JMLR: Workshop and Conference Proceedings vol 35:1–17, 2014

lil' UCB : An Optimal Exploration Algorithm for Multi-Armed Bandits *

Kevin Jamieson KGJAMIESON@WISC.EDU
Matthew Malloy MMALLOY@WISC.EDU
Robert Nowak NOWAK@ECE.WISC.EDU
University of Wisconsin

Sébastien Bubeck SBUBECK@PRINCETON.EDU
Princeton University

Abstract

The paper proposes a novel upper confidence bound (UCB) procedure for identifying the arm with the largest mean in a multi-armed bandit game in the fixed confidence setting using a small number of total samples. The procedure cannot be improved in the sense that the number of samples required to identify the best arm is within a constant factor of a lower bound based on the law of the iterated logarithm (LIL). Inspired by the LIL, we construct our confidence bounds to explicitly account for the infinite time horizon of the algorithm. In addition, by using a novel stopping time for the algorithm we avoid a union bound over the arms that has been observed in other UCB-type algorithms. We prove that the algorithm is optimal up to constants and also show through simulations that it provides superior performance with respect to the state-of-the-art.

Keywords: Multi-armed bandit, upper confidence bound (UCB), iterated logarithm

1. Introduction

This paper introduces a new algorithm for the *best arm* problem in the stochastic multi-armed bandit (MAB) setting. Consider a MAB with n arms, each with unknown mean payoff $\mu_1, \dots, \mu_n$ in $[0, 1]$. A sample of the ith arm is an independent realization of a sub-Gaussian random variable with mean μ_i. In the *fixed confidence setting*, the goal of the best arm problem is to devise a sampling procedure with a single input δ that, regardless of the values of $\mu_1, \dots, \mu_n$, finds the arm with the largest mean with probability at least $1 - \delta$. More precisely, best arm procedures must satisfy $\sup_{\mu_1,\dots,\mu_n} \mathbb{P}(\widehat{i} \neq i^*) \leq \delta$, where i^* is the best arm, $\widehat{i}$ an estimate of the best arm, and the supremum is taken over all set of means such that there exists a unique best arm. In this sense, best arm procedures must automatically adjust sampling to ensure success when the mean of the best and second best arms are arbitrarily close. Contrast this with the *fixed budget setting* where the total number of samples remains a constant and the confidence in which the best arm is identified within the given budget varies with the setting of the means. While the fixed budget and fixed confidence settings are related (see Gabillon et al. (2012) for a discussion) this paper focuses on the fixed confidence setting only.

The best arm problem has a long history dating back to the '50s with the work of Paulson (1964); Bechhofer (1958). In the fixed confidence setting, the last decade has seen a flurry of

* Part of the research described here was carried out at the Simons Institute for the Theory of Computing. We are grateful to the Simons Institute for providing a wonderful research environment.

activity providing new upper and lower bounds. In 2002, the *successive elimination* procedure of Even-Dar et al. (2002) was shown to find the best arm with order $\sum_{i \neq i^*} \Delta_i^{-2} \log(n\Delta_i^{-2})$ samples, where $\Delta_i = \mu_{i^*} - \mu_i$, coming within a logarithmic factor of the lower bound of $\sum_{i \neq i^*} \Delta_i^{-2}$, shown in 2004 in Mannor and Tsitsiklis (2004). A similar bound was also obtained using a procedure known as *LUCB1* that was originally designed for finding the m-best arms (Kalyanakrishnan et al., 2012). Recently, Jamieson et al. (2013) proposed a procedure called *PRISM* which succeeds with $\sum_i \Delta_i^{-2} \log\log\left(\sum_j \Delta_j^{-2}\right)$ or $\sum_i \Delta_i^{-2} \log\left(\Delta_i^{-2}\right)$ samples depending on the parameterization of the algorithm, improving the result of Even-Dar et al. (2002) by at least a factor of $\log(n)$. The best sample complexity result for the fixed confidence setting comes from a procedure similar to PRISM, called *exponential-gap elimination* (Karnin et al., 2013), which guarantees best arm identification with high probability using order $\sum_i \Delta_i^{-2} \log\log \Delta_i^{-2}$ samples, coming within a doubly logarithmic factor of the lower bound of Mannor and Tsitsiklis (2004). While the authors of Karnin et al. (2013) conjecture that the $\log\log$ term cannot be avoided, it remained unclear as to whether the upper bound of Karnin et al. (2013) or the lower bound of Mannor and Tsitsiklis (2004) was loose.

The classic work of Farrell (1964) answers this question. It shows that the doubly logarithmic factor is necessary, implying that order $\sum_i \Delta_i^{-2} \log\log \Delta_i^{-2}$ samples are necessary and sufficient in the sense that no procedure can satisfy $\sup_{\Delta_1,\ldots,\Delta_n} \mathbb{P}(\widehat{i} \neq i^*) \leq \delta$ and use fewer than $\sum_i \Delta_i^{-2} \log\log \Delta_i^{-2}$ samples in expectation for all $\Delta_1, \ldots, \Delta_n$. The doubly logarithmic factor is a consequence of the law of the iterated logarithm (LIL) (Darling and Robbins, 1985). The LIL states that if X_ℓ are i.i.d. sub-Gaussian random variables with $\mathbb{E}[X_\ell] = 0$, $\mathbb{E}[X_\ell^2] = \sigma^2$ and we define $S_t = \sum_{\ell=1}^t X_\ell$ then

$$\limsup_{t\to\infty} \frac{S_t}{\sqrt{2\sigma^2 t \log\log(t)}} = 1 \quad \text{and} \quad \liminf_{t\to\infty} \frac{S_t}{\sqrt{2\sigma^2 t \log\log(t)}} = -1$$

almost surely. Here is the basic intuition behind the lower bound. Consider the two-arm problem and let Δ be the difference between the means. In this case, it is reasonable to sample both arms equally and consider the sum of differences of the samples, which is a random walk with drift Δ. The deterministic drift crosses the LIL bound above when $t\,\Delta = \sqrt{2t \log\log t}$. Solving this equation for t yields $t \approx 2\Delta^{-2} \log\log \Delta^{-2}$. This intuition will be formalized in the next section.

The LIL also motivates a novel approach to the best arm problem. Specifically, the LIL suggests a natural scaling for confidence bounds on empirical means, and we follow this intuition to develop a new algorithm for the best-arm problem. The algorithm is an Upper Confidence Bound (UCB) procedure (Auer et al., 2002) based on a finite sample version of the LIL. The new algorithm, called lil'UCB, is described in Figure 1. By explicitly accounting for the $\log\log$ factor in the confidence bound and using a novel stopping criterion, our analysis of lil'UCB avoids taking naive union bounds over time, as encountered in some UCB algorithms (Kalyanakrishnan et al., 2012; Audibert et al., 2010), as well as the wasteful "doubling trick" often employed in algorithms that proceed in epochs, such as the PRISM and exponential-gap elimination procedures (Even-Dar et al., 2002; Karnin et al., 2013; Jamieson et al., 2013). Also, in some analyses of best arm algorithms the upper confidence bounds of each arm are designed to hold with high probability for all arms uniformly, incurring a $\log(n)$ term in the confidence bound as a result of the necessary union bound over the n arms (Even-Dar et al., 2002; Kalyanakrishnan et al., 2012; Audibert et al., 2010). However, our stopping time allows for a tighter analysis so that arms with larger gaps are allowed larger confidence bounds than those arms with smaller gaps where higher confidence is required. Like exponential-gap elimination, lil'UCB is order optimal in terms of sample complexity.

It is easy to show that without the stopping condition (and with the right δ) our algorithm achieves a cumulative regret of the same order as standard UCB. Thus for the expert it may be surprising that such an algorithm can achieve optimal sample complexity for the best arm identification problem given the lower bound of Bubeck et al. (2009). As it was empirically observed in the latter paper there seems to be a transient regime, before this lower bound applies, where the performance in terms of best arm identification is excellent. In some sense the results in the present paper can be viewed as a formal proof of this transient regime: if stopped at the right time performance of UCB for best arm identification is near-optimal (or even optimal for lil'UCB).

One of the main motivations for this work was to develop an algorithm that exhibits great practical performance in addition to optimal sample complexity. While the sample complexity of exponential-gap elimination is optimal up to constants, and PRISM up to small $\log\log$ factors, the empirical performance of these methods is rather disappointing, even when compared to non-sequential sampling. Both PRISM and exponential-gap elimination employ *median elimination* (Even-Dar et al., 2002) as a subroutine. Median elimination is used to find an arm that is within $\varepsilon > 0$ of the largest, and has sample complexity within a constant factor of optimal for this sub-problem. However, the constant factors tend to be quite large, and repeated applications of median elimination within PRISM and exponential-gap elimination are extremely wasteful. On the contrary, lil'UCB does not invoke wasteful subroutines. As we will show, in addition to having the best theoretical sample complexities bounds known to date, lil'UCB also exhibits superior performance in practice with respect to state-of-the-art algorithms.

2. Lower Bound

Before introducing the lil'UCB algorithm, we show that the $\log\log$ factor in the sample complexity is necessary for best-arm identification. It suffices to consider a two armed bandit problem with a gap Δ. If a lower bound on the gap is unknown, then the $\log\log$ factor is necessary, as shown by the following result.

Theorem 1 *Consider the best arm problem in the fixed confidence setting with $n = 2$, difference between the two means Δ, and expected number of samples $\mathbb{E}_\Delta[T]$. Any procedure with $\sup_{\Delta\neq 0}\ \mathbb{P}(\widehat{i} \neq i^*) \leq \delta$, $\delta \in (0, 1/2)$, then has*

$$\limsup_{\Delta\to 0} \frac{\mathbb{E}_\Delta[T]}{\Delta^{-2}\log\log\Delta^{-2}} \geq 2 - 4\delta.$$

Proof The proof follows readily from Theorem 1 of Farrell (1964) by considering a reduction of the best arm problem with $n = 2$ in which the value of one arm is known. In this case, the only strategy available is to sample the other arm some number of times to determine if it is less than or greater than the known value. We have reduced the problem to the setting of (Farrell, 1964, Theorem 1), and stated it in Appendix A. ■

Theorem 1 implies that in the fixed confidence setting, no best arm procedure can have $\sup \mathbb{P}(\widehat{i} \neq i^*) \leq \delta$ *and* use fewer than $(2 - 4\delta)\sum_i \Delta_i^{-2}\log\log\Delta_i^{-2}$ samples in expectation for all Δ_i.

In brief, the result of Farrell follows by showing a generalized sequential probability ratio test, which compares the running empirical mean of X after t samples against a series of thresholds,

is an optimal test. In the limit as t increases, if the thresholds are not at least $\sqrt{(2/t)\log\log(t)}$ then the LIL implies the procedure will fail with probability approaching 1/2 for small values of Δ. Setting the thresholds to be just greater than $\sqrt{(2/t)\log\log(t)}$, in the limit, one can show the expected number of samples must scale as $\Delta^{-2}\log\log\Delta^{-2}$. As the proof in Farrell (1964) is quite involved, we provide a short argument for a slightly simpler result than above in Appendix A.

3. Procedure

This section introduces lil'UCB. The procedure operates by sampling the arm with the largest upper confidence bound; the confidence bounds are defined to account for the implications of the LIL. The procedure terminates when one of the arms has been sampled more than a constant times the number of samples collected from all other arms combined. Fig. 1 details the algorithm and Theorem 2 quantifies performance. In what follows, let $X_{i,s}$, $s = 1, 2, \ldots$ denote independent samples from arm i and let $T_i(t)$ denote the number of times arm i has been sampled up to time t. Define $\widehat{\mu}_{i,T_i(t)} := \frac{1}{T_i(t)}\sum_{s=1}^{T_i(t)} X_{i,s}$ to be the empirical mean of the $T_i(t)$ samples from arm i up to time t. The algorithm of Fig. 1 assumes that the centered realizations of the ith arm are sub-Gaussian[1] with known scale parameter σ.

lil' UCB

input: confidence $\delta > 0$, algorithm parameters ε, λ, $\beta > 0$

initialize: sample each of the n arms once, set $T_i(t) = 1$ for all i and set $t = n$

while $T_i(t) < 1 + \lambda\sum_{j\neq i} T_j(t)$ for all i

sample arm

$$I_t = \operatorname*{argmax}_{i\in\{1,\ldots,n\}} \left\{ \widehat{\mu}_{i,T_i(t)} + (1+\beta)(1+\sqrt{\varepsilon})\sqrt{\frac{2\sigma^2(1+\varepsilon)\log\left(\frac{\log((1+\varepsilon)T_i(t))}{\delta}\right)}{T_i(t)}} \right\}.$$

set $T_i(t+1) = T_i(t) + 1$ if $I_t = i$, otherwise set $T_i(t+1) = T_i(t)$.

else stop and output $\arg\max_{i\in\{1,\ldots,n\}} T_i(t)$

Figure 1: The lil' UCB algorithm.

Define

$$\mathbf{H}_1 = \sum_{i\neq i^*} \frac{1}{\Delta_i^2} \quad \text{and} \quad \mathbf{H}_3 = \sum_{i\neq i^*} \frac{\log\log_+(1/\Delta_i^2)}{\Delta_i^2}$$

where $\log\log_+(x) = \log\log(x)$ if $x \geq e$, and 0 otherwise. Our main result is the following.

Theorem 2 *For $\varepsilon \in (0,1)$, let $c_\varepsilon = \frac{2+\varepsilon}{\varepsilon}(1/\log(1+\varepsilon))^{1+\varepsilon}$ and fix $\delta \in (0, \log(1+\varepsilon)/(ec_\varepsilon))$. Then for any $\beta \in (0,3]$, there exists a constant $\lambda > 0$ such that with probability at least $1 - 4\sqrt{c_\varepsilon\delta} - 4c_\varepsilon\delta$ lil' UCB stops after at most $c_1\mathbf{H}_1\log(1/\delta) + c_3\mathbf{H}_3$ samples and outputs the optimal arm, where $c_1, c_3 > 0$ are known constants that depend only on $\varepsilon, \beta, \sigma^2$.*

1. A zero-mean random variable X is said to be sub-Gaussian with scale parameter σ if for all $t \in \mathbb{R}$ we have $\mathbb{E}[\exp\{tX\}] \leq \exp\{\sigma^2t^2/2\}$. If $a \leq X \leq b$ almost surely than it suffices to take $\sigma^2 = (b-a)^2/4$.

Note that the algorithm obtains the optimal query complexity of $\mathbf{H}_1 \log(1/\delta) + \mathbf{H}_3$ up to constant factors. We remark that the theorem holds with any value of λ satisfying (7). Inspection of (7) shows that as $\delta \to 0$ we can let λ tend to $\left(\frac{2+\beta}{\beta}\right)^2$. We point out that the sample complexity bound in the theorem can be optimized by choosing ε and β. For a setting of these parameters in a way that is more or less faithful to the theory, we recommend taking $\varepsilon = 0.01$, $\beta = 1$, and $\lambda = \left(\frac{2+\beta}{\beta}\right)^2$. For improved performance in practice, we recommend applying footnote 2 and setting $\varepsilon = 0$, $\beta = 0.5$, $\lambda = 1 + 10/n$ and $\delta \in (0, 1)$, which do not meet the requirements of the theorem, but work very well in our experiments presented later. We prove the theorem via two lemmas, one for the total number of samples taken from the suboptimal arms and one for the correctness of the algorithm. In the lemmas we give precise constants.

4. Proof of Theorem 2

Before stating the two main lemmas that imply the result, we first present a finite form of the law of iterated logarithm. This finite LIL bound is necessary for our analysis and may also prove useful for other applications.

Lemma 3 *Let $X_1, X_2, \ldots$ be i.i.d. centered sub-Gaussian random variables with scale parameter σ. For any $\varepsilon \in (0, 1)$ and $\delta \in (0, \log(1+\varepsilon)/e)$[2] one has with probability at least $1 - \frac{2+\varepsilon}{\varepsilon}\left(\frac{\delta}{\log(1+\varepsilon)}\right)^{1+\varepsilon}$ for all $t \geq 1$,*

$$\sum_{s=1}^{t} X_s \leq (1+\sqrt{\varepsilon})\sqrt{2\sigma^2(1+\varepsilon)t \log\left(\frac{\log((1+\varepsilon)t)}{\delta}\right)}.$$

Proof We denote $S_t = \sum_{s=1}^{t} X_s$, and $\psi(x) = \sqrt{2\sigma^2 x \log\left(\frac{\log(x)}{\delta}\right)}$. We also define by induction the sequence of integers (u_k) as follows: $u_0 = 1$, $u_{k+1} = \lceil (1+\varepsilon)u_k \rceil$.

Step 1: Control of $S_{u_k}, k \geq 1$. The following inequalities hold true thanks to an union bound together with Chernoff's bound, the fact that $u_k \geq (1+\varepsilon)^k$, and a simple sum-integral comparison:

$$\begin{aligned}
\mathbb{P}\left(\exists k \geq 1 : S_{u_k} \geq \sqrt{1+\varepsilon}\,\psi(u_k)\right) &\leq \sum_{k=1}^{\infty} \exp\left(-(1+\varepsilon)\log\left(\frac{\log(u_k)}{\delta}\right)\right) \\
&\leq \sum_{k=1}^{\infty} \left(\frac{\delta}{k\log(1+\varepsilon)}\right)^{1+\varepsilon} \leq \left(1+\frac{1}{\varepsilon}\right)\left(\frac{\delta}{\log(1+\varepsilon)}\right)^{1+\varepsilon}.
\end{aligned}$$

Step 2: Control of $S_t, t \in (u_k, u_{k+1})$. Adopting the notation $[n] = \{1, \ldots, n\}$, recall that Hoeffding's maximal inequality[3] states that for any $m \geq 1$ and $x > 0$ one has

$$\mathbb{P}(\exists\, t \in [m] \text{ s.t. } S_t \geq x) \leq \exp\left(-\frac{x^2}{2\sigma^2 m}\right).$$

2. Note δ is restricted to guarantee that $\log(\frac{\log((1+\varepsilon)t)}{\delta})$ is well defined. This makes the analysis cleaner but in practice one can allow the full range of δ by using $\log(\frac{\log((1+\varepsilon)t+2)}{\delta})$ instead and obtain the same theoretical guarantees.
3. It is an easy exercise to verify that Azuma-Hoeffding holds for martingale differences with sub-Gaussian increments, which implies Hoeffding's maximal inequality for sub-Gaussian distributions.

Thus the following inequalities hold true (by using trivial manipulations on the sequence (u_k)):

$$\begin{aligned}
&\mathbb{P}\big(\exists\, t \in \{u_k+1,\dots,u_{k+1}-1\} : S_t - S_{u_k} \geq \sqrt{\varepsilon}\,\psi(u_{k+1})\big) \\
&\quad = \mathbb{P}\left(\exists\, t \in [u_{k+1}-u_k-1] : S_t \geq \sqrt{\varepsilon}\,\psi(u_{k+1})\right) \leq \exp\left(-\varepsilon \tfrac{u_{k+1}}{u_{k+1}-u_k-1} \log\left(\tfrac{\log(u_{k+1})}{\delta}\right)\right) \\
&\quad \leq \exp\left(-(1+\varepsilon)\log\left(\tfrac{\log(u_{k+1})}{\delta}\right)\right) \leq \left(\tfrac{\delta}{(k+1)\log(1+\varepsilon)}\right)^{1+\varepsilon}.
\end{aligned}$$

Step 3: By putting together the results of Step 1 and Step 2 we obtain that with probability at least $1 - \frac{2+\varepsilon}{\varepsilon}\left(\frac{\delta}{\log(1+\varepsilon)}\right)^{1+\varepsilon}$, one has for any $k \geq 0$ and any $t \in \{u_k+1,\dots,u_{k+1}\}$,

$$\begin{aligned}
S_t &= S_t - S_{u_k} + S_{u_k} \\
&\leq \sqrt{\varepsilon}\,\psi(u_{k+1}) + \sqrt{1+\varepsilon}\,\psi(u_k) \\
&\leq \sqrt{\varepsilon}\,\psi((1+\varepsilon)t) + \sqrt{1+\varepsilon}\,\psi(t) \\
&\leq (1+\sqrt{\varepsilon})\,\psi((1+\varepsilon)t),
\end{aligned}$$

which concludes the proof. ■

Without loss of generality we assume that $\mu_1 > \mu_2 \geq \ldots \geq \mu_n$. To shorten notation we denote

$$U(t,\omega) = (1+\sqrt{\varepsilon})\sqrt{\tfrac{2\sigma^2(1+\varepsilon)}{t}\log\left(\tfrac{\log((1+\varepsilon)t)}{\omega}\right)}.$$

The following events will be useful in the analysis:

$$\mathcal{E}_i(\omega) = \{\forall t \geq 1, |\widehat{\mu}_{i,t} - \mu_i| \leq U(t,\omega)\}$$

where $\widehat{\mu}_{i,t} = \frac{1}{t}\sum_{j=1}^{t} x_{i,j}$. Note that Lemma 3 shows $\mathbb{P}(\mathcal{E}_i(\omega)^c) = O(\omega)$. The following trivial inequalities will also be useful (the second one is derived from the first inequality and the fact that $\frac{x+a}{x+b} \leq \frac{a}{b}$ for $a \geq b$, $x \geq 0$). For $t \geq 1$, $\varepsilon \in (0,1)$, $c > 0$, $0 < \omega \leq 1$,

$$\frac{1}{t}\log\left(\frac{\log((1+\varepsilon)t)}{\omega}\right) \geq c \Rightarrow t \leq \frac{1}{c}\log\left(\frac{2\log((1+\varepsilon)/(c\omega))}{\omega}\right), \tag{1}$$

and for $t \geq 1$, $s \geq 3$, $\varepsilon \in (0,1)$, $c \in (0,1]$, $0 < \omega \leq \delta \leq e^{-e}$,

$$\frac{1}{t}\log\left(\frac{\log((1+\varepsilon)t)}{\omega}\right) \geq \frac{c}{s}\log\left(\frac{\log((1+\varepsilon)s)}{\delta}\right) \text{ and } \omega \leq \delta \Rightarrow t \leq \frac{s}{c}\frac{\log\left(2\log\left(\frac{1}{c\omega}\right)/\omega\right)}{\log(1/\delta)}. \tag{2}$$

Lemma 4 *Let $\beta, \varepsilon, \delta$ be set as in Theorem 2 and let $\gamma = 2(2+\beta)^2(1+\sqrt{\varepsilon})^2\sigma^2(1+\varepsilon)$ and $c_\varepsilon = \frac{2+\varepsilon}{\varepsilon}\left(\frac{1}{\log(1+\varepsilon)}\right)^{1+\varepsilon}$. Then we have with probability at least $1 - 2c_\varepsilon\delta$ and any $t \geq 1$,*

$$\sum_{i=2}^{n} T_i(t) \leq n + 5\gamma\mathbf{H}_1\log(e/\delta) + \sum_{i=2}^{n}\gamma\frac{\log(2\max\{1,\log(\gamma(1+\varepsilon)/\Delta_i^2/\delta)\})}{\Delta_i^2}.$$

The proof relies crucially on the fact that the realizations from each arm are independent of each other. This means that if we condition on the event that the realizations from the optimal arm are well-behaved, it is shown that the number of times the ith suboptimal arm is pulled is an independent sub-exponential random variable with mean on the order of $\Delta_i^{-2}\log(\log(\Delta_i^{-2})/\delta)$. We then apply a standard tail bound to the sum of independent sub-exponential random variables to obtain the result.

Proof We decompose the proof in two steps.

Step 1. Let $i > 1$. Assuming that $\mathcal{E}_1(\delta)$ and $\mathcal{E}_i(\omega)$ hold true and that $I_t = i$ one has

$$\mu_i + U(T_i(t),\omega) + (1+\beta)U(T_i(t),\delta) \geq \widehat{\mu}_{i,T_i(t)} + (1+\beta)U(T_i(t),\delta) \geq \widehat{\mu}_{1,T_1(t)} + (1+\beta)U(T_1(t),\delta) \geq \mu_1,$$

which implies $(2+\beta)U(T_i(t),\min(\omega,\delta)) \geq \Delta_i$. If $\gamma = 2(2+\beta)^2(1+\sqrt{\varepsilon})^2\sigma^2(1+\varepsilon)$ then using (1) with $c = \frac{\Delta_i^2}{\gamma}$ one obtains that if $\mathcal{E}_1(\delta)$ and $\mathcal{E}_i(\omega)$ hold true and $I_t = i$ then

$$\begin{aligned} T_i(t) &\leq \frac{\gamma}{\Delta_i^2}\log\left(\frac{2\log(\gamma(1+\varepsilon)/\Delta_i^2/\min(\omega,\delta))}{\min(\omega,\delta)}\right) \\ &\leq \tau_i + \frac{\gamma}{\Delta_i^2}\log\left(\frac{\log(e/\omega)}{\omega}\right) \leq \tau_i + \frac{2\gamma}{\Delta_i^2}\log\left(\frac{1}{\omega}\right), \end{aligned}$$

where $\tau_i = \frac{\gamma}{\Delta_i^2}\log\left(\frac{2\max\{1,\log(\gamma(1+\varepsilon)/\Delta_i^2/\delta)\}}{\delta}\right)$.

Since $T_i(t)$ only increases when i is played the above argument shows that the following inequality is true for any time $t \geq 1$:

$$T_i(t)\mathbb{1}\{\mathcal{E}_1(\delta)\cap\mathcal{E}_i(\omega)\} \leq 1 + \tau_i + \frac{2\gamma}{\Delta_i^2}\log\left(\frac{1}{\omega}\right). \tag{3}$$

Step 2. We define the following random variable:

$$\Omega_i = \max\{\omega \geq 0 : \mathcal{E}_i(\omega) \text{ holds true}\}.$$

Note that Ω_i is well-defined and by Lemma 3 it holds that $\mathbb{P}(\Omega_i < \omega) \leq c_\varepsilon\omega$ where $c_\varepsilon = \frac{2+\varepsilon}{\varepsilon}\left(\frac{1}{\log(1+\varepsilon)}\right)^{1+\varepsilon}$. Furthermore one can rewrite (3) as

$$T_i(t)\mathbb{1}\{\mathcal{E}_1(\delta)\} \leq 1 + \tau_i + \frac{2\gamma}{\Delta_i^2}\log\left(\frac{1}{\Omega_i}\right). \tag{4}$$

We use this equation as follows:

$$\begin{aligned} \mathbb{P}\left(\sum_{i=2}^n T_i(t) > x + \sum_{i=2}^n(\tau_i+1)\right) &\leq c_\varepsilon\delta + \mathbb{P}\left(\sum_{i=2}^n T_i(t) > x + \sum_{i=2}^n(\tau_i+1)\,\middle|\,\mathcal{E}_1(\delta)\right) \\ &\leq c_\varepsilon\delta + \mathbb{P}\left(\sum_{i=2}^n \frac{2\gamma}{\Delta_i^2}\log\left(\frac{1}{\Omega_i}\right) > x\right). \end{aligned} \tag{5}$$

Let $Z_i = \frac{2\gamma}{\Delta_i^2}\log\left(\frac{c_\varepsilon^{-1}}{\Omega_i}\right)$, $i \in [n]\setminus 1$. Observe that these are independent random variables and since $\mathbb{P}(\Omega_i < \omega) \leq c_\varepsilon\omega$ it holds that $\mathbb{P}(Z_i > x) \leq \exp(-x/a_i)$ with $a_i = 2\gamma/\Delta_i^2$. Using standard

techniques to bound the sum of sub-exponential random variables one directly obtains that

$$\mathbb{P}\left(\sum_{i=2}^{n}(Z_i - a_i) \geq z\right) \leq \exp\left(-\min\left\{\frac{z^2}{4\|a\|_2^2}, \frac{z}{4\|a\|_\infty}\right\}\right) \leq \exp\left(-\min\left\{\frac{z^2}{4\|a\|_1^2}, \frac{z}{4\|a\|_1}\right\}\right). \tag{6}$$

Putting together (5) and (6) with $z = 4\|a\|_1 \log(1/(c_\varepsilon \delta))$, $x = z + ||a||_1 \log(ec_\varepsilon)$ one obtains

$$\mathbb{P}\left(\sum_{i=2}^{n} T_i(t) > \sum_{i=2}^{n}\left(\frac{4\gamma \log(e/\delta)}{\Delta_i^2} + \tau_i + 1\right)\right) \leq 2c_\varepsilon \delta,$$

which concludes the proof. ■

Lemma 5 *Let $\beta, \varepsilon, \delta$ be set as in Theorem 2 and let $c_\varepsilon = \frac{2+\varepsilon}{\varepsilon}\left(\frac{1}{\log(1+\varepsilon)}\right)^{1+\varepsilon}$. If*

$$\lambda \geq \frac{1 + \frac{\log\left(2\log\left(\left(\frac{2+\beta}{\beta}\right)^2/\delta\right)\right)}{\log(1/\delta)}}{1 - (c_\varepsilon\delta) - \sqrt{(c_\varepsilon\delta)^{1/4}\log(1/(c_\varepsilon\delta))}}\left(\frac{2+\beta}{\beta}\right)^2, \tag{7}$$

then for all $i = 2, \ldots, n$ and $t = 1, 2, \ldots$, we have $T_i(t) < 1 + \lambda \sum_{j \neq i} T_j(t)$ with probability at least $1 - 2c_\varepsilon\delta + 4\sqrt{c_\varepsilon\delta}$.

Note that the right hand side of (7) can be bound by a universal constant for all allowable δ which leads to the simplified statement of Theorem 2. Moreover, for any $\nu > 0$ there exists a sufficiently small $\delta \in (0, 1)$ such that the right hand side of (7) is less than or equal to $(1 + \nu)\left(\frac{2+\beta}{\beta}\right)^2$.

Essentially, the proof relies on the fact that given any two arms $j < i$ (i.e. $\mu_j \geq \mu_i$), $T_i(t)$ cannot be larger than a constant times $T_j(t)$ with probability at least $1 - \delta$. Considering this fact, it is reasonable to suppose that the probability that $T_i(t)$ is larger than a constant times $\sum_{j=1}^{i-1} T_j(T)$ is decreasing exponentially fast in i. Consequently, our stopping condition is not based on a uniform confidence bound for all arms. Rather, it is based on confidence bounds that grow in size as the arm index i increases.

Proof We decompose the proof in two steps.

Step 1. Let $i > j$. Assuming that $\mathcal{E}_i(\omega)$ and $\mathcal{E}_j(\delta)$ hold true and that $I_t = i$ one has

$$\begin{aligned}\mu_i + U(T_i(t), \omega) + (1+\beta)U(T_i(t), \delta) &\geq \widehat{\mu}_{i,T_i(t)} + (1+\beta)U(T_i(t), \delta) \\ &\geq \widehat{\mu}_{j,T_j(t)} + (1+\beta)U(T_j(t), \delta) \geq \mu_j + \beta U(T_j(t), \delta),\end{aligned}$$

which implies $(2+\beta)U(T_i(t), \min(\omega, \delta)) \geq \beta U(T_j(t), \delta)$. Thus using (2) with $c = \left(\frac{\beta}{2+\beta}\right)^2$ one obtains that if $\mathcal{E}_i(\omega)$ and $\mathcal{E}_j(\delta)$ hold true and $I_t = i$ then

$$T_i(t) \leq \left(\frac{2+\beta}{\beta}\right)^2 \frac{\log\left(2\log\left(\left(\frac{2+\beta}{\beta}\right)^2/\min(\omega,\delta)\right)/\min(\omega,\delta)\right)}{\log(1/\delta)} T_j(t).$$

Similarly to Step 1 in the proof of Lemma 4 we use the fact that $T_i(t)$ only increases when I_t is played and the above argument to obtain the following inequality for any time $t \geq 1$:

$$(T_i(t) - 1)\mathbb{1}\{\mathcal{E}_i(\omega) \cap \mathcal{E}_j(\delta)\} \leq \left(\frac{2+\beta}{\beta}\right)^2 \frac{\log\left(2\log\left(\left(\frac{2+\beta}{\beta}\right)^2/\min(\omega,\delta)\right)/\min(\omega,\delta)\right)}{\log(1/\delta)} T_j(t). \tag{8}$$

Step 2. Using (8) with $\omega = \delta^{i-1}$ we see that

$$\mathbb{1}\{\mathcal{E}_i(\delta^{i-1})\}\frac{1}{i-1}\sum_{j=1}^{i-1}\mathbb{1}\{\mathcal{E}_j(\delta)\} > 1-\alpha \;\Rightarrow\; (1-\alpha)(T_i(t)-1) \leq \kappa\sum_{j\neq i} T_j(t)$$

where $\kappa = \left(\frac{2+\beta}{\beta}\right)^2\left(1+\frac{\log\left(2\log\left(\left(\frac{2+\beta}{\beta}\right)^2/\delta\right)\right)}{\log(1/\delta)}\right)$. This implies the following, using that $\mathbb{P}(\mathcal{E}_i(\omega)) \geq 1 - c_\varepsilon\omega$,

$$\begin{aligned}
&\mathbb{P}\left(\exists\,(i,t)\in\{2,\dots,n\}\times\{1,\dots\} : (1-\alpha)(T_i(t)-1) \geq \kappa\sum_{j\neq i}T_j(t)\right)\\
&\leq \mathbb{P}\left(\exists\, i\in\{2,\dots,n\} : \mathbb{1}\{\mathcal{E}_i(\delta^{i-1})\}\frac{1}{i-1}\sum_{j=1}^{i-1}\mathbb{1}\{\mathcal{E}_j(\delta)\} \leq 1-\alpha\right)\\
&\leq \sum_{i=2}^{n}\mathbb{P}(\mathcal{E}_i(\delta^{i-1}) \text{ does not hold}) + \sum_{i=2}^{n}\mathbb{P}\left(\frac{1}{i-1}\sum_{j=1}^{i-1}\mathbb{1}\{\mathcal{E}_j(\delta)\} \leq 1-c_\varepsilon\delta-(\alpha-c_\varepsilon\delta)\right).
\end{aligned}$$

Let $\delta' = c_\varepsilon\delta$. Note that by a simple Hoeffding's inequality and a union bound one has

$$\mathbb{P}\left(\frac{1}{i-1}\sum_{j=1}^{i-1}\mathbb{1}\{\mathcal{E}_j(\delta)\} \leq 1-\delta'-(\alpha-\delta')\right) \leq \min((i-1)\delta', \exp(-2(i-1)(\alpha-\delta')^2),$$

and thus if we define $j_* = \lceil \delta'^{-1/4}/2\rceil$ we obtain with the above calculations

$$\begin{aligned}
&\mathbb{P}\left(\exists\,(i,t)\in\{2,\dots,n\}\times\{1,\dots\} : \left(1-\delta'-\sqrt{\delta'^{1/4}\log(1/\delta')}\right)(T_i(t)-1) \geq \kappa\sum_{j\neq i}T_j(t)\right)\\
&\leq \sum_{i=2}^{n}\left(\delta'^{i-1} + \min\left((i-1)\delta', e^{-2(i-1)\delta'^{1/4}\log\left(\frac{1}{\delta'}\right)}\right)\right) \leq \frac{\delta'}{1-\delta'} + \delta' j_*^2 + \frac{e^{-2j_*\delta'^{1/4}\log\left(\frac{1}{\delta'}\right)}}{1-e^{-2\delta'^{1/4}\log\left(\frac{1}{\delta'}\right)}}\\
&\leq \frac{\delta'}{1-\delta'} + \tfrac{9}{4}\delta'^{1/2} + \tfrac{3}{2}\delta'^{3/4} \leq 2c_\varepsilon\delta + 4\sqrt{c_\varepsilon\delta}.
\end{aligned}$$

■

Treating ε, σ^2 and factors of $\log\log(\beta)$ as constants, Lemma 4 says that the total number of times the suboptimal arms are sampled does not exceed $(\beta+2)^2\left(c_1\mathbf{H}_1\log(1/\delta) + c_3\mathbf{H}_3\right)$. Lemma 5 states that only the optimal arm will meet the stopping condition with $\lambda = c_\lambda\left(\frac{2+\beta}{\beta}\right)^2$ for some c_λ constant defined in the lemma. Combining these results, we observe that the total number of times all the arms are sampled does not exceed $(\beta+2)^2\left(c_1\mathbf{H}_1\log(1/\delta) + c_3\mathbf{H}_3\right)\left(1+c_\lambda\left(\frac{2+\beta}{\beta}\right)^2\right)$, completing the proof of the theorem. We also observe using the approximation $c_\lambda = 1$, the optimal choice of $\beta \approx 1.66$.

5. Implementation and Simulations

In this section we investigate how the state of the art methods for solving the best arm problem behave in practice. Before describing each of the algorithms in the comparison, we briefly describe a LIL-based stopping criterion that can be applied to any of the algorithms.

LIL Stopping (LS) : For any algorithm and $i \in [n]$, after the t-th time we have that the i-th arm has been sampled $T_i(t)$ times and accumulated a mean $\widehat{\mu}_{i,T_i(t)}$. We can apply Lemma 3 (with a union bound) so that with probability at least $1 - \frac{2+\varepsilon}{\varepsilon}\left(\frac{\delta}{\log(1+\varepsilon)}\right)^{1+\varepsilon}$

$$\left|\widehat{\mu}_{i,T_i(t)} - \mu_i\right| \leq B_{i,T_i(t)} := (1+\sqrt{\varepsilon})\sqrt{\frac{2\sigma^2(1+\varepsilon)\log\left(\frac{2\log((1+\varepsilon)T_i(t)+2)}{\delta/n}\right)}{T_i(t)}} \tag{9}$$

for all $t \geq 1$ and all $i \in [n]$. We may then conclude that if $\widehat{i} := \arg\max_{i\in[n]} \widehat{\mu}_{i,T_i(t)}$ and $\widehat{\mu}_{\widehat{i},T_{\widehat{i}}(t)} - B_{\widehat{i},T_{\widehat{i}}(t)} \geq \widehat{\mu}_{j,T_j(t)} + B_{j,T_j(t)}\ \forall j \neq \widehat{i}$ then with high probability we have that $\widehat{i} = i_*$.

The LIL stopping condition is somewhat naive but often quite effective in practice for smaller size problems when $\log(n)$ is negligible. To implement the strategy for any algorithm with fixed confidence ν, simply run the algorithm with $\nu/2$ in place of ν and assign the other $\nu/2$ confidence to the LIL stopping criterion. Note that to for the LIL bound to hold with probability at least $1-\nu$, one should use $\delta = \log(1+\varepsilon)\left(\frac{\nu\varepsilon}{2+\varepsilon}\right)^{1/(1+\varepsilon)}$. The algorithms compared were:

- *Nonadaptive + LS* : Draw a random permutation of $[n]$ and sample the arms in an order defined by cycling through the permutation until the LIL stopping criterion is met.
- *Exponential-Gap Elimination (+LS)* (Karnin et al., 2013) : This procedure proceeds in stages where at each stage, *median elimination* (Even-Dar et al., 2002) is used to find an ε-optimal arm whose mean is guaranteed (with large probability) to be within a specified $\varepsilon > 0$ of the mean of the best arm, and then arms are discarded if their empirical mean is sufficiently below the empirical mean of the ε-optimal arm. The algorithm terminates when there is only one arm that has not yet been discarded (or when the LIL stopping criterion is met).
- *Successive Elimination* (Even-Dar et al., 2002) : This procedure proceeds in the same spirit as *Exponential-Gap Elimination* except the ε-optimal arm is equal to $\widehat{i} := \arg\max_{i\in[n]} \widehat{\mu}_{i,T_i(t)}$.
- *lil'UCB (+LS)* : The procedure of Figure 1 is run with $\varepsilon = 0.01$, $\beta = 1$, $\lambda = (2+\beta)^2/\beta^2 = 9$, and $\delta = \frac{(\sqrt{1+\nu(/2)}-1)^2}{4c_\varepsilon}$ for input confidence ν. The algorithm terminates according to Fig. 1 (or when the LIL stopping criterion is met). Note that δ is defined as prescribed by Theorem 2 but we approximate the leading constant in (7) by 1 to define λ.
- *lil'UCB Heuristic* : The procedure of Figure 1 is run with $\varepsilon = 0$, $\beta = 1/2$, $\lambda = 1 + 10/n$, and $\delta = \nu/5$ for input confidence ν. These parameter settings do not satisfy the conditions of Theorem 2, and thus there is no guarantee that this algorithm will find the best arm.
- *LUCB1 (+ LS)* (Kalyanakrishnan et al., 2012) : This procedure pulls two arms at each time: the arm with the highest empirical mean and the arm with the highest upper confidence bound among the remaining arms. The upper confidence bound was of the form prescribed in the simulations section of Kaufmann and Kalyanakrishnan (2013) and is guaranteed to return the arm with the highest mean with confidence $1-\delta$.

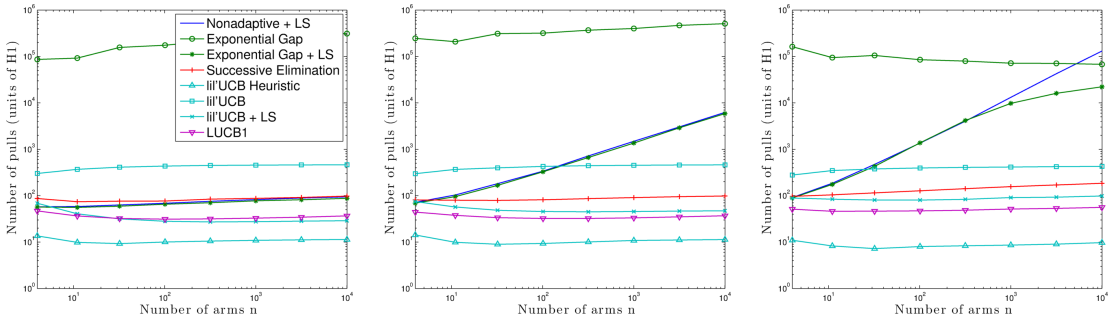
We did not compare to *PRISM* of Jamieson et al. (2013) because the algorithm and its empirical performance are very similar to *Exponential-Gap Elimination* so its inclusion in the comparison would provide very little added value. We remark that the first three algorithms require $O(1)$ amortized computation per time step, the lil'UCB algorithms require $O(\log(n))$ computation per time step using smart data structures[4], and LUCB1 requires $O(n)$ computation per time step. LUCB1 was not run on all problem sizes due to poor computational scaling with respect to the problem size.

Three problem scenarios were considered over a variety problem sizes (number of arms). The "1-sparse" scenario sets $\mu_1 = 1/2$ and $\mu_i = 0$ for all $i = 2, \ldots, n$ resulting in a hardness of $\mathbf{H}_1 = 4n$. The "$\alpha = 0.3$" and "$\alpha = 0.6$" scenarios consider $n + 1$ arms with $\mu_0 = 1$ and $\mu_i = 1 - (i/n)^\alpha$ for all $i = 1, \ldots, n$ with respective hardnesses of $\mathbf{H}_1 \approx 3/2n$ and $\mathbf{H}_1 \approx 6n^{1.2}$. That is, the $\alpha = 0.3$ case should be about as hard as the sparse case with increasing problem size while the $\alpha = 0.6$ is considerably more challenging and grows super linearly with the problem size. See Jamieson et al. (2013) for an in-depth study of the α parameterization. All experiments were run with input confidence $\delta = 0.1$. All realizations of the arms were Gaussian random variables with mean μ_i and variance $1/4$[5].

Each algorithm terminates at some finite time with high probability so we first consider the relative stopping times of each of the algorithms in Figure 2. Each algorithm was run on each problem scenario and problem size, repeated 50 times. The first observation is that *Exponential-Gap Elimination (+LS)* appears to barely perform better than nonadaptive sampling with the LIL stopping criterion. This confirms our suspicion that the constants in *median elimination* are just too large to make this algorithm practically relevant. While the LIL stopping criterion seems to have measurably improved the *lil'UCB* algorithm, it had no impact on the *lil'UCB Heuristic* variant (not plotted). While *lil'UCB Heuristic* has no theoretical guarantees of outputting the best arm, we remark that over the course of all of our tens of thousands of experiments, the algorithm never failed to terminate with the best arm. The *LUCB* algorithm, despite having worse theoretical guarantees than the *lil'UCB* algorithm, performs surprisingly well. We conjecture that this is because UCB style algorithms tend to lean towards exploiting the top arm versus focusing on increasing the gap between the top two arms, which is the goal of *LUCB*.

In reality, one cannot always wait for an algorithm to run until it terminates on its own so we now explore how the algorithms perform if the algorithm must output an arm at every time step before termination (this is similar to the setting studied in Bubeck et al. (2009)). For each algorithm, at each time we output the arm with the highest empirical mean. Clearly, the probability that a sub-optimal arm is output by any algorithm should very close to 1 in the beginning but then eventually decrease to at least the desired input confidence, and likely, to zero. Figure 3 shows the "anytime" performance of the algorithms for the three scenarios and unlike the empirical stopping times of the algorithms, we now observe large differences between the algorithms. Each experiment was repeated 5000 times. Again we see essentially no difference between nonadaptive sampling and the exponential-gap procedure. While in the stopping time plots of Figure 2 the *successive elimination* appears competitive with the UCB algorithms, we observe in Figure 3 that the UCB algorithms are collecting

4. The sufficient statistic for lil'UCB to decide which arm to sample depends only on $\widehat{\mu}_{i,T_i(t)}$ and $T_i(t)$ which only changes for an arm if that particular arm is pulled. Thus, it suffices to maintain an ordered list of the upper confidence bounds in which deleting, updating, and reinserting the arm requires just $O(\log(n))$ computation. Contrast this with a UCB procedure in which the upper confidence bounds depend explicitly on t so that the sufficient statistics for pulling the next arm changes for all arms after each pull, requiring $\Omega(n)$ computation per time step.
5. The variance was chosen such that the analyses of algorithms that assumed realizations were in $[0, 1]$ and used Hoeffding's inequality were still valid using sub-Gaussian tail bounds with scale parameter $1/2$.



1-sparse, $\mathbf{H}_1 = 4n$ $\alpha = 0.3$, $\mathbf{H}_1 \approx \frac{3}{2}n$ $\alpha = 0.6$, $\mathbf{H}_1 \approx 6n^{1.2}$

Figure 2: Stopping times of the algorithms for three scenarios for a variety of problem sizes. The problem scenarios from left to right are the 1-sparse problem ($\mu_1 = 0.5$, $\mu_i = 0 \; \forall i > 1$), $\alpha = 0.3$ ($\mu_i = 1 - (i/n)^\alpha$, $i = 0, 1, \ldots, n$), and $\alpha = 0.6$.

sufficient information to output the best arm at least twice as fast as *successive elimination*. This tells us that the stopping conditions for the UCB algorithms are still too conservative in practice which motivates the use of the *lil'UCB Heuristic* algorithm which appears to perform very strongly across all metrics. The *LUCB* algorithm again performs strongly here suggesting that LUCB-style algorithms are very well-suited for exploration tasks.

6. Discussion

This paper proposed a new procedure for identifying the best arm in a multi-armed bandit problem in the fixed confidence setting, a problem of pure exploration. However, there are some scenarios where one wishes to balance exploration with exploitation and the metric of interest is the *cumulative regret*. We remark that the techniques developed here can be easily extended to show that the lil'UCB algorithm obtains bounded regret with high probability, improving upon the result of Abbasi-Yadkori et al. (2011).

In this work we proved upper and lower bounds over the class of distributions with bounded means and sub-Guassian realizations and presented our results just in terms of the difference between the means of the arms. In contrast to just considering the means of the distributions, Kaufmann and Kalyanakrishnan (2013) studied the Chernoff information between distributions, a quantity related to the KL divergence, that is sharper and can result in improved rates in identifying the best arm in theory and practice (for instance if the realizations from the arms have very different variances). Pursuing methods that exploit distributional characteristics beyond the mean is a good direction for future work.

Finally, an obvious extension of this work is to consider finding the top-m arms instead of just the best arm. This idea has been explored in both the fixed confidence setting Kaufmann and Kalyanakrishnan (2013) and the fixed budget setting Bubeck et al. (2012) but we believe both of these sample complexity results to be suboptimal. It may be possible to adapt the approach developed in this paper to find the top-m arms and obtain gains in theory and practice.

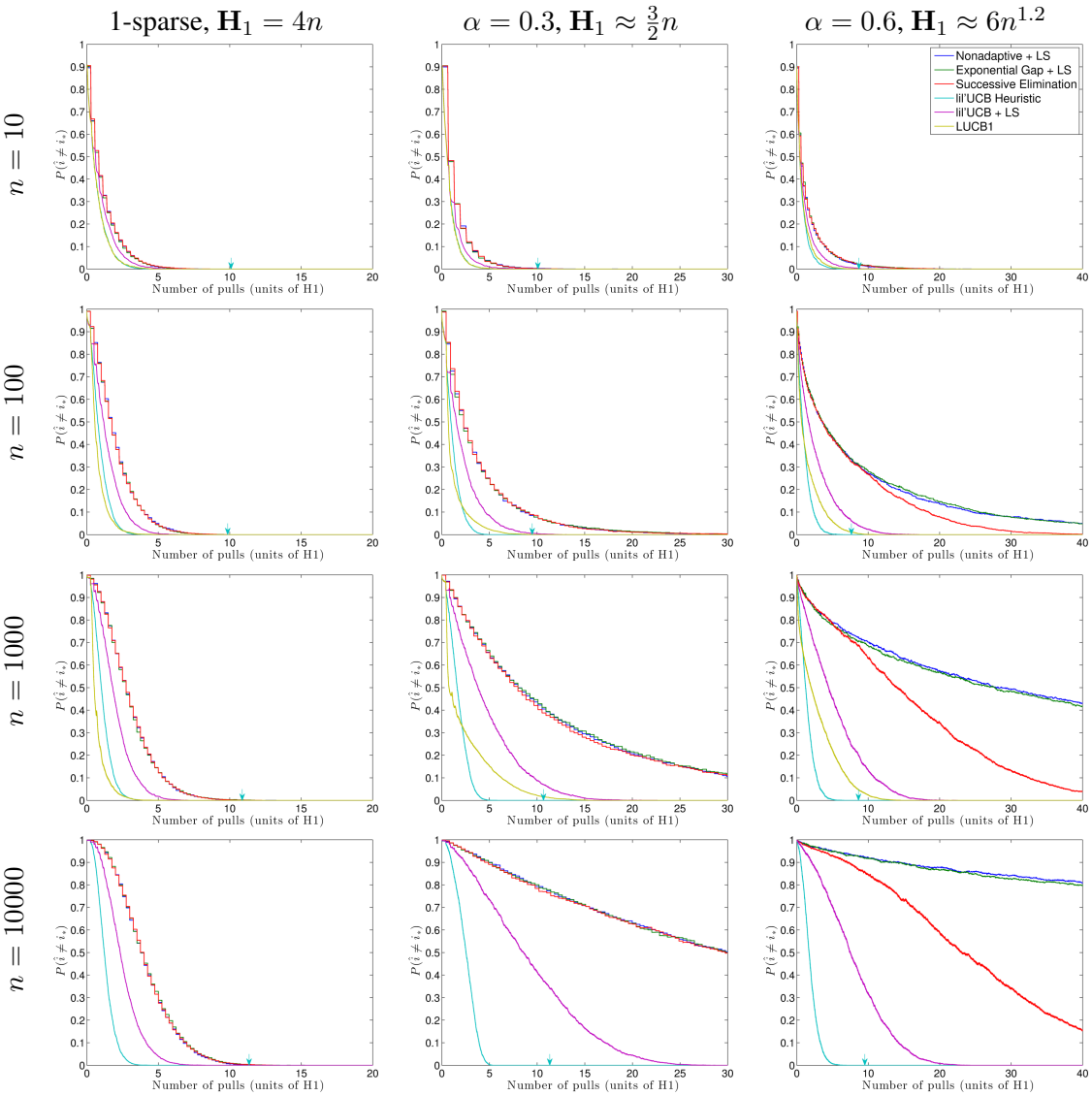


Figure 3: At every time, each algorithm outputs an arm $\hat{i}$ that has the highest empirical mean. The $\mathbb{P}(\hat{i} \neq i_*)$ is plotted with respect to the total number of pulls by the algorithm. The problem sizes (number of arms) increase from top to bottom. The problem scenarios from left to right are the 1-sparse problem ($\mu_1 = 0.5$, $\mu_i = 0\ \forall i > 1$) , $\alpha = 0.3$ ($\mu_i = 1 - (i/n)^\alpha$, $i = 0, 1, \ldots, n$), and $\alpha = 0.6$. The arrows indicate the stopping times (if not shown, those algorithms did not terminate within the time window shown). Note that LUCB1 is not plotted for $n = 10000$ due to computational constraints (see text for explanation). Also note that in some plots it is difficult to distinguish between the nonadaptive sampling procedure, the exponential-gap algorithm, and successive elimination due to the curves being on top of each other.

References

Yasin Abbasi-Yadkori, Csaba Szepesvári, and David Tax. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320, 2011.

Jean-Yves Audibert, Sébastien Bubeck, and Rémi Munos. Best arm identification in multi-armed bandits. *COLT 2010-Proceedings*, 2010.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.

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.

S. Bubeck, R. Munos, and G. Stoltz. Pure exploration in multi-armed bandits problems. In *Proceedings of the 20th International Conference on Algorithmic Learning Theory (ALT)*, 2009.

Sébastien Bubeck, Tengyao Wang, and Nitin Viswanathan. Multiple identifications in multi-armed bandits. *arXiv preprint arXiv:1205.3181*, 2012.

DA Darling and Herbert Robbins. Iterated logarithm inequalities. In *Herbert Robbins Selected Papers*, pages 254–258. Springer, 1985.

Eyal Even-Dar, Shie Mannor, and Yishay Mansour. PAC bounds for multi-armed bandit and markov decision processes. In *Computational Learning Theory*, pages 255–270. Springer, 2002.

R. H. Farrell. Asymptotic behavior of expected sample size in certain one sided tests. *The Annals of Mathematical Statistics*, 35(1):pp. 36–72, 1964. ISSN 00034851.

Victor Gabillon, Mohammad Ghavamzadeh, Alessandro Lazaric, and Team SequeL. Best arm identification: A unified approach to fixed budget and fixed confidence. In *NIPS*, pages 3221–3229, 2012.

Kevin Jamieson, Matthew Malloy, Robert Nowak, and Sebastien Bubeck. On finding the largest mean among many. *arXiv preprint arXiv:1306.3917*, 2013.

Shivaram Kalyanakrishnan, Ambuj Tewari, Peter Auer, and Peter Stone. PAC subset selection in stochastic multi-armed bandits. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pages 655–662, 2012.

Zohar Karnin, Tomer Koren, and Oren Somekh. Almost optimal exploration in multi-armed bandits. In *Proceedings of the 30th International Conference on Machine Learning*, 2013.

Emilie Kaufmann and Shivaram Kalyanakrishnan. Information complexity in bandit subset selection. COLT, 2013.

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

Edward Paulson. A sequential procedure for selecting the population with the largest mean from k normal populations. *The Annals of Mathematical Statistics*, 35(1):174–180, 1964.

Appendix A. Condensed Proof of Lower Bound

We first re-state the main result of Farrell (1964).

Theorem 6 *(Farrell, 1964, Theorem 1). Let* $X_i \overset{i.i.d.}{\sim} \mathcal{N}(\Delta, 1)$*, where* $\Delta \neq 0$ *is unknown. Consider testing whether* $\Delta > 0$ *or* $\Delta < 0$*. Let* $Y \in \{-1, 1\}$ *be the decision of any such test based on* T *samples (possibly a random number) and let* $\delta \in (0, 1/2)$*. If* $\sup_{\Delta \neq 0} \mathbb{P}(Y \neq sign(\Delta)) \leq \delta$*, then*

$$\limsup_{\Delta \to 0} \frac{\mathbb{E}_\Delta[T]}{\Delta^{-2} \log \log \Delta^{-2}} \geq 2 - 4\delta.$$

In the following we show a weaker result than what is shown in Farrell (1964); nonetheless, it shows the $\log \log$ term is necessary.

Theorem 7 *Let* $X_i \overset{i.i.d.}{\sim} \mathcal{N}(\Delta, 1)$*, where* $\Delta \neq 0$ *is unknown. Consider testing whether* $\Delta > 0$ *or* $\Delta < 0$*. Let* $Y \in \{-1, 1\}$ *be the decision of any such test based on* T *samples (possibly a random number). If* $\sup_{\Delta \neq 0} \mathbb{P}(Y \neq \text{sign}(\Delta)) < 1/2$*, then*

$$\limsup_{\Delta \to 0} \frac{\mathbb{E}[T]}{\Delta^{-2} \log \log \Delta^{-2}} \quad > \quad 0\,.$$

We rely on two intuitive facts, each which justified more formally in Farrell (1964).

Fact 1. The form of *an* optimal test is a *generalized sequential probability ratio test* (GSPRT), which continues sampling while

$$-B_t \leq \sum_{j=1}^{t} X_i \leq B_t$$

and stops otherwise, declaring $\Delta > 0$ if $\sum_{j=1}^{t} X_j \geq B_t$, and $\Delta < 0$ if $\sum_{j=1}^{t} X_j \leq -B_t$ where $B_t > 0$ is non-decreasing in t. This is made formal in Farrell (1964).

Fact 2. If

$$\lim_{t \to \infty} \frac{B_t}{\sqrt{2t \log \log t}} \leq 1 \tag{10}$$

then Y, the decision output by the GSPRT, satisfies $\sup_{\Delta \neq 0} \mathbb{P}_\Delta(Y \neq \text{sign}\,\Delta) = 1/2$. This follows from the LIL and a continuity argument (and note the limit exists as B_t is non-decreasing). Intuitively, if the thresholds satisfy (10), a zero mean random walk will eventually hit either the upper or lower threshold. The upper threshold is crossed first with probability one half, as is the lower. By arguing that the error probabilities are continuous functions of Δ, one concludes this assertion is true.

The argument proceeds as follows. If (10) is holds, then the error probability is $1/2$. So we can focus on threshold sequences satisfying $\lim_{t \to \infty} \frac{B_t}{\sqrt{2t \log \log t}} \geq (1 + \varepsilon)$ for some $\varepsilon > 0$. In other words, for all $t > t_1$ some $\varepsilon > 0$, some sufficiently large t_1

$$B_t \geq (1 + \varepsilon)\sqrt{2t \log \log t}.$$

Define the function

$$t_0(\Delta) = \frac{\varepsilon^2 \Delta^{-2}}{2} \log\log\left(\frac{\Delta^{-2}}{2}\right)$$

and let T be the stopping time:

$$T := \inf\left\{ t \in \mathbb{N} : \left| \sum_{i=1}^{t} X_i \right| \geq B_t \right\}.$$

Let $S_t^{(\Delta)} = \sum_{j=1}^{t} X_j$ for $X_j \overset{iid}{\sim} \mathcal{N}(\Delta, 1)$. Without loss of generality, assume $\Delta > 0$. Additionally, suppose Δ is sufficiently small, such that both $t_0(\Delta) > t_1(\varepsilon)$ and $\Delta \leq \varepsilon$ (in the following steps we consider the limit as $\Delta \to 0$). We have

$$\begin{aligned}
&\mathbb{P}_\Delta(T \geq t_0(\Delta)) \\
&= \mathbb{P}\left(\bigcap_{t=1}^{t_0(\Delta)-1} |S_t^{(\Delta)}| < B_t \right) \\
&= \mathbb{P}\left(\bigcap_{t=1}^{t_1(\varepsilon)} \{|S_t^{(\Delta)}| < B_t\} \cap \bigcap_{t=t_1(\varepsilon)+1}^{t_0(\Delta)-1} \{S_t^{(0)} < B_t - \Delta t\} \cap \{S_t^{(0)} > -B_t - \Delta t\} \right) \\
&\geq \mathbb{P}\left(\bigcap_{t=1}^{t_1(\varepsilon)} \{|S_t^{(\Delta)}| < B_t\} \cap \bigcap_{t=t_1(\varepsilon)+1}^{t_0(\Delta)-1} \{|S_t^{(0)}| < (1+\varepsilon/2)\sqrt{2t\log\log t}\} \right) \qquad (11) \\
&= \mathbb{P}\left(\bigcap_{t=1}^{t_1(\varepsilon)} |S_t^{(\Delta)}| < B_t \right) \mathbb{P}\left(\bigcap_{t=t_1(\varepsilon)+1}^{t_0(\Delta)-1} |S_t^{(0)}| \leq (1+\varepsilon/2)\sqrt{2t\log\log t} \,\middle|\, \bigcap_{t=1}^{t_1(\varepsilon)} |S_t^{(0)}| < B_t \right) \\
&\geq \mathbb{P}\left(\bigcap_{t=1}^{t_1(\varepsilon)} |S_t^{(\Delta)}| < B_t \right) \mathbb{P}\left(\bigcap_{t=t_1(\varepsilon)+1}^{\infty} |S_t^{(0)}| < (1+\varepsilon/2)\sqrt{2t\log\log t} \right) \qquad (12)
\end{aligned}$$

where (11) holds when $\varepsilon \geq \Delta$ and (12) holds by removing the conditioning, and then by increasing the number of terms in the intersection. To see that (11) holds, note that $\frac{2\log\log t}{t} \geq \left(\frac{2\Delta}{\varepsilon}\right)^2$ for all $t \leq t_0(\Delta)$, which is easily verified when $\varepsilon \geq \Delta$ since

$$\frac{\log\log\left(\frac{\varepsilon^2\Delta^{-2}}{2}\log\log\left(\frac{\Delta^{-2}}{2}\right)\right)}{\log\log\left(\frac{\Delta^{-2}}{2}\right)} \quad \geq \quad 1.$$

Taking the limit as $\Delta \to 0$, for any $\varepsilon > 0$, gives

$$\lim_{\Delta\to 0} \mathbb{P}_\Delta(T \geq t_0(\Delta)) \quad \geq \quad c(\varepsilon) > 0$$

where $c(\varepsilon)$ is a non-zero constant, and the inequality follows from (12), as the first term is non-zero for any Δ (including $\Delta = 0$) since $t_1(\varepsilon) < \infty$ and $B_t > 0$, and the second term is non-zero by the LIL for any $\varepsilon > 0$. Note that a finite bound on the second term can be obtained as in Section 2.

By Markov, $\mathbb{E}_\Delta[T]/t_0(\Delta) \geq \mathbb{P}_\Delta(T \geq t_0(\Delta))$, and we conclude

$$\lim_{\Delta \to 0} \frac{\mathbb{E}_\Delta[T]}{\Delta^{-2} \log\log \Delta^{-2}} \geq \varepsilon^2 \, c(\varepsilon) > 0$$

for any test with $\sup_{\Delta \neq 0} \mathbb{P}(Y \neq \text{sign}(\Delta)) < 1/2$.