Almost Optimal Exploration in Multi-Armed Bandits

Zohar Karnin ZKARNIN@YMAIL.COM
Yahoo! Labs, Haifa, Israel

Tomer Koren[†] TOMERK@TECHNION.AC.IL
Technion—Israel Institute of Technology, Haifa, Israel

Oren Somekh ORENS@YAHOO-INC.COM
Yahoo! Labs, Haifa, Israel

Abstract

We study the problem of exploration in stochastic Multi-Armed Bandits. Even in the simplest setting of identifying the best arm, there remains a logarithmic multiplicative gap between the known lower and upper bounds for the number of arm pulls required for the task. This extra logarithmic factor is quite meaningful in nowadays large-scale applications. We present two novel, parameter-free algorithms for identifying the best arm, in two different settings: given a target confidence and given a target budget of arm pulls, for which we prove upper bounds whose gap from the lower bound is only doubly-logarithmic in the problem parameters. We corroborate our theoretical results with experiments demonstrating that our algorithm outperforms the state-of-the-art and scales better as the size of the problem increases.

1. Introduction

Exploration problems in stochastic Multi-Armed Bandits (MAB) have received much attention recently (see e.g. Even-Dar et al. 2006; Bubeck et al. 2009; Audibert et al. 2010; Gabillon et al. 2012). In the best-arm identification problem, the player repeatedly chooses one arm (corresponding to an action), and receives a reward drawn from a fixed probability distribution corresponding to the chosen arm. At the end of this exploration phase, the player must commit to a single arm. Unlike the standard MAB setup where the player's strategy is evaluated in terms of its regret (e.g. Lai & Robbins 1985; Auer et al. 2002), in the best-arm identification setup the goal of the player is to maximize her probability of choosing the best arm having the maximal expected reward.

The complexity of a MAB exploration problem is determined by the total number of arms and difference between their expected rewards and that of the optimal arm. To date, in the setting where one is required to find the optimal (or almost optimal) arm, there remains a multiplicative gap between the lower and upper bounds over the number of arm pulls required for the task, which is logarithmic in the problem parameters.

Such extra logarithmic terms might entail a meaningful drawback in large-scale applications, and algorithms that are more asymptotically efficient are likely to be the better choice. Indeed, over the past few years MAB algorithms have been employed in an increasing amount of web-scale applications. These include online ad selection (Chakrabarti et al., 2008), web content optimization (Agarwal et al., 2009), user-content matching (Pandey et al., 2007), learning search-ranking from usage data (Radlinski et al., 2008) and more. In many of these applications, the number of arms may be very large and the number of available arm pulls even more so. For example, an arm may represent a *specific combination* of content items or advertisements to be displayed on a web page, giving rise to an extremely large number of arms.

In this work, we shorten the theoretic gap almost entirely, leaving only doubly-logarithmic extra terms. We consider two settings of recent interest: *fixed confi-*

[†]Most of this work was done while the author was at Yahoo! Research.

Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28.

dence and *fixed budget*, in which either the error probability or the number of arm pulls are fixed and the goal is to minimize the other (see Gabillon et al. 2012). In each setting, we present a new parameter-free algorithm achieving the improved bound. Additionally, we evaluate our algorithm empirically and compare it to state of the art algorithms, in several realistic scenarios and various scales. The experiments demonstrate that our algorithm outperforms previous algorithms and scales better as the complexity of the problem grows.

1.1. Previous Work

The fixed confidence setting was first considered by Even-Dar et al. (2002), who presented the Successive Elimination strategy for the best-arm and ε-best arm identification problems. Mannor & Tsitsiklis (2004) provided tight, distribution-dependent lower bounds for several variants of the PAC model. The setting was revisited by Bubeck et al. (2009), who indicated that regret-minimizing algorithms (like UCB of Auer et al. 2002) are not well-suited for pure exploration tasks. More recently, Audibert et al. (2010) addressed the fixed budget setting and presented the algorithms UCB-E and Successive Rejects for best-arm identification, proved their optimality up to logarithmic factors, and demonstrated their effectiveness in simulations.

Recently, some extensions of the explorative MAB problem were studied by several authors. Bubeck et al. (2013) considered the problem of multiple best-arm identification and presented a variant of the Successive Rejects algorithm that with high probability outputs a set of m-best arms given a fixed budget of pulls. They also presented extensions to their setting that deal with arms having different variances. The related works of Kalyanakrishnan & Stone (2010) and Kalyanakrishnan et al. (2012) studied the same problem in a fixed confidence PAC setup. Another related setting was introduced by Gabillon et al. (2011), who consider a multi-bandit MAB problem in which each player has to identify the best arm in his sub-problem and the challenge is to distribute the limited resources at hand (namely T arm pulls) so as to maximize the confidence of all players simultaneously.

Our work is related to (Gabillon et al., 2012), that also consider both the fixed confidence and fixed budget variants of the problem. While they aim at a unified algorithmic approach for both setups, our goal is to derive better algorithms and improved analysis for the problem, treating each variant separately.

1.2. Overview of Our Approach

In this section we give a technical overview of our approach and methods, that enable us to avoid logarithmic factors in our upper bounds.

Our algorithms, in both the fixed confidence and fixed budget settings, are based on sequential elimination of arms. Specifically, the algorithms proceed in rounds, where within a round the remaining arms are sampled uniformly. At the end of each round, the arms are eliminated according to some criteria; the process continues until only a single arm remains.

The challenge we face when designing an efficient elimination algorithm is the following: each elimination of an arm should be made with high confidence, since an eliminated arm is never revived. Hence, before an arm is ruled out its reward should be estimated with high confidence. Usually, this is accomplished by ensuring that before each elimination, *all* arms are estimated with high confidence, thereby requiring a union-bound argument in the analysis and giving rise to logarithmic factors in the resulting bound.

In order to circumvent this difficulty, we first aim at reducing the number of elimination rounds. More importantly, we design our estimators within each round so as to only ensure that *most* (typically, a constant fraction) of the remaining arms are sampled with high confidence. Taking this path, we are able to avoid most union-bound arguments being used in previous works.

In the fixed confidence setting, we face an additional difficulty. There, in order to eliminate a suboptimal arm, we require an accurate estimation of its gap from the best arm having the maximal reward. Instead of establishing that by taking the maximum of the empirical rewards of the surviving arms (as done in previous works), we employ a subprocedure based on the MedianElimination algorithm (see Section 2.1), giving rise to a more efficient estimation.

2. Background and Statement of Our Results

In the Multi-Armed Bandit (MAB) problem, a player is given n arms, enumerated by $[n] := \{1, 2, \ldots, n\}$. Each arm $i \in [n]$ is associated with a reward, which is a random variable bounded in the interval $[0, 1]$ with expectation p_i. For convenience, we assume that the arms are ordered by their expected rewards, that is $p_1 \geq p_2 \geq \cdots \geq p_n$. At every time step $t = 1, 2, \ldots$, the player pulls one arm of her choice and observes an independent sample of its reward. We use the notation

$\Delta_i := p_1 - p_i$ to denote the suboptimality gap of arm i, and occasionally use $\Delta := \Delta_2$ for denoting the minimal gap. Since we assume that the best arm is unique, we have $\Delta_i > 0$ for all $i > 1$.

In the explorative setting, at the end of the game the player must commit to a single arm; the goal of the player is to choose the *best arm* (i.e. the arm having the maximal expected reward) with maximal confidence. As is standard, we shall henceforth assume for simplicity that this best arm is unique[1].

We now distinguish between two settings of interest:

- **Fixed confidence**: The player is given a target confidence δ and his goal is to pull arms as few times as possible in order to identify the best arm with probability at least $1 - \delta$.
- **Fixed budget**: Given a total budget of T arm pulls, the player's target is to maximize his probability of identifying the best arm correctly while not pulling arms more than T times.

For the sake of consistency, we measure the quality of our results in both the fixed confidence and fixed budget settings by expressing the required budget of arm pulls T as a function of the target error probability δ. Mannor & Tsitsiklis (2004) have shown that for a wide variety of reward setups and any MAB policy we have that $T = \Omega(H \log(1/\delta))$, where

$$H := \sum_{i=2}^{n} \frac{1}{\Delta_i^2} \ .$$

As demonstrated in previous works (and in our experiments), it seems that H indeed captures the complexity of the problem. However, our analysis in the fixed budget setting relies on the following related complexity measure, introduced by Audibert et al. (2010):

$$H_2 := \max_{i \neq 1} \frac{i}{\Delta_i^2} \ .$$

It is not difficult to prove that the ratio $\tilde{H} := H/H_2$ has $1 \leq \tilde{H} \leq \ln 2n$, and that both inequalities are in general tight. As a result, our upper bound in the fixed budget setting is not directly comparable to the above lower bound; see Section 6 for a discussion.

Table 1 summarizes our theoretical results along with a comparison to the previously known state-of-the-art. The bounds are stated in comparison to the lower bound $T = \Omega(H \log(1/\delta))$. That is, instead of stating the required number of pulls we give the multiplicative gap from this lower bound. The exact statements of our results are given in Theorems 3.1 and 4.1 below.

Setting	New Gap	Previous Gap
Fixed δ	$O(\log\log(1/\Delta))$	$O(\log(n/\Delta))$
Fixed T	$O(\tilde{H} \log\log n)$	$O(\tilde{H} \log n)$

Table 1. The performance of our algorithms compared to the state-of-the-art. In both cases T is considered as a function of δ and the measure presented is the multiplicative gap from the lower bound. The variable $\tilde{H}$ in the second row depends on the relative structure of the expected rewards and is bounded between 1 and $\log 2n$.

2.1. The Median Elimination Algorithm

For our algorithm in the fixed confidence setting, we require an approximate estimation procedure for the maximal expected reward of a subset of arms, which is more efficient than the empirical maximal reward. To this end we employ the (ε, δ)-PAC algorithm MedianElimination (see Even-Dar et al. 2006) described in the following lemma.

Lemma 2.1 (Even-Dar et al. 2006)**.** *Given parameters $\varepsilon, \delta > 0$ and a set S of n bandit arms,* MedianElimination(S, ε, δ) *emits an ε-optimal arm with probability at least $1 - \delta$ by using a budget of at most $O((n/\varepsilon^2)\log(1/\delta))$ pulls.*

We note that the analysis of MedianElimination is inherently worst-case, and is independent of the reward distributions. In Section 3 we show how this algorithm can be used in conjunction with adaptive elimination techniques, in order to improve the distribution-dependent upper bounds.

3. Fixed Confidence Setting

In this section we study the fixed confidence setting, where the player is given a target confidence level δ with a goal of using as few arm pulls as possible in order to find the best arm with probability no less than $1 - \delta$.

The algorithm we propose in this setting is reminiscent of sequential elimination algorithms; see Algorithm 1 for its precise description. The algorithm proceeds in rounds, where in round r it aims at eliminating $(1/2)^r$-suboptimal arms[2], i.e., arms i having $\Delta_i > (1/2)^r$.

[1] The case of multiple best-arms is more naturally captured by the PAC setup, in which we are interested in finding an ε-best arm. Some of our results can be readily extended to this setup; we defer details to the full version of the paper.

[2] A similar scheme was used by Auer & Ortner (2010) for improving the regret bound of the UCB algorithm with

Since we do not know the suboptimality of each arm, we estimate it to the required accuracy by means of the MEDIANELIMINATION algorithm (see Section 2.1). Also, as we show in the analysis below, our estimators in each round are tuned only to ensure that *most* of the suboptimal arms are eliminated with high probability, instead of requiring that all of them are ruled out in the particular round.

Algorithm 1 EXPONENTIAL-GAP ELIMINATION

input confidence $\delta > 0$

initialize $S_1 \leftarrow [n]$, $r \leftarrow 1$
while $|S_r| > 1$ **do**
  let $\varepsilon_r = 2^{-r}/4$ and $\delta_r = \delta/(50r^3)$
  sample each arm $i \in S_r$ for $t_r = (2/\varepsilon_r^2)\ln(2/\delta_r)$ times, and let $\hat{p}_i^r$ be the average reward
  invoke $i_r \leftarrow$ MEDIANELIMINATION$(S_r, \varepsilon_r/2, \delta_r)$ and let $\hat{p}_\star^r = \hat{p}_{i_r}^r$
  set $S_{r+1} \leftarrow S_r \setminus \{i \in S_r : \hat{p}_i^r < \hat{p}_\star^r - \varepsilon_r\}$
  update $r \leftarrow r + 1$
end while

output arm in S_r

Formally, we prove the following sample complexity guarantee.

Theorem 3.1. *With probability at least $1-\delta$, Algorithm 1 identifies the optimal arm using*

$$O\left(\sum_{i=2}^{n} \frac{1}{\Delta_i^2} \log\left(\frac{1}{\delta}\log\frac{1}{\Delta_i}\right)\right)$$

arm pulls.

We note that a simple modification[3] of Algorithm 1 give rise to an (ε,δ)-PAC algorithm (i.e., algorithm that finds an ε-optimal arm with probability $\geq 1-\delta$) with similar, almost-optimal guarantees.

Theorem 3.2. *There exists an algorithm that with probability at least $1-\delta$, finds an ε-optimal arm using at most*

$$O\left(\sum_{i=2}^{n} \frac{1}{(\Delta_i^\varepsilon)^2} \log\left(\frac{1}{\delta}\log\frac{1}{\Delta_i^\varepsilon}\right)\right)$$

arm pulls, where $\Delta_i^\varepsilon := \max\{\Delta_i, \varepsilon\}$.

For proving Theorem 3.1, we first establish few lemmas. First, it is easy to prove that with high probability, the best arm is not eliminated by the algorithm.

Lemma 3.3. *With probability at least $1-\delta/5$, we have $\hat{p}_1^r \geq \hat{p}_\star^r - \varepsilon_r$ for all r.*

Proof. First observe that for any arm i and round r, we have by Hoeffding's inequality

$$\Pr[|\hat{p}_i^r - p_i| \geq \varepsilon_r/2] \leq 2\exp(-\varepsilon_r^2 t_r/2) = \delta_r\,. \tag{1}$$

Now consider round r, assuming that the best arm was not eliminated previously. Since $p_{i_r} \leq p_1$, from (1) we have $\hat{p}_\star^r < p_1 + \varepsilon_r/2$ with probability at least $1-\delta_r$. However, (1) also gives that $p_1 < \hat{p}_1^r + \varepsilon_r/2$ with probability at least $1-\delta_r$, from which we conclude that $\hat{p}_1^r > \hat{p}_\star^r - \varepsilon_r$ with probability $\geq 1-2\delta_r$. The lemma now follows from a union bound, noting that $\sum_{r=1}^{\infty} 2\delta_r \leq \sum_{r=1}^{\infty} 2\delta/(50r^2) \leq \delta/5$. □

For the rest of the analysis we will need some additional notation. For all $0 \leq s \leq \lceil \log_2(1/\Delta)\rceil$, we define the set

$$A_s = \{i \in [n] \;:\; 2^{-s} \leq \Delta_i < 2^{-s+1}\} \tag{2}$$

and let $n_s = |A_s|$. Also, we denote the set of arms from A_s surviving after round r by $S_{r,s} = S_r \cap A_s$, for all $r, s \geq 0$.

Our next step is to prove that from round s onwards, a constant fraction of the surviving arms of the set A_s is eliminated on each round.

Lemma 3.4. *Assume that the optimal arm is not eliminated by the algorithm. Then with probability at least $1-4\delta/5$, we have $|S_{r,s}| \leq \frac{1}{8}|S_{r-1,s}|$ for all $1 \leq s \leq r$.*

Proof. Consider round r. The guarantees of the MEDIANELIMINATION algorithm (see Lemma 2.1) imply that $p_{i_r} \geq p_1 - \varepsilon_r/2$ with probability at least $1-\delta_r$, assuming that the best arm reached round r. Together with (1), we have

$$\Pr[\hat{p}_\star^r \leq p_1 - \varepsilon_r] \leq 2\delta_r\,. \tag{3}$$

For any arm i with $\Delta_i \geq 2^{-r} = 4\varepsilon_r$, the event $\hat{p}_i^r \geq \hat{p}_\star^r - \varepsilon_r$ implies that either $\hat{p}_i^r \geq p_i + \varepsilon_r$ or $\hat{p}_\star^r \leq p_1 - \varepsilon_r$, since otherwise

$$\hat{p}_i^r < p_i + \varepsilon_r \leq p_1 - \Delta_i + 2\varepsilon_r \leq p_1 - 2\varepsilon_r < \hat{p}_\star^r - \varepsilon_r\,.$$

Hence, (1) and (3) give

$$\begin{aligned}\Pr[\hat{p}_i^r \geq \hat{p}_\star^r - \varepsilon_r] &\leq \Pr[\hat{p}_i^r \geq p_i + \varepsilon_r] + \Pr[\hat{p}_\star^r \leq p_1 - \varepsilon_r] \\ &\leq 3\delta_r\end{aligned}$$

which means that an arm $i \in A_s$ survives round $r \geq s$ with probability at most $3\delta_r$. Consequently, we have

respect to the gaps Δ_i.

[3] Essentially, the only modification required is limiting the number of rounds performed by the algorithm to $O(\log(1/\varepsilon))$.

$\mathbf{E}[|S_{r,s}|] \leq 3\delta_r|S_{r-1,s}|$, and by applying Markov's inequality we obtain that

$$\Pr[|S_{r,s}| > \tfrac{1}{8}|S_{r-1,s}|] < \frac{3\delta_r|S_{r-1,s}|}{\frac{1}{8}|S_{r-1,s}|} = 24\delta_r\,.$$

The lemma now follows via a union bound, by which the probability of failure is bounded by $\sum_{r=1}^{\infty}\sum_{s=1}^{r} 24\delta_r = \sum_{r=1}^{\infty} 24\delta/(50r^2) < 4\delta/5$. □

We next calculate how many times an arm from A_s is pulled by the algorithm, ignoring (for now) the pulls spent due to invocations of MEDIANELIMINATION[4].

Lemma 3.5. *With probability at least $1-\delta$, the total number of times an arm from A_s is sampled in line 4 is $O\left(4^s n_s \log(s/\delta)\right)$ for all s.*

Proof. Let T_s denote the total number of times an arm from A_s is pulled. By Lemma 3.4, if the algorithm is successful we have

$$\begin{aligned} T_s &= \sum_{r=1}^{\infty}|S_{r,s}|\cdot t_r \leq \sum_{r=1}^{s-1}|A_s|\cdot t_r + \sum_{r=s}^{\infty}|S_{r,s}|\cdot t_r \\ &\leq n_s\sum_{r=1}^{s-1} t_r + n_s\sum_{r=0}^{\infty}(\tfrac{1}{8})^{r+1}t_{r+s}\,. \end{aligned}$$

Now, the first sum is upper bounded by

$$\sum_{r=1}^{s-1} t_r = 32\sum_{r=1}^{s-1} 4^r \ln\frac{100r^3}{\delta} = O\left(4^s\log\frac{s}{\delta}\right).$$

For the second sum, we have

$$\begin{aligned} \sum_{r=0}^{\infty}(\tfrac{1}{8})^{r+1}t_{r+s} &= 4^{s+1}\sum_{r=0}^{\infty}2^{-r}\ln\frac{100(r+s)^3}{\delta} \\ &\leq 3\cdot 4^{s+1}\ln\frac{4s}{\delta}\sum_{r=0}^{\infty}2^{-r} \\ &\quad + 3\cdot 4^{s+1}\sum_{r=1}^{\infty}2^{-r}\ln(5r) \\ &= O\left(4^s\log\frac{s}{\delta}\right). \end{aligned}$$

This completes the proof. □

We are now ready to prove Theorem 3.1.

Proof. We first prove that the algorithm is correct with probability $1-\delta$. Lemma 3.4 imply that any suboptimal arm is eliminated eventually, while 3.3 ensures that the best arm is never eliminated. Hence, the algorithm terminates at some point and returns the correct arm. By a union bound, this happens with probability at least $1-\delta$.

We turn to compute the sample complexity of the algorithm. Notice that the number of pulls spent by the invocation of MEDIANELIMINATION on round r is $O(|S_{r-1}|\cdot t_r)$, which is the same as the total number of pulls used by the algorithm itself (in line 4) on that round. Hence, the invocations of MEDIANELIMINATION affect the overall sample complexity only by a constant factor, and we may safely ignore the arm-pulls associated with them. Now, Lemma 3.5 asserts that, if the algorithm is successful, the number of times it pulls an arm from A_s is $T_s = O\left(4^s n_s\log(s/\delta)\right)$ for all s. Recalling that $2^s < 2/\Delta_i$ for all $i \in A_s$ (see (2)), we obtain

$$T_s = O\left(4^s n_s\log\frac{s}{\delta}\right) = O\left(\sum_{i\in A_s}\frac{1}{\Delta_i^2}\log\left(\frac{1}{\delta}\log\frac{1}{\Delta_i}\right)\right)$$

and the theorem follows by summing over s. □

[4]It is possible to re-use the same arms pulls issued by Algorithm 1 (in line 4) within invocations of ME, without any major changes in the analysis (independence of the RVs in question is not needed).

4. Fixed Budget Setting

In this section we turn to discuss the setting in which instead of targeting a given confidence δ, we are given a fixed budget of T arm pulls with the goal of maximizing the probability of correct identification.

The algorithm we propose, which we call SEQUENTIAL HALVING, is given in Algorithm 2. The strategy is simple: we split the given budget evenly across $\log_2 n$ elimination rounds, and within a round we pull arms in a uniform manner. At the end of a round, we rule out the worst half of the arms. By inspecting the empirical third quartile of the surviving arms at each round, we are able to bound the probability of this strategy to erroneously eliminate the best arm. Below we prove that this algorithm achieves the following bounds.

Theorem 4.1. *Algorithm 2 correctly identifies the best arm with probability at least*

$$1 - 3\log_2 n\cdot\exp\left(-\frac{T}{8H_2\log_2 n}\right).$$

Alternatively, for succeeding with probability at least $1-\delta$, the algorithm needs a total of at most

$$T = O\left(H_2\log n\log\left(\frac{\log n}{\delta}\right)\right)$$

arm pulls.

To avoid technicalities and ease the reading, we henceforth assume that n is a power of 2. It is easy to verify that the analysis holds for any n. We begin with

Algorithm 2 SEQUENTIAL HALVING

input total budget T

initialize $S_0 \leftarrow [n]$

for $r = 0$ to $\lceil \log_2 n \rceil - 1$ **do**

sample each arm $i \in S_r$ for

$$t_r = \left\lfloor \frac{T}{|S_r| \lceil \log_2 n \rceil} \right\rfloor$$

times, and let $\hat{p}_i^r$ be the average reward

let S_{r+1} be the set of $\lceil |S_r|/2 \rceil$ arms in S_r with the largest empirical average

end for

output arm in $S_{\lceil \log_2 n \rceil}$

the following simple lemma, which follows immediately from Hoeffding's inequality (thus the proof is omitted).

Lemma 4.2. *Assume that the best arm was not eliminated prior to round r. Then for any arm $i \in S_r$,*

$$\Pr[\hat{p}_1^r < \hat{p}_i^r] \leq \exp(-\tfrac{1}{2} t_r \Delta_i^2).$$

We next bound the probability that the algorithm errs and excludes the best arm on round r.

Lemma 4.3. *The probability that the best arm is eliminated on round r is at most*

$$3 \exp\left(-\frac{T}{8 \log_2 n} \cdot \frac{\Delta_{i_r}^2}{i_r}\right)$$

where $i_r = n/2^{r+2}$.

Proof. Define S_r' as the set of arms in S_r, excluding the $\frac{1}{4}|S_r| = n/2^{r+2}$ arms with the largest mean. If the best arm is eliminated in round r, it must be the case that at least half the arms of S_r (i.e., $\frac{1}{2}|S_r| = n/2^{r+1}$ arms) have their empirical average larger than its empirical average. In particular, the empirical means of at least $\frac{1}{3}|S_r'| = n/2^{r+2}$ of the arms in S_r' must be larger than that of the best arm at the end of round r. Letting N_r denote the number of arms in S_r' whose empirical average is larger than that of the optimal arm, we have by Lemma 4.2:

$$\begin{aligned}
\mathbf{E}[N_r] &= \sum_{i \in S_r'} \Pr[\hat{p}_1^r < \hat{p}_i^r] \leq \sum_{i \in S_r'} \exp(-\tfrac{1}{2} t_r \Delta_i^2) \\
&\leq \sum_{i \in S_r'} \exp\left(-\tfrac{1}{2}\Delta_i^2 \cdot \frac{T}{|S_r| \log_2 n}\right) \\
&\leq |S_r'| \max_{i \in S_r'} \exp\left(-\tfrac{1}{2}\Delta_i^2 \cdot \frac{2^r T}{n \log_2 n}\right) \\
&\leq |S_r'| \exp\left(-\frac{T}{8 \log_2 n} \cdot \frac{\Delta_{i_r}^2}{i_r}\right)
\end{aligned}$$

Where the last inequality follows from the fact that there are at least $i_r - 1$ arms that are not in S_r' with average reward greater than that of any arm in S_r'. We now apply Markov's inequality to obtain

$$\begin{aligned}
\Pr[N_r > \tfrac{1}{3}|S_r'|] &\leq 3\mathbf{E}[N_r]/|S_r'| \\
&\leq 3 \exp\left(-\frac{T}{8 \log_2 n} \cdot \frac{\Delta_{i_r}^2}{i_r}\right),
\end{aligned}$$

and the lemma follows. □

The proof of Theorem 4.1 is now immediate.

Proof. Clearly, the algorithm does not exceed the budget of T arm pulls. Also, if the best arm survives the execution, then the algorithm succeeds as all other arms must be eliminated after $\log_2 n$ rounds. Finally, by Lemma 4.3 and a union bound, the best arm is eliminated in one of the $\log_2 n$ rounds of the algorithm with probability at most

$$\begin{aligned}
&3 \sum_{r=1}^{\log_2 n} \exp\left(-\frac{T}{8 \log_2 n} \cdot \frac{\Delta_{i_r}^2}{i_r}\right) \\
&\quad \leq 3 \log_2 n \cdot \exp\left(-\frac{T}{8 \log_2 n} \cdot \frac{1}{\max_i \; i \Delta_i^{-2}}\right) \\
&\quad \leq 3 \log_2 n \cdot \exp\left(-\frac{T}{8 H_2 \log_2 n}\right),
\end{aligned}$$

which gives the theorem. □

5. Experiments

In this section we present a few simple experimental setups to illustrate our theoretical results. We focus on algorithms with a fixed budget setting, as those are more practical and easier to work with. Our baselines include the state-of-the-art SUCCESSIVE REJECTS, UCB-E and ADAPTIVE UCB-E (AUCB-E) algorithms of Audibert et al. (2010). It is noted that the UCB-E algorithm requires the knowledge of a parameter depending on the underlying rewards in advance; its adaptive (heuristic) counterpart AUCB-E calculates this parameter on-the-fly.

We considered six different setups, where in each the reward distributions are Bernoulli and the best arm has expected reward equal to 0.5. Each setup extends to any number of arms; we ran the experiments with $n = 20,\ 40,\ 80$ arms in order to examine how the performance of each algorithm scales as the number of arms grow. The exact setups we examined are as follows:

1. **One group of suboptimal arms**: $p_i = 0.45$ for $i \geq 2$.

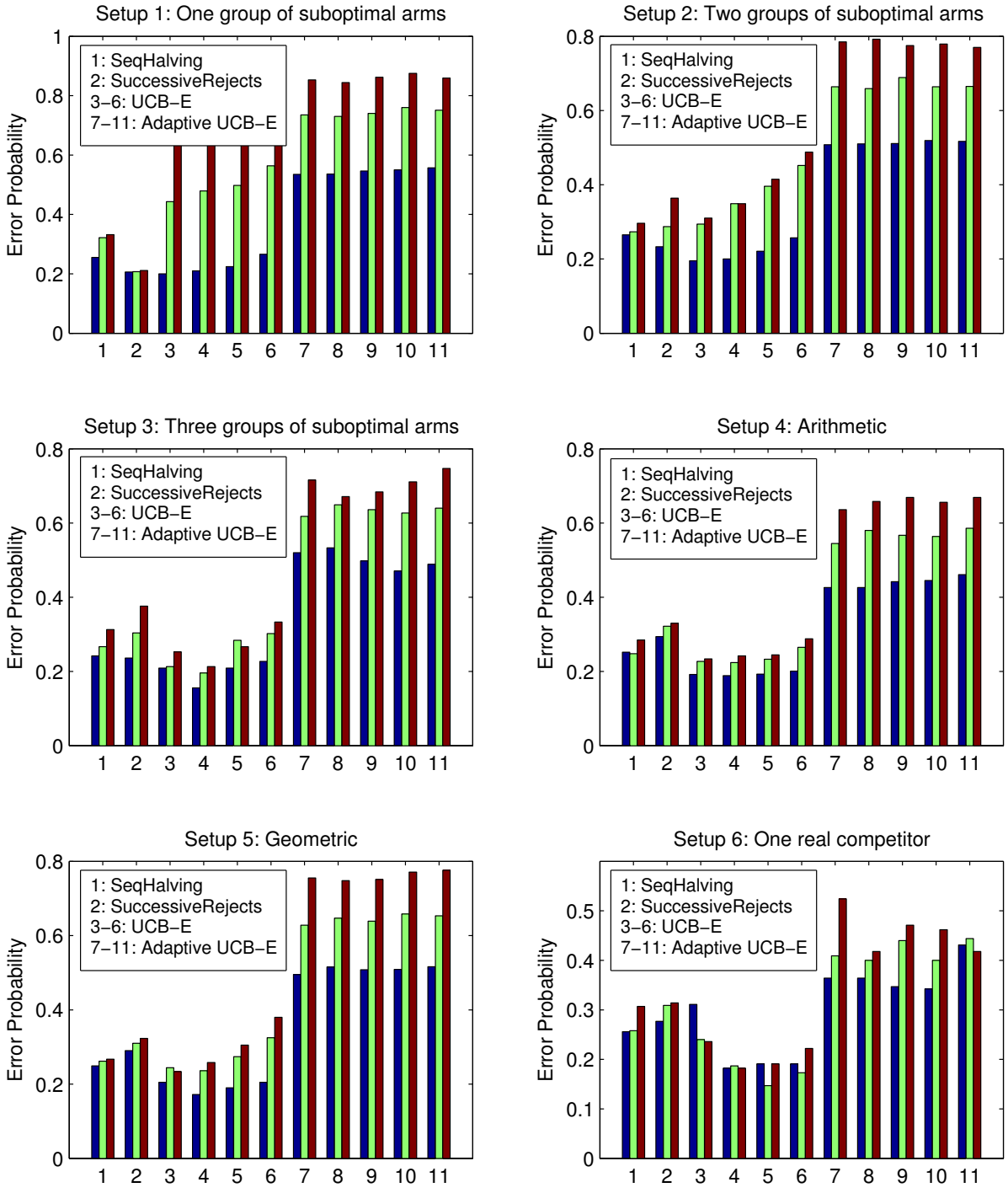

Figure 1. The error probability of the different algorithms in six different setups, averaged over 3000 independent executions (results in standard deviations of less than 1%). For each setup, we repeated the experiments with 20, 40 and 80 arms (left, middle, and right sub-columns respectively).

2. **Two groups of suboptimal arms**: $p_i = 0.5 - \frac{1}{2n}$ for $i = 2, \ldots, \lceil\sqrt{n}\rceil + 1$ and $p_i = 0.45$ otherwise.
3. **Three groups of suboptimal arms**: $p_i = 0.5 - \frac{1}{5n}$ for $i = 2, \ldots, 6$, $p_i = 0.49$ for $i = 7, \ldots, 6 + 2\lceil\sqrt{n}\rceil$ and $p_i = 0.35$ otherwise.
4. **Arithmetic**: The suboptimality of the arms form an arithmetic series where $p_2 = 0.5 - \frac{1}{5n}$ and $p_n = 0.25$.
5. **Geometric**: The suboptimality of the arms form a geometric series where $p_2 = 0.5 - \frac{1}{5n}$ and $p_n = 0.25$.
6. **One real competitor**: $p_2 = 0.5 - \frac{1}{10n}$ and $p_i = 0.4$ otherwise.

For each setup we consider eleven executions:

- Execution 1: Sequential Halving algorithm
- Execution 2: Successive Reject algorithm
- Executions 3-6: UCB-E algorithm with parameters 1, 2, 4, and 8, respectively
- Executions 7-11: AUCB-E algorithm with parameters 0.25, 0.5, 1, 2, and 4, respectively

Figure 1 presents the error probability of the different algorithms in each setup, for $n = 20$, 40 arms and $n = 80$ arms respectively. Following Audibert et al. (2010), each algorithm was given a total budget of H arm pulls which corresponds to the complexity of the underlying problem. The exact budget in each of the setups is given in Table 2.

Setup	20 arms	40 arms	80 arms
1	7600	15600	31600
2	13599	57599	258400
3	150177	340888	982488
4	14005	57430	232579
5	33220	214888	1436445
6	41799	163799	647800

Table 2. The value of H in each setup we experimented.

The experiments demonstrate a few interesting insights. In all cases, it is clear that as the number of arms grow, the algorithms perform more poorly when given the respective budget of H pulls. This suggests that the budget required for the algorithms to achieve a fixed confidence is indeed asymptotically (strictly) larger than H. The most rapid decrease in the performance as the number of arms increases occurs in UCB-E and AUCB-E. This may be explained by the fact that the theoretical bounds of these algorithms exhibit an additional logarithmic factor in the budget T (as opposed to the number of arms n as in the case of Sequential Halving and Successive Reject).

Notice that while UCB-E performs best in some settings, yet is not a practical algorithm as it requires knowledge of H. Its practical counterpart AUCB-E does not perform as well as our algorithm nor Successive Reject. The opposite occurred in the experiments of (Audibert et al., 2010), most likely due to the difference in scale. Our Sequential Halving algorithm has similar performance to that of Successive Reject, yet presents better performance overall and scales better with the number of arms.

6. Summary and Discussion

We have considered the best-arm identification problem in Multi-Armed Bandits, in two settings of recent interest: fixed confidence and fixed budget. We have investigated the tightness of the upper bounds over the number of arm pulls needed to reach a target confidence, in each of these settings. Our main contribution is in proposing two new algorithms, Exponential-Gap Elimination that is optimal up to a doubly-logarithmic factor in the problem parameters (namely, $\Delta_2, \ldots, \Delta_n$ and n), and Sequential Halving, that under a wide family of settings is optimal up to a doubly-logarithmic factors. By that, we improve upon previous works that are tight only to within logarithmic factors.

In addition, we report experimental results that support our theoretical findings and suggest that the logarithmic factors arising in the upper bounds of algorithms for the problem are evident in practice and are not just artifacts of the analyses. In particular, we demonstrate the disadvantage of a logarithmic factor in the budget. Indeed, the theoretical improvement in the analysis of our Sequential Halving algorithm is visible in the experiments we have conducted.

In the fixed confidence setting, we have closed the gap between the upper and lower bounds up to a doubly-logarithmic factor in the parameters Δ_i. We do not believe that this gap can be removed algorithmically; in fact, even in a very simple MAB problem with merely two arms, it is not clear how this can be accomplished without prior knowledge of the underlying parameters. In the fixed budget setting our analysis relies on the surrogate H_2 of the problem complexity, and thus our upper bound does not relate directly to the lower bound $\Omega(H \log(1/\delta))$. However, it is important to note that any algorithm whose upper bound is expressed as a function of H_2 cannot perform better than $O(H_2 \log n \log(1/\delta))$, and probably a quite different approach is needed in order to lower the gap any further.

References

Agarwal, D., Chen, B., and Elango, P. Explore/exploit schemes for web content optimization. In *Proc. Ninth IEEE International Conference on Data Mining (ICDM'2009)*, pp. 1–10, 2009.

Audibert, J.Y., Bubeck, S., and Munos, R. Best arm identification in multi-armed bandits. In *COLT*, pp. 41–53, 2010.

Auer, P. and Ortner, R. UCB revisited: Improved regret bounds for the stochastic multi-armed bandit problem. *Periodica Mathematica Hungarica*, 61(1-2):55–65, 2010.

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

Bubeck, S., Munos, R., and Stoltz, G. Pure exploration in multi-armed bandits problems. In *Algorithmic Learning Theory*, pp. 23–37. Springer, 2009.

Bubeck, S., Wang, T., and Viswanathan, N. Multiple identifications in multi-armed bandits. In *Proceedings of the 30th International Conference on Machine Learning*, 2013.

Chakrabarti, D., Kumar, R., Radlinski, F., and Upfal, E. Mortal multi-armed bandits. In *Proceedings of the 22nd Annual Conference on Neural Information Processing Systems (NIPS'2008)*, pp. 273–280, 2008.

Even-Dar, E., Mannor, S., and Mansour, Y. PAC bounds for multi-armed bandit and markov decision processes. In *COLT*, pp. 193–209. Springer, 2002.

Even-Dar, E., Mannor, S., and Mansour, Y. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *The Journal of Machine Learning Research*, 7:1079–1105, 2006.

Gabillon, V., Ghavamzadeh, M., Lazaric, A., and Bubeck, S. Multi-bandit best arm identification. In *Advances in Neural Information Processing Systems 24*, pp. 2222–2230. 2011.

Gabillon, V., Ghavamzadeh, M., and Lazaric, A. Best arm identification: A unified approach to fixed budget and fixed confidence. In *Advances in Neural Information Processing Systems 25*, pp. 3221–3229, 2012.

Kalyanakrishnan, S. and Stone, P. Efficient selection of multiple bandit arms: Theory and practice. In *Proceedings of the 27th International Conference on Machine Learning (ICML 2010)*, pp. 511–518, 2010.

Kalyanakrishnan, S., Tewari, A., Auer, P., and Stone, P. PAC subset selection in stochastic multi-armed bandits. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012*, 2012.

Lai, Tze Leung and Robbins, Herbert. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

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

Pandey, S., Agarwal, D., Chakrabarti, D., and Josifovski, V. Bandits for taxonomies: A model based approach. In *In Proceedings of the SIAM International Conference on Data Mining*. SDM, 2007.

Radlinski, F., Kleinberg, R., and Joachims, T. Learning diverse rankings with multi-armed bandits. In *Proceedings of the 25th International Conference on Machine Learning*, pp. 784–791. ACM, 2008.