least $1 - \delta$ by Jun and Zhang [18, Lemma 5]:

$$\mathsf{conc}(\theta', \theta) := \left\{ \sum_{i=1}^{n}\sum_{t=1}^{B} \ln\left(\frac{p_{\theta'}(X_{i,t}|A_t = i)}{p_\theta(X_{i,t}|A_t = i)}\right) - (1+\rho)B\sum_{i=1}^{n}\mathsf{KL}(\theta'_i, \theta_i) \le \frac{1}{\rho}\ln(1/\delta)\right\}.$$

We now employ a change of measure argument to $\mathbb{P}_\theta(\hat{S} = \mathsf{first}(\tilde{Q}_a))$ and switch θ with θ' that would result in $\tilde{M}_{\theta'} = a - 3\gamma =: b$. For $a \in [3\gamma : 5\gamma]$, this can be achieved with the choice of $\theta' = (0, \dots, 0, \epsilon, \dots, \epsilon, 0, \dots, 0)$ that is supported on $[3\gamma + 1 : m + 3\gamma]$. Let Θ be the set of

[2] Note that the assumption of a symmetric selection $\hat{S}$ is equivalent to considering a lower bound for general selections $\hat{S}$ in the face of a uniform mixture over all permutations of θ (see [28] for a more in-depth discussion)

permutations of θ. Then,

$$
\begin{aligned}
&\mathbb{P}_\theta(\tilde{M}_\theta = a) \\
&= |\tilde{Q}_a|\,\mathbb{P}_\theta(\hat{S} = \mathsf{first}(\tilde{Q}_a)) \\
&\geq |\tilde{Q}_a|\,\mathbb{P}_{\theta'}(\hat{S} = \mathsf{first}(\tilde{Q}_a), \mathsf{conc}(\theta', \theta))\exp(-(1+\rho)2mB\cdot(\epsilon^2/2) - \rho^{-1}\ln(1/\delta)) \\
&\qquad\qquad \text{(change of measure; } \rho > 0) \\
&\geq |\tilde{Q}_a|\,\mathbb{P}_{\theta'}(\hat{S} = \mathsf{first}(\tilde{Q}_a), \cap_{\sigma\in\Sigma}\mathsf{conc}(\theta', \sigma(\theta')))\exp(-(1+\rho)mB\epsilon^2 - \rho^{-1}\ln(1/\delta)) \\
&\qquad\qquad (\Sigma\text{: symmetric group of } [n]) \\
&\geq |\tilde{Q}_a|\,\mathbb{P}_\theta(\hat{S} = \mathsf{first}(\tilde{Q}_b), \cap_{\sigma\in\Sigma}\mathsf{conc}(\theta, \sigma(\theta)))\exp(-(1+\rho)mB\epsilon^2 - \rho^{-1}\ln(1/\delta)) \quad \text{(symmetry)} \\
&= \frac{|\tilde{Q}_a|}{|\tilde{Q}_b|}\,\mathbb{P}_\theta(\tilde{M}_\theta = b, \cap_{\sigma\in\Sigma}\mathsf{conc}(\theta, \sigma(\theta)))\exp(-(1+\rho)mB\epsilon^2 - \rho^{-1}\ln(1/\delta)) \quad \text{(symmetry)} \\
&\geq \frac{|\tilde{Q}_a|}{|\tilde{Q}_b|}\left(\mathbb{P}_\theta(\tilde{M}_\theta = b) - |\Theta|\delta\right)\exp\left(-(1+\rho)mB\epsilon^2 - \rho^{-1}\ln(1/\delta)\right)\,. \\
&\qquad\qquad (\mathbb{P}(A,B) \geq \mathbb{P}(A) - \mathbb{P}(B^c)\text{; union bound over } \{\sigma(\theta) : \sigma \in \Sigma\})
\end{aligned}
$$

Thus,

$$
\begin{aligned}
&\sum_{a=3\gamma}^{5\gamma} \mathbb{P}_\theta(\tilde{M}_\theta = a) \\
&\geq \underbrace{\min_{a\in[3\gamma:5\gamma]} \frac{|\tilde{Q}_a|}{|\tilde{Q}_{a-3\gamma}|}}_{=:\,Y}\left(\underbrace{\mathbb{P}_\theta(\tilde{M}_\theta \in [0:2\gamma])}_{\geq 1-\xi} - (2\gamma+1)|\Theta|\delta\right)\exp\left(-(1+\rho)mB\epsilon^2 - \rho^{-1}\ln(1/\delta)\right)\,.
\end{aligned}
$$

Let us choose $\delta = \frac{1}{2}\cdot\frac{1-\xi}{(2\gamma+1)|\Theta|}$. Since the LHS above is at most ξ, we have

$$
\xi \geq Y\frac{1-\xi}{2}\exp\left(-(1+\rho)mB\epsilon^2 - \rho^{-1}\ln\left(\frac{2(2\gamma+1)|\Theta|}{1-\xi}\right)\right)
$$

One can consider two cases, namely $\xi \geq \frac{1}{2}$ and $\xi < \frac{1}{2}$, to arrive at

$$
\xi \geq \min\left(\frac{1}{2},\ \ \exp\left(-(1+\rho)mB\epsilon^2 - \rho^{-1}\ln\left(4(2\gamma+1)|\Theta|\right) + \ln(4Y)\right)\right)
$$

It remains to find an appropriate value of ρ. One simple choice is

$$
\rho^{-1} = \frac{1}{2}\frac{\ln(4Y)}{\ln\left(4(2\gamma+1)|\Theta|\right)}\,,
$$

which satisfies that $\rho^{-1} > 0$ as show later. Thus, we have

$$
\xi \geq \min\left(\frac{1}{2},\ \ 2\sqrt{Y}\exp\left(-\left(1 + 2\frac{\ln\left(4(2\gamma+1)|\Theta|\right)}{\ln(4Y)}\right)mB\epsilon^2\right)\right)
$$

It remains to figure out bounds for Y and $|\Theta|$. For Y, note that $|\tilde{Q}_a| = \binom{m}{m-a}\binom{n-m}{s-(m-a)}$. So, for $a \in [3\gamma : 5\gamma]$ and $b = a - 3\gamma$,

$$\begin{aligned}
\min_a \frac{|\tilde{Q}_a|}{|\tilde{Q}_b|} &= \frac{\binom{m}{m-a}\binom{n-m}{s-m+a}}{\binom{m}{m-b}\binom{n-m}{s-m+b}} \\
&= \frac{(m-b)(m-b-1)\cdots(m-a+1)}{a(a-1)\cdots(b+1)} \cdot \frac{(n-s-b)(n-s-b-1)\cdots(n-s-a+1)}{(s-m+a)(s-m+a-1)\cdots(s-m+b+1)} \\
&\overset{(a)}{\geq} \left(\frac{m-b}{a}\right)^{a-b} \cdot \left(\frac{n-s-b}{s-m+a}\right)^{a-b} \\
&\geq \left(\frac{6}{5}\right)^{3\gamma} \cdot \left(\frac{n-s-2\gamma}{s-3\gamma}\right)^{3\gamma} \geq \left(\frac{6}{5}\right)^{3\gamma} \cdot \left(\frac{n-s}{s}\right)^{3\gamma} \\
\implies Y = \min_{a\in[3\gamma:5\gamma]} \frac{|\tilde{Q}_a|}{|\tilde{Q}_{a-2\gamma}|} &\geq \left(\frac{6}{5}\right)^{3\gamma} \cdot \left(\frac{n-s}{s}\right)^{3\gamma}
\end{aligned}$$

where (a) is due to the fact that $\frac{m-b}{a} > 1$ implies $(m-b-i)/(a-i) \geq (m-b)/a$ for $i \in [0 : a-b-1]$ and, with a similar reasoning, $n \geq s/2 \implies \frac{n-s-a+3\gamma}{s-m+a} > 1 \implies (n-s-b-i)/(s-m+a-i) \geq (n-s-b)/(s-m+a)$. Then, using $|\Theta| = \binom{n}{m} \leq \left(\frac{en}{m}\right)^m$, we have

$$\begin{aligned}
\rho = 2\frac{\ln\left(4(2\gamma+1)|\Theta|\right)}{\ln(4Y)} &\leq 2\frac{\ln(m+4) + m\ln\left(\frac{en}{m}\right)}{\ln(4) + \frac{m}{4}\ln\left(\frac{6}{5}\cdot\frac{n-s}{s}\right)} \\
&\overset{(a)}{\leq} 4\frac{m\ln\left(\frac{en}{m}\right)}{\ln(4) + \frac{m}{4}\ln\left(\frac{6}{5}\cdot\frac{n-s}{s}\right)} \\
&\leq 16\frac{\ln(en/m)}{\ln\left(\frac{6}{5}\cdot\frac{n-s}{s}\right)}
\end{aligned}$$

where (a) is by $m \leq n/3 \implies \ln(m+4) \leq m\ln(en/m)$.

Altogether,

$$\mathbb{P}_\theta(\tilde{M} \geq \frac{1}{4}m) \geq \min\left(\frac{1}{2},\ 2\left(\frac{6}{5}\cdot\frac{n-s}{s}\right)^{\frac{3}{16}m} \exp\left(-\left(1 + 16\cdot\frac{\ln(en/m)}{\ln\left(\frac{6}{5}\cdot\frac{n-s}{s}\right)}\right)mB\epsilon^2\right)\right)$$

To verify that our choice of ρ is nonnegative, it suffices to show that $\frac{6}{5}\cdot\frac{n-s}{s} > 1$, which is true for $s < \frac{6}{11}n$.

□

C Simple Regret in the Data-Poor Regime

C.1 The number of opened brackets

Lemma 8. *The number of opened brackets till round t, denoted by L_t, satisfies*

$$0.63\cdot\log_2\big(1+\ln(2)t\big) < L_t \leq 1 + \log_2\big(1+\ln(2)t\big).$$

Proof. Recall that whenever $t \geq B2^B$ is true, we open a new bracket and increase B by 1. $B = 0$ before the game starts. First, note that for $t = 1$ we have $L_t = 1$. Then, let us focus on $t \geq 2$. It is not hard to verify that

$$L_t = \max\Big\{B : t \geq (B-1)2^{B-1}\Big\}.$$

Fix $t \geq 2$. Then, we can define

$$t' := (L_t - 1)2^{L_t - 1}, \tag{10}$$

the first time point the bracket L_t was opened. Clearly $t' \le t$. Solve (10) for L_t.

$$\begin{aligned} t' &= (L_t - 1)\exp\big(\ln(2)(L_t - 1)\big), \\ \ln(2)t' &= \ln(2)(L_t - 1)\exp\big(\ln(2)(L_t - 1)\big) \\ &=: X\exp(X). \end{aligned}$$

Solving it for X is exactly Lambert W function. By Lemma 17 of Orabona and Pal [27], we have

$$0.6321 \cdot \ln\big(1 + \ln(2)t'\big) \le X \le \ln\big(1 + \ln(2)t'\big).$$

This implies that

$$X = \ln(2)(L_t - 1) \le \ln\big(1 + \ln(2)t'\big) \le \ln\big(1 + \ln(2)t\big) \implies L_t \le 1 + \log_2\big(1 + \ln(2)t\big).$$

To obtain a lower bound, define

$$t'' := L_t 2^{L_t},$$

which is strictly larger than t. Using similar techniques as above, we can derive

$$L_t \ge 0.6321 \cdot \log_2\big(1 + \ln(2)t''\big) > 0.63 \cdot \log_2\big(1 + \ln(2)t\big).$$

Together, we have

$$0.63 \cdot \log_2\big(1 + \ln(2)t\big) < L_t \le 1 + \log_2\big(1 + \ln(2)t\big).$$

□

C.2 The number of pulls for the representative arms

Lemma 9. *At round t, the number of times the representative arm of bracket B has been pulled in the latest finished SH, denoted as $D_{B,t}$, satisfies,*

$$D_{B,t} \ge \text{const} \cdot \frac{t}{\ln t \log_2 n}.$$

Proof. We call the round interval $(L_t - 1)2^{L_t - 1} \le t < L_t 2^{L_t}$ as *phase* L_t, whose length is $L_t 2^{L_t} - (L_t - 1)2^{L_t - 1} = (L_t + 1)2^{L_t - 1}$. Correspondingly, sampling budget of $\frac{(L_t+1)2^{L_t-1}}{L_t}$ is assigned to each bracket. For each opened bracket $A_B, B \in [L_t - 1]$, the sampling budget for it is accumulated from phase B to phase L_t, while phase L_t is not finished yet. The sampling budget bracket A_B receives satisfies

$$\begin{aligned} T_{B,t} &= \sum_{i=B}^{L_t - 1} \frac{2^{i-1}(i+1)}{i} + \frac{t - 2^{L_t - 1}(L_t - 1)}{L_t} \\ &> \frac{2^{L_t - 2} L_t}{L_t - 1} + \frac{t - 2^{L_t - 1}(L_t - 1)}{L_t} \\ &> \frac{t}{4L_t} \\ &> \frac{t}{4 + 4\log\big(1 + \ln(2)t\big)} > \frac{t}{4\ln t}. \end{aligned} \qquad (t > 16)$$

Recall we run the SH with doubling trick on each bracket individually, we claim that the latest finished SH receives budget at least $T_{B,t}/4$. To see why, we initialize Algorithm 2 with budget staring from $\lceil n \log_2 n \rceil$. Then restart the SH algorithm with budget $2 \cdot \lceil n \log_2 n \rceil$, and so on. Before finishing the k-th doubling trick, the best arm from $(k-1)$-th doubling trick is output. Thus the portion of the lasted finished SH ranges in

$$\left[\frac{|\mathcal{T}_{k-1}|}{\sum_{i=1}^{k}|\mathcal{T}_i|}, \frac{|\mathcal{T}_{k-1}|}{\sum_{i=1}^{k-1}|\mathcal{T}_i|}\right),$$

which, by some simple calculations, is

$$\left[\frac{b_{r-1}}{4b_{r-1} - b_0}, \frac{b_{r-1}}{2b_{r-1} - b_0}\right). \qquad (11)$$

That is to say the lasted finished SH, which gives current output, has budget at least $T_{B,t}/4$. We notice that $T_{B,t}$ is essentially irrelevant to B. This means each bracket receives an order-wisely equal amount of budget to query the SH algorithm. In the SH algorithm, the output arm is pulled for fixed times when the number of arms and the sampling budget are fixed. It is pulled $\Theta\Big(\frac{T}{\log_2 n}\Big)$ times for budget T. Then we have all need to find out the lower bound for $D_{B,t}$.

$$D_{B,t} \geq \text{const} \cdot \frac{t}{\ln t \log_2 n}.$$

□

C.3 An example for the lower bound of the (ϵ, δ)-unverifiable sample complexity

While Katz-Samuels and Jamieson [22] shows a lower bound for the (ϵ, δ)-unverifiable sample complexity that can match the upper bound in special instances, it is loose and can even go to 0 in certain instances. Showing a tight lower bound for the (ϵ, δ)-unverifiable sample complexity is an interesting future work. Recall the definition of $\tau_{\epsilon,\delta}$, the (ϵ, δ)-unverifiable sample complexity (Definition 1).

Theorem 10 (Theorem 1 of Katz-Samuels and Jamieson [22])**.** *Fix $\epsilon > 0, \delta \in (0, 1/16)$, and a vector $\mu \in \mathbb{R}^n$. Consider n arms where rewards from the i-th arm are distributed according to $\mathcal{N}(\mu_i, 1)$. For every permutation $\pi \in \mathbb{S}^n$, let $(\mathcal{F}_t^\pi)_{t\in\mathbb{N}}$ be the filtration generated by the algorithm playing on instance $\pi(\mu)$. Then the (ϵ, δ)-unverifiable sample complexity satisfies,*

$$\mathbb{E}\big[\tau_{\epsilon,\delta}\big] \geq \frac{1}{64}\left(-\big(\mu_1 - \mu_{m(\epsilon)+1}\big)^{-2} + \frac{1}{m(\epsilon)} \sum_{i=m(\epsilon)+1}^{n} (\mu_1 - \mu_i)^{-2}\right),$$

where the expectation is with respect to $\pi \in \mathbb{S}^n$ and $\pi(\mu)$.

Consider the instance,

$$\mu_1 = 1, \mu_2 = \cdots = \mu_{n/2} = \frac{1}{2} - \frac{1}{n}, \mu_{n/2+1} = \cdots = \mu_n = \frac{1}{2} + \frac{1}{n}.$$

Take $\epsilon = 1/2$. Thus $m(\epsilon) = n/2$. The lower bound is vacuous,

$$\begin{aligned}
\mathbb{E}[\tau_\pi] \geq & \frac{1}{64}\left(-\big(\mu_1 - \mu_{m(\epsilon)+1}\big)^{-2} + \frac{1}{m(\epsilon)} \sum_{i=m(\epsilon)+1}^{n} (\mu_1 - \mu_i)^{-2}\right) \\
= & \frac{1}{64}\left(-\left(\frac{1}{2} + \frac{1}{n}\right)^{-2} + \frac{2}{n} \sum_{i=\frac{n}{2}+1}^{n} \left(\frac{1}{2} + \frac{1}{n}\right)^{-2}\right) \\
= & \frac{1}{64}\left(-\left(\frac{2n}{n+2}\right)^{2} + \frac{2}{n}\frac{n}{2}\left(\frac{2n}{n+2}\right)^{2}\right). \\
= & 0.
\end{aligned}$$

C.4 Proof of Theorem 3

Theorem 3. *For any $\epsilon \in (0, 1)$, the error probability of Bracketing SH satisfies*

$$\mathbb{P}\big(\mu_{J_t} < \mu_1 - \epsilon\big) \leq \exp\left(-\tilde{\Theta}\left(\min\left\{\max_{i\in[m(\epsilon/2)]} \frac{i\Delta_{i,m(\epsilon/2)+1}^2}{n}, \epsilon^2\right\} t'\right)\right),$$

where $t' = \frac{t}{4 \ln t}$. Accordingly, $\tilde{\mathcal{O}}\left(\max\left\{\min_{i\in[m(\epsilon/2)]} \frac{n}{i\Delta_{i,m(\epsilon/2)+1}^2} \log\big(\frac{1}{\delta}\big), \frac{1}{\epsilon^2}\log\big(\frac{1}{\delta}\big)\right\}\right)$ samples are sufficient for Bracketing SH to output an ϵ-good arm with probability $1 - \delta$. Here, $m(\epsilon) = \Big|\big\{i \in [n] : \mu_i \geq \mu_1 - \epsilon\big\}\Big|$.

Proof. We finish the complete proof based on the proof sketch in the main content. At round t, define the best bracket as bracket $r_i^*(t) = \left\lceil \log_2\left(c\Delta^2_{i,m(\epsilon/2)+1}t'\right)\right\rceil$, where $i \in [m(\epsilon/2)]$ is a free parameter and c is a logarithmic term for shorthand $c = \left(16\ln(4e)\log_2^2(\frac{i\Delta^2_{i,m(\epsilon/2)+1}t'}{n})\log_2\left(2\Delta^2_{i,m(\epsilon/2)+1}t'\right)\right)^{-1}$. Thus the size of bracket $r_i^*(t)$ satisfies,

$$c\Delta^2_{i,m(\epsilon/2)+1}t' \le \left|A_{r_i^*(t)}\right| \le 2^{\log_2\left(c\Delta^2_{i,m(\epsilon/2)+1}t'\right)+1} = 2c\Delta^2_{i,m(\epsilon/2)+1}t'.$$

Define the following events:

- E_1: the number of arms with average reward at least μ_i included in bracket $r_i^*(t)$ is at least j_t, where $j_t = \left\lfloor \frac{ci\Delta^2_{i,m(\epsilon/2)+1}t'}{2n} \right\rfloor$.
- E_2: $\mu_{a_{r_i^*(t)}} \ge \mu_1 - \epsilon/2$, where $a_{r_i^*(t)}$ is the arm representing bracket $r_i^*(t)$ at round t.

Then the error probability of Bracketing SH can be expressed as follows,

$$\begin{aligned}\mathbb{P}\left(\mu_{J_t} < \mu_1 - \epsilon\right) &= \mathbb{P}\left(\mu_{J_t} < \mu_1 - \epsilon, E_1^c\right) + \mathbb{P}\left(\mu_{J_t} < \mu_1 - \epsilon, E_1, E_2^c\right) + \mathbb{P}\left(\mu_{J_t} < \mu_1 - \epsilon, E_1, E_2\right) \\ &\le \mathbb{P}\left(E_1^c\right) + \mathbb{P}\left(E_1, E_2^c\right) + \mathbb{P}\left(\mu_{J_t} < \mu_1 - \epsilon, E_2\right). \end{aligned} \tag{12}$$

We bound the three terms respectively. For the first term, we use the same technique as in the proof sketch.

$$\begin{aligned}\mathbb{P}\left(E_1^c\right) &\le \exp\left(-\left|A_{r_i^*(t)}\right| \cdot \mathsf{KL}\left(\frac{j_t}{\left|A_{r_i^*(t)}\right|}, \frac{i}{n}\right)\right) \\ &\overset{(a_1)}{\le} \exp\left(-\frac{\left|A_{r_i^*(t)}\right| n}{2i}\left(\frac{j_t}{\left|A_{r_i^*(t)}\right|} - \frac{i}{n}\right)^2\right) \\ &\le \exp\left(-\frac{\Delta^2_{i,m(\epsilon/2)+1}t'n}{2i}\left(\frac{\frac{i\Delta^2_{i,m(\epsilon/2)+1}t'}{2n}}{\Delta^2_{i,m(\epsilon/2)+1}t'} - \frac{i}{n}\right)^2\right) \\ &\le \exp\left(-\frac{ci\Delta^2_{i,m(\epsilon/2)+1}t'}{8n}\right) \\ &\le \exp\left(-\tilde{\Theta}\left(\frac{i\Delta^2_{i,m(\epsilon/2)+1}t'}{n}\right)\right). \end{aligned} \tag{13}$$

The inequality (a_1) is due to $\mathsf{KL}(p,q) \ge \frac{(p-q)^2}{2\max\{p,q\}}$. For the second term of (12), we bound the event that SH returns a non-$(j_t, \Delta_{i,m(\epsilon/2)+1})$-good arm, where j_t is with respect to the best bracket. We use $\mu_{(j_t)}$ for the average reward of the j_t-th best arm in bracket $r_i^*(t)$. Thus $\mu_{(j_t)} - \Delta_{i,m(\epsilon/2)+1} \ge \mu_1 - \epsilon/2$.

By Corollary 1.1,

$$\begin{aligned}
\mathbb{P}\big(E_1, E_2^c\big) &\le \mathbb{P}\Big(\mu_{a_{r_i^*(t)}} < \mu_1 - \epsilon/2 \mid E_1\Big) \\
&\le \mathbb{P}\Big(\mu_{a_{r_i^*(t)}} < \mu_{(j_t)} - \Delta_{i,m(\epsilon/2)+1} \mid E_1\Big) \\
&\overset{(a_1)}{\le} \log_2\Big|A_{r_i^*(t)}\Big| \cdot \exp\left(-\text{const}\cdot j_t\left(\frac{\Delta^2_{i,m(\epsilon/2)+1} t'}{4\Big|A_{r_i^*(t)}\Big| \log_2^2(2j_t)\log_2\Big|A_{r_i^*(t)}\Big|} - \ln(4e)\right)\right) \\
&\le \log_2\Big|A_{r_i^*(t)}\Big| \cdot \exp\left(-\text{const}\cdot\frac{ci\Delta^2_{i,m(\epsilon/2)+1} t'}{n}\right) \\
&\le \exp\left(-\tilde{\Theta}\left(\frac{i\Delta^2_{i,m(\epsilon/2)+1} t'}{n}\right)\right).
\end{aligned} \tag{14}$$

For the inequality (a_1), we bound the term in parenthesis as a constant,

$$\begin{aligned}
&\frac{\Delta^2_{i,m(\epsilon/2)+1} t'}{4\Big|A_{r_i^*(t)}\Big| \log_2^2(2j_t)\log_2\Big|A_{r_i^*(t)}\Big|} - \ln(4e) \\
>&\frac{\Delta^2_{i,m(\epsilon/2)+1} t'}{8c\Delta^2_{i,m(\epsilon/2)+1} t' \log_2^2(2c\frac{i\Delta^2_{i,m(\epsilon/2)+1} t'}{2n})\log_2\Big(2c\Delta^2_{i,m(\epsilon/2)+1} t'\Big)} - \ln(4e) \\
>&\frac{1}{8c\log_2^2(\frac{i\Delta^2_{i,m(\epsilon/2)+1} t'}{n})\log_2\Big(2\Delta^2_{i,m(\epsilon/2)+1} t'\Big)} - \ln(4e) \\
>&2\ln(4e) - \ln(4e) \\
=&\ln(4e).
\end{aligned}$$

The third term of (12) means an arm whose average reward is less than $\mu_1 - \epsilon$ has a larger empirical reward than an arm whose average reward is larger than $\mu_1 - \epsilon/2$. Note the recently opened brackets may not have finished their first SH yet. In this case, the DSH always returns an empirical reward of negative infinity. Thus the arms returned from these brackets will never be selected by BSH. Denote $\dot{A}(t) := \left\{a \in \bigcup_{k=1,k\neq r_i^*(t)}^{L_t} A_k, \mu_a < \mu_1 - \epsilon\right\}$. We have

$$\Big|\dot{A}(t)\Big| \le \sum_{k=1}^{L_t} 2^k \le 2\cdot\Big(2^{1+\log_2(1+\ln(2)t)} - 1\Big) \le 8\cdot\big(1 + \ln(2)t\big).$$

By the number of arm pulls in Lemma 9,

$$\begin{aligned}
\mathbb{P}\big(\mu_{J_t} < \mu_1 - \epsilon, E_2\big) &\le \mathbb{P}\Big(\exists a_1 \in A_{r_i^*(t)}, \exists a_2 \in \dot{A}(t), \text{s.t.}, \mu_{a_1} \ge \mu_1 - \epsilon/2, \hat{\mu}_{a_1} < \hat{\mu}_{a_2}\Big) \\
&\le \Big|\dot{A}(t)\Big|\Big|A_{r_i^*(t)}\Big| \exp\left(-\left(\frac{\epsilon}{2}\right)^2 \cdot \text{const}\cdot\frac{t'}{\log_2 n}\right) \\
&\le \exp\left(-\text{const}\cdot\epsilon^2\frac{t'}{\log_2 n} + \ln\left(\Big|\dot{A}(t)\Big|\Big|A_{r_i^*(t)}\Big|\right)\right) \\
&\le \exp\left(-\text{const}\cdot\epsilon^2\frac{t'}{\log_2 n} + \ln\Big(8\cdot\big(1+\ln(2)t\big)2\Delta^2_{i,m(\epsilon/2)+1} t'\Big)\right) \\
&\le \exp\left(-\text{const}\cdot\epsilon^2\frac{t'}{\log_2 n} + \mathcal{O}(\ln t)\right) \\
&\le \exp\Big(-\tilde{\Theta}\big(\epsilon^2 t'\big)\Big).
\end{aligned} \tag{15}$$

Combine (13)(14)(15) and the fact $i \in [m(\epsilon/2)]$ is a free parameter, then we have

$$\mathbb{P}\Big(\mu_{J_t} < \mu_1 - \epsilon\Big) \le \exp\left(-\tilde{\Theta}\left(\min\left\{\max_{i\in[m(\epsilon/2)]} \frac{i\Delta^2_{i,m(\epsilon/2)+1}}{n}, \epsilon^2\right\} t'\right)\right).$$

□

C.5 (ϵ, δ)-unverifiable sample complexity of algorithms with ϵ-error probability bound

Definition 1 ((ϵ, δ)-unverifiable sample complexity [22])**.** *For an algorithm π and an instance ρ. Let $\tau_{\epsilon,\delta}$ be a stopping time such that*

$$\mathbb{P}\Big(\forall t \ge \tau_{\epsilon,\delta} : \mu_{J_t} > \mu_1 - \epsilon\Big) \ge 1 - \delta.$$

Then, $\tau_{\epsilon,\delta}$ is called (ϵ, δ)-unverifiable sample complexity of the algorithm with respect to ρ.

The (ϵ, δ)-unverifiable sample complexity requires all the outputs at and after $t = \tau_{\epsilon,\delta}$ are ϵ-good. This is slightly different from the sample complexity of BSH. However, we show they are order-wise equivalent. To see this, let us consider an anytime algorithm that achieves an exponentially decreasing error probability,

$$\mathbb{P}\Big(\mu_{J_t} < \mu_1 - \epsilon\Big) \le \exp(-c{\cdot}t).$$

Let $t_{\epsilon,\delta}$ be the first time step that satisfies $\mathbb{P}\Big(\mu_{J_{t_{\epsilon,\delta}}} < \mu_1 - \epsilon\Big) \le \delta$. Note this only guarantees the output at round $t_{\epsilon,\delta}$ is ϵ-good with probability $1 - \delta$, instead of all the outputs at and after $t_{\epsilon,\delta}$. We prove our claim by showing the probability that there is a non-ϵ-good output at or after $t_{\epsilon,\delta}$ has an exponentially decreasing rate with same parameter.

$$\begin{aligned}
\mathbb{P}\Big(\exists t \ge t_{\epsilon,\delta} : \mu_{J_t} < \mu_1 - \epsilon\Big) &\le \sum_{t=t_{\epsilon,\delta}}^{\infty} \mathbb{P}\Big(\mu_{J_t} < \mu_1 - \epsilon\Big) \\
&\le \sum_{t=t_{\epsilon,\delta}}^{\infty} \exp(-c{\cdot}t) \\
&= \lim_{t\to\infty} \exp\Big(-c{\cdot}t_{\epsilon,\delta}\Big) \frac{1 - \exp(-c{\cdot}t)}{1 - \exp(-c)} \\
&< \exp\Big(-c{\cdot}t_{\epsilon,\delta}\Big) \frac{1}{1 - \exp(-c)} \\
&= \exp\Big(-\Theta(c{\cdot}t_{\epsilon,\delta})\Big).
\end{aligned}$$

C.6 Proof of Corollary 3.1

Corollary 3.1. *Consider the equal gap instance (3). The Bracketing UCB achieves an expected (ϵ, δ)-unverifiable sample complexity as*

$$\mathbb{E}\big[\tau_{\epsilon,\delta}\big] \le \tilde{\mathcal{O}}\left(\frac{n}{\epsilon^2 m} \log\left(\frac{1}{\delta}\right)\right).$$

In the same instance, Bracketing SH achieves an (ϵ, δ)-unverifiable sample complexity as

$$\tau_{\epsilon,\delta} \le \tilde{\mathcal{O}}\left(\frac{n}{\epsilon^2 m} \log\left(\frac{1}{\delta}\right)\right)$$

with probability $1 - \delta$.

Proof. For BUCB, Theorem 7 of Katz-Samuels and Jamieson [22] gives the following instance dependent upper bound for the (ϵ, δ)-unverifiable sample complexity,

$$\mathbb{E}\big[\tau_{\epsilon,\delta}\big] \le \min_{j\in[m(\epsilon)]} \frac{1}{j}\left(\sum_{i=1}^{m(\epsilon)} \big(\Delta_{i\vee j, m(\epsilon)+1}\big)^{-2} \ln\left(\frac{n}{j\delta}\right) + \sum_{i=m(\epsilon)+1}^{n} \Delta_{j,i}^{-2} \ln\left(\frac{1}{\delta}\right)\right).$$

Plug the instance

$$\mu_1 = (3/2)\epsilon, \mu_i = \epsilon, \forall i \in \{2, \dots, m\}, \mu_i = 0, \forall i \geq m+1 \text{ for some } m \geq 2, \epsilon > 0 \,.$$

We easily have

$$\mathbb{E}\big[\tau_{\epsilon,\delta}\big] \leq \tilde{\mathcal{O}}\left(\frac{n}{\epsilon^2 m}\log\left(\frac{1}{\delta}\right)\right).$$

For BSH,

$$\min_{i\in[m(\epsilon/2)]} \frac{n}{i\Delta^2_{i,m(\epsilon/2)+1}}\log\left(\frac{1}{\delta}\right) \leq \min\left\{\frac{n}{(\epsilon/2)^2}\log\left(\frac{1}{\delta}\right), \frac{n}{m\epsilon^2}\log\left(\frac{1}{\delta}\right)\right\} = \frac{n}{m\epsilon^2}\log\left(\frac{1}{\delta}\right).$$

□

C.7 Proof of Corollary 3.2

Corollary 3.2. *Consider the polynomial instance; i.e.,* $\Delta_i = \left(\frac{i}{n}\right)^\alpha$ *with* $\alpha > 0$*. For any* $\epsilon \in (0,1)$*, let* $\hat{\tau}_{\epsilon,\delta}$ *be the upper bound of the expected* (ϵ,δ)*-unverifiable sample complexity reported in Katz-Samuels and Jamieson [22, Theorem 7] for Bracketing UCB. Then, BUCB satisfies*

$$\mathbb{E}\big[\tau_{\epsilon,\delta}\big] \leq \hat{\tau}_{\epsilon,\delta} = \tilde{\Theta}\left(\epsilon^{-\frac{2\alpha-1}{\alpha}} n \log\left(\frac{1}{\delta}\right)\right).$$

On the other hand, BSH satisfies, with probability $1-\delta$*,*

$$\tau_{\epsilon,\delta} \leq \tilde{\mathcal{O}}\left(\epsilon^{-\frac{2\alpha+1}{\alpha}} \log\left(\frac{1}{\delta}\right)\right).$$

Proof. We first show the result for Bracketing UCB. Theorem 7 of Katz-Samuels and Jamieson [22] gives the following instance dependent upper bound for the (ϵ,δ)-unverifiable sample complexity,

$$\hat{\tau}_{\epsilon,\delta} = \min_{j\in[m(\epsilon)]} \frac{1}{j}\left(\sum_{i=1}^{m(\epsilon)} \left(\Delta_{i\vee j, m(\epsilon)+1}\right)^{-2} \ln\left(\frac{n}{j\delta}\right) + \sum_{i=m(\epsilon)+1}^{n} \Delta_{j,i}^{-2} \ln\left(\frac{1}{\delta}\right)\right). \tag{16}$$

Plug the instance $\Delta_i = \left(\frac{i}{n}\right)^\alpha$ into the above,

$$
\begin{aligned}
\hat{\tau}_{\epsilon,\delta} &= \min_{j\in[m(\epsilon)]} \frac{1}{j}\left(\sum_{i=1}^{m(\epsilon)} \left(\Delta_{i\vee j, m(\epsilon)+1}\right)^{-2} \ln\left(\frac{n}{j\delta}\right) + \sum_{i=m(\epsilon)+1}^{n} \Delta_{j,i}^{-2} \ln\left(\frac{1}{\delta}\right)\right) \\
&> \min_{j\in[m(\epsilon)]} \frac{1}{j}\left(\sum_{i=1}^{m(\epsilon)} \left(\Delta_{i\vee j, m(\epsilon)+1}\right)^{-2} \ln\left(\frac{n}{j\delta}\right)\right) \\
&> \min_{j\in[m(\epsilon)]} \frac{1}{j} \ln\left(\frac{1}{\delta}\right)\left(\sum_{i=1}^{m(\epsilon)} \left(\Delta_{i\vee j, m(\epsilon)+1}\right)^{-2}\right) \\
&> \min_{j\in[m(\epsilon)]} \frac{1}{j} \ln\left(\frac{1}{\delta}\right)\left(\left(\Delta_{m(\epsilon), m(\epsilon)+1}\right)^{-2}\right) \\
&> \min_{j\in[m(\epsilon)]} \frac{n^{2\alpha}}{j} \ln\left(\frac{1}{\delta}\right)\left(\frac{1}{\left((m(\epsilon)+1)^{\alpha} - (m(\epsilon))^{\alpha}\right)^2}\right) \\
&\overset{(a_1)}{>} \min_{j\in[m(\epsilon)]} \frac{n^{2\alpha}}{j} \ln\left(\frac{1}{\delta}\right)\left(\frac{1}{\left(\alpha\left(m(\epsilon)+1\right)^{\alpha-1}\right)^2}\right) \\
&> \min_{j\in[m(\epsilon)]} \frac{n^{2\alpha}}{j} \ln\left(\frac{1}{\delta}\right)\left(\frac{1}{\alpha^2\left(2m(\epsilon)\right)^{2\alpha-2}}\right) \\
&= \frac{2^{-2\alpha+2}\alpha^{-2}n^{2\alpha}}{\left(m(\epsilon)\right)^{2\alpha-1}} \ln\left(\frac{1}{\delta}\right) \\
&\overset{(a_2)}{=} \frac{2^{-2\alpha+2}\alpha^{-2}n}{\epsilon^{\frac{2\alpha-1}{\alpha}}} \ln\left(\frac{1}{\delta}\right).
\end{aligned}
$$

Note here we lower bound the upper bound in Theorem 7 of Katz-Samuels and Jamieson [22]. The inequality (a_1) results by

$$
(m+1)^{\alpha} - m^{\alpha} = \int_{m}^{m+1} \alpha x^{\alpha-1}\,\mathrm{d}x \le \alpha(m+1)^{\alpha-1}.
$$

The equation (a_2) is because of $m(\epsilon) = n\epsilon^{1/\alpha}$ by the definition of the instance.

We also upper bound (16). Note that for $i > j$, we have

$$
i^{\alpha} - j^{\alpha} = \int_{j}^{i} \alpha x^{\alpha-1}\,\mathrm{d}x \ge \alpha(i-j)\cdot j^{\alpha-1}\ .
$$

With this, we have

$$
\hat{\tau}_{\epsilon,\delta} \le n^{2\alpha}\left(\min_{j\le m(\epsilon)} \frac{1}{(m(\epsilon)+1-j)^2 j^{2\alpha-2}} + \frac{1}{j}\sum_{i=j+1}^{m(\epsilon)} \frac{1}{(m(\epsilon)+1-i)^2 i^{2\alpha-2}} + \frac{1}{j}\sum_{i=m(\epsilon)+1}^{n} \frac{1}{(i-j)^2 j^{2\alpha-2}}\right)
$$

We can bound the third sum by

$$
\frac{1}{j}\sum_{i=m(\epsilon)+1}^{n} \frac{1}{(i-j)^2 j^{2\alpha-2}} \lesssim \frac{1}{j^{2\alpha-1}}\cdot\frac{1}{m(\epsilon)+1-j}
$$

Take $j = m(\epsilon) - \sqrt{m(\epsilon)}$. This means that $j = \Theta(m(\epsilon))$ and $m(\epsilon) - j = \sqrt{m(\epsilon)}$. With this choice,

$$\begin{aligned}
\hat{\tau}_{\epsilon,\delta} &\lesssim n^{2\alpha}\left(\frac{1}{m(\epsilon)\cdot m(\epsilon)^{2\alpha-2}} + \frac{1}{m(\epsilon)}\cdot\left(\sum_{i=m(\epsilon)-\sqrt{m(\epsilon)}+1}^{m(\epsilon)} \frac{1}{(m(\epsilon)+1-i)^2}\cdot\frac{1}{m(\epsilon)^{2\alpha-2}}\right) + \frac{1}{m(\epsilon)^{2\alpha-1}}\cdot\frac{1}{1+\sqrt{m(\epsilon)}}\right) \\
&\leq n^{2\alpha}\left(\frac{1}{m(\epsilon)^{2\alpha-1}} + \frac{1}{m(\epsilon)^{2\alpha-1}}\cdot\frac{\pi^2}{6} + \frac{1}{m(\epsilon)^{2\alpha-1}}\cdot\frac{1}{1+\sqrt{m(\epsilon)}}\right) \\
&\lesssim n^{2\alpha}\cdot\frac{1}{m(\epsilon)^{2\alpha-1}} = n^{2\alpha}\frac{1}{(n\epsilon^{1/\alpha})^{2\alpha-1}} = \frac{n}{\epsilon^{\frac{2\alpha-1}{\alpha}}}.
\end{aligned}$$

For Bracketing SH, first notice that

$$\begin{aligned}
\Delta_{i,j+1}^2 > \Delta_{i,j}^2 =& \left(\left(\frac{j}{n}\right)^\alpha - \left(\frac{i}{n}\right)^\alpha\right)^2 \\
=& \left(\frac{j}{n}\right)^{2\alpha} + \left(\frac{i}{n}\right)^{2\alpha} - 2\left(\frac{ij}{n^2}\right)^\alpha.
\end{aligned}$$

The sample complexity satisfies,

$$\begin{aligned}
\min_{i\in[m(\epsilon/2)]} \frac{n}{i\Delta_{i,m(\epsilon/2)+1}^2}\log\left(\frac{1}{\delta}\right) &\leq \min_{i\in[m(\epsilon/2)]} \frac{n}{i\left(\left(\frac{m(\epsilon/2)}{n}\right)^{2\alpha} + \left(\frac{i}{n}\right)^{2\alpha} - 2\left(\frac{im(\epsilon/2)}{n^2}\right)^\alpha\right)}\log\left(\frac{1}{\delta}\right) \\
&= \min_{i\in[m(\epsilon/2)]} \frac{n\cdot n^{2\alpha}}{i\left(\left(m(\epsilon/2)\right)^{2\alpha} + i^{2\alpha} - 2\left(im(\epsilon/2)\right)^\alpha\right)}\log\left(\frac{1}{\delta}\right) \\
&\overset{(a_1)}{\leq} \frac{2\cdot n^{2\alpha+1}}{(m(\epsilon/2))^{2\alpha+1}(1+2^{-2\alpha}-2^{-\alpha+1})}\log\left(\frac{1}{\delta}\right) \\
&= \frac{2\cdot n^{2\alpha+1}}{(n(\epsilon/2)^{1/\alpha})^{2\alpha+1}(1+2^{-2\alpha}-2^{-\alpha+1})}\log\left(\frac{1}{\delta}\right) \\
&= 2^{\frac{2\alpha+1}{\alpha}+1}\left(1+2^{-2\alpha}-2^{-\alpha+1}\right)^{-1}\epsilon^{-\frac{2\alpha+1}{\alpha}}\log\left(\frac{1}{\delta}\right).
\end{aligned}$$

The inequality (a_1) is by taking $i = m(\epsilon/2)/2$.

□

D Discussion on the Practical Algorithm Implementation

The design of DSH (Algorithm 2) and BSH (Algorithm 3) involves two key ideas, bracketing and doubling trick. In practice, both two techniques could be implemented in a more efficient way. In addition, the base algorithm SH (Algorithm 1) has a commonly used implementation for reusing samples [4, 17]. We summarize these strategies as follows for the practitioner's consideration. However, they do not make any order-wise difference (not more than logarithmic factors) in terms of the theoretical guarantee.

- SH: The whole procedure of SH is divided into $\log_2 n$ stages. All the stages are independent from each other, since the samples of prior stages are abandoned when a new stage starts as described in Algorithm 1. In practical usage, one can keep all the samples since the first stage for the surviving arms.
- DSH: The common doubling trick does not require to keep the samples after finishing one invocation of the base algorithm. Not requiring to keep the samples is convenient to implement as we just need to repeatedly initialize a new instance of the base algorithm with

the doubled budget parameter. In fact, for multi-armed bandit problems, it is beneficial to keep all the samples of prior invocations. For the implementation, one can create a class for the bash algorithm SH, and we initialize a new instance of the class with initial empirical rewards equal to the empirical rewards saved in the prior instance.

- BSH: The bracketing technique does not promise to avoid overlap with the already-opened brackets. A more practical implementation of BSH is to share the samples of the same arm across different brackets. Such that the empirical reward is more accurate. However, reusing sampling across different brackets is meaningless, if we consider the infinitely-armed bandit models. Because, for the infinitely-armed bandit models, we will not draw exactly the same arm more than once. Thus it is barely possible to have overlap among the opened brackets.

Note our implementation reported in section 5 only uses the first reusing strategy for the base algorithm SH.

References

[1] Jean-Yves Audibert and Sébastien Bubeck. Best arm identification in multi-armed bandits. In *Proceedings of the Conference on Learning Theory (COLT)*, pages 41–66, 2010.

[2] Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning (ICML)*, pages 263–272. PMLR, 2017.

[3] Maryam Aziz, Jesse Anderton, Emilie Kaufmann, and Javed Aslam. Pure exploration in infinitely-armed bandit models with fixed-confidence. In *Algorithmic Learning Theory*, pages 3–24. PMLR, 2018.

[4] Tavor Baharav and David Tse. Ultra fast medoid identification via correlated sequential halving. *Advances in Neural Information Processing Systems (NeurIPS)*, 32, 2019.

[5] Romain Camilleri, Kevin Jamieson, and Julian Katz-Samuels. High-dimensional experimental design and kernel bandits. In *International Conference on Machine Learning (ICML)*, pages 1227–1237. PMLR, 2021.

[6] Alexandra Carpentier and Andrea Locatelli. Tight (lower) bounds for the fixed budget best arm identification bandit problem. *Journal of Machine Learning Research*, 49(June):590–604, 2016.

[7] Alexandra Carpentier and Michal Valko. Simple regret for infinitely many armed bandits. In *International Conference on Machine Learning (ICML)*, pages 1133–1141. PMLR, 2015.

[8] Arghya Roy Chaudhuri and Shivaram Kalyanakrishnan. PAC identification of a bandit arm relative to a reward quantile. In *Thirty-First AAAI Conference on Artificial Intelligence*, pages 1777–1783, 2017.

[9] Arghya Roy Chaudhuri and Shivaram Kalyanakrishnan. PAC identification of many good arms in stochastic multi-armed bandits. In *International Conference on Machine Learning (ICML)*, pages 991–1000. PMLR, 2019.

[10] Lijie Chen, Jian Li, and Mingda Qiao. Nearly instance optimal sample complexity bounds for top-k arm selection. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 101–110. PMLR, 2017.

[11] Lijie Chen, Jian Li, and Mingda Qiao. Towards instance optimal bounds for best arm identification. In *Conference on Learning Theory (COLT)*, pages 535–592. PMLR, 2017.

[12] Christoph Dann, Tor Lattimore, and Emma Brunskill. Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017.

[13] Eyal Even-Dar, Shie Mannor, Yishay Mansour, and Sridhar Mahadevan. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of machine learning research*, 7(6), 2006.

[14] Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *Proceedings of the Conference on Learning Theory (COLT)*, pages 998–1027, 2016.

[15] Avinatan Hassidim, Ron Kupfer, and Yaron Singer. An optimal elimination algorithm for learning a best arm. *Advances in Neural Information Processing Systems (NeurIPS)*, 33: 10788–10798, 2020.

[16] Kevin Jamieson, Matthew Malloy, Robert Nowak, and Sébastien Bubeck. lil'UCB: An optimal exploration algorithm for multi-armed bandits. In *Proceedings of the Conference on Learning Theory (COLT)*, pages 423–439, 2014.

[17] K.-S. Jun and R. Nowak. Anytime exploration for multi-armed bandits using confidence information. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 48, pages 974–982, 2016. ISBN 9781510829008.

[18] Kwang-Sung Jun and Chicheng Zhang. Crush Optimism with Pessimism: Structured Bandits Beyond Asymptotic Optimality. *ICML Workshop on Theoretical Foundations of Reinforcement Learning (arXiv:2006.08754)*, 2020.

[19] Shivaram Kalyanakrishnan and Peter Stone. Efficient selection of multiple bandit arms: Theory and practice. In *International Conference on Machine Learning (ICML)*, 2010.

[20] Shivaram Kalyanakrishnan, Ambuj Tewari, Peter Auer, and Peter Stone. PAC subset selection in stochastic multi-armed bandits. In *International Conference on Machine Learning (ICML)*, volume 12, pages 655–662, 2012.

[21] Zohar Karnin, Tomer Koren, and Oren Somekh. Almost optimal exploration in multi-armed bandits. In *International Conference on Machine Learning (ICML)*, pages 1238–1246. PMLR, 2013.

[22] Julian Katz-Samuels and Kevin Jamieson. The true sample complexity of identifying good arms. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 108, pages 1781–1791, 2020.

[23] Emilie Kaufmann and Shivaram Kalyanakrishnan. Information complexity in bandit subset selection. In *Conference on Learning Theory (COLT)*, pages 228–251. PMLR, 2013.

[24] Emilie Kaufmann, Pierre Ménard, Omar Darwiche Domingues, Anders Jonsson, Edouard Leurent, and Michal Valko. Adaptive reward-free exploration. In *Algorithmic Learning Theory*, pages 865–891. PMLR, 2021.

[25] Levente Kocsis and Csaba Szepesvári. Bandit based monte-carlo planning. In *European conference on machine learning*, pages 282–293. Springer, 2006.

[26] Shie Mannor and John N Tsitsiklis. The sample complexity of exploration in the multi-armed bandit problem. *Journal of Machine Learning Research*, 5(Jun):623–648, 2004.

[27] Francesco Orabona and David Pal. Coin betting and parameter-free online learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 29, pages 577–585, 2016.

[28] Max Simchowitz, Kevin Jamieson, and Benjamin Recht. The Simulator: Understanding Adaptive Sampling in the Moderate-Confidence Regime. *Proceedings of Machine Learning Research*, 65:1–41, 2017.

[29] Marta Soare, Alessandro Lazaric, and Rémi Munos. Best-arm identification in linear bandits. *Advances in Neural Information Processing Systems (NeurIPS)*, 27, 2014.

[30] Yinglun Zhu, Julian Katz-Samuels, and Robert Nowak. Near instance optimal model selection for pure exploration linear bandits. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 6735–6769. PMLR, 2022.