On Sequential Elimination Algorithms for Best-Arm Identification in Multi-Armed Bandits

Shahin Shahrampour, Mohammad Noshad, and Vahid Tarokh

Abstract—We consider the best-arm identification problem in multi-armed bandits, which focuses purely on exploration. A player is given a fixed budget to explore a finite set of arms, and the rewards of each arm are drawn independently from a fixed, unknown distribution. The player aims to identify the arm with the largest expected reward. We propose a general framework to unify sequential elimination algorithms, where the arms are dismissed iteratively until a unique arm is left. Our analysis reveals a novel performance measure expressed in terms of the sampling mechanism and number of eliminated arms at each round. Based on this result, we develop an algorithm that divides the budget according to a nonlinear function of remaining arms at each round. We provide theoretical guarantees for the algorithm, characterizing the suitable nonlinearity for different problem environments described by the number of competitive arms. Matching the theoretical results, our experiments show that the nonlinear algorithm outperforms the state-of-the-art. We finally study the side-observation model, where pulling an arm reveals the rewards of its related arms, and we establish improved theoretical guarantees in the pure-exploration setting.

Index Terms—Multi-armed bandits, best-arm identification, sequential decision-making, pure-exploration, online learning.

I. INTRODUCTION

MULTI-ARMED Bandits (MAB) is a sequential decision-making framework addressing the exploration-exploitation dilemma [1]–[4]. A player explores a finite set of arms sequentially, and pulling each of them results in a *reward* to the player. The problem has been studied for different reward models. In the stochastic MAB, the rewards of each arm are assumed to be i.i.d. samples of an *unknown*, fixed distribution. Then, the player's goal is to exploit the arm with the largest expected reward as many times as possible to maximize the gain. In the literature, this objective has been translated to minimizing the *cumulative* regret, a comparison measure between the actual performance of the player versus

a clairvoyant knowing the best arm in advance. Early studies on MAB dates back to a few decades ago, but the problem has surged a lot of renewed interest due to its modern applications. Though generally addressing sequential decision-making problems, MAB has been studied in several engineering contexts such as web search and advertising, wireless cognitive radios, multi-channel communication systems, and Big-Data streaming (see e.g. [5]–[11] and references therein).

Departing from its classical setting (exploration-exploitation), many researchers have studied MAB in the pure-exploration framework. In this case, the player aims to minimize the *simple* regret which can be related to finding the best arm with a high probability [12]. As a result, the best-arm identification problem has received a considerable attention in the literature of machine learning [12]–[19]. The problem has been viewed from two main standpoints: **(i)** the *fixed-confidence* setting, where the objective is to minimize the number of trials to find the best arm with a certain confidence, and **(ii)** the *fixed-budget* setting, where the player attempts to maximize the probability of correct identification given a fixed number of arm pulls. Best-arm identification has several applications including channel allocation as originally proposed in [12]. Consider the problem of channel allocation for mobile phone communication. Before the outset of communication, a cellphone (player) can explore the set of channels (arms) to find the best one to operate. Each channel feedback is noisy, and the number of trials (budget) is limited. The problem is hence an instance of best-arm identification, and minimizing the cumulative regret is not the right approach to the problem.

While both pure-exploration and exploration-exploitation setups are concerned with finding the best arm, they are quite different problems in nature. In fact, Bubeck *et al.* [17] establish that methods designed to minimize the cumulative regret (exploration-exploitation) can perform poorly for the simple-regret minimization (pure-exploration). More specifically, they proved that upper bounds on the cumulative regret yield lower bounds on the simple regret, i.e., the smaller the cumulative regret, the larger the simple regret. Therefore, one must adopt different strategies for optimal best-arm recommendation.

A. Our Contribution

In this paper, we address the best-arm identification problem in the fixed-budget setting. We restrict our attention to a class of algorithms that work based on sequential elimination. Recently, it is proved in [20] that some existing strategies based on the sequential elimination of the arms are *optimal*. However,

the notion of optimality is defined with respect to the *worst-case* allocation of the reward distributions. The main focus of this paper is not the worst-case scenario. On the contrary, given certain regimes for rewards, our goal is to propose an algorithm outperforming the state-of-the-art in these regimes. We characterize these reward regimes in terms of the number of competitive arms, and we prove the superiority of our algorithm both theoretically and empirically.

Of particular relevance to the current study is the works of [12], [18], where two iterative algorithms are proposed for sequential elimination: Successive Rejects (Succ-Rej) [12] and Sequential Halving (Seq-Halv) [18]. In both algorithms, the player must sample the arms in rounds to discard the arms sequentially, until a single arm is left, one that is perceived as the best arm. However, we recognize two distinctions between the algorithms: **(i)** Succ-Rej eliminates one arm at each round, until it is left with a single arm, whereas Seq-Halv discards roughly half of the remaining arms at each round to identify the best arm. **(ii)** At each round, Seq-Halv samples the remaining arms uniformly (excluding previous rounds), whereas Succ-Rej samples them uniformly once including the previous rounds.

Inspired by these works, our first contribution is to propose a general framework to bring sequential elimination algorithms (including Succ-Rej and Seq-Halv) under the same umbrella. Our analysis reveals a novel performance bound which relies on the sampling design as well as the number of eliminated arms at each around. Following this general framework, we extend Succ-Rej to an algorithm that divides the budget by a nonlinear function of remaining arms at each round, unlike Succ-Rej that does so in a linear fashion. We prove theoretically that we can gain advantage from the nonlinearity. In particular, we consider several well-studied reward regimes and exhibit the suitable nonlinearity for each environment. Tuning with a proper nonlinearity, our algorithm outperforms Succ-Rej and Seq-Halv in these regimes. Interestingly, our numerical experiments support our theoretical results, while showing that our algorithm is competitive with UCB-E [12] which requires prior knowledge of a problem-dependent parameter.

Finally, we consider sequential elimination in the presence of side observations. In this model, a graph encodes the connection between the arms, and pulling one arm reveals the rewards of all neighboring arms [21]. While the impact of side observations is well-known for exploration-exploitation [22], [23], we consider the model in the pure-exploration setting. Given a partition of arms to a few blocks, we propose an algorithm that eliminates blocks consecutively and selects the best arm from the final block. Naturally, we provide an improved theoretical guarantee comparing to the full bandit setting where there is no side observation.

B. Related Work

Pure-exploration in the PAC-learning setup was studied in [13], where Successive Elimination for finding an ϵ-optimal arm with probability $1 - \delta$ (fixed-confidence setting) was proposed. Seminal works of [14], [15] provide matching lower bounds

for the problem, which present a sufficient number of arm pulls to reach the confidence $1 - \delta$. Many algorithms for pure-exploration are inspired by the classical UCB1 for exploration-exploitation [2]. For instance, Audibert *et al.* [12] propose UCB-E, which modifies UCB1 for pure-exploration. UCB-E needs prior knowledge of a problem-dependent parameter, so the authors also propose its adaptive counterpart AUCB-E to address the issue. In addition, Jamieson *et al.* [24] propose an optimal algorithm for the fixed confidence setting, inspired by the law of the iterated logarithm. We refer the reader to [25] and the references therein for recent advances in the fixed-confidence setting, while remarking that Gabillon *et al.* [19] present a unifying approach for fixed-budget and fixed-confidence settings. As another interesting direction, various works in the literature introduce information-theoretic measures for best-arm identification. Kaufmann *et al.* [26] study the identification of multiple top arms using KL-divergence-based confidence intervals. The authors of [27] investigate both settings to show that the complexity of the fixed-budget setting may be smaller than that of the fixed-confidence setting. Recently, Russo [28] develops three Bayesian algorithms to examine asymptotic complexity measure for the fixed-confidence setting. There also exists extensive literature on identification of multiple top arms in MAB (see e.g. [26], [27], [29]–[32]). Finally, we remark that simple-regret minimization has been successfully used in the context of Monte-Carlo Tree Search [33], [34] as well.

C. Organization

The rest of the paper is organized as follows. Section II is dedicated to nomenclature, problem formulation, and a summary of our results. In Sections III and IV, we discuss our main theoretical results and their consequences, while we extend these results to side-observation model in Section V. In Section VI, we describe our numerical experiments, and the concluding remarks are provided in Section VII. We include the proofs in the Appendix (Section VIII).

II. PRELIMINARIES

Notation: For integer K, we define $[K] := \{1, \ldots, K\}$ to represent the set of positive integers smaller than or equal to K. We use $|S|$ to denote the cardinality of the set S. Throughout, the random variables are denoted in bold letters.

A. Problem Statement

Consider the stochastic Multi-armed Bandit (MAB) problem, where a player explores a finite set of K arms. When the player samples an arm, only the corresponding *payoff* or *reward* of that arm is observed. The reward sequence for each arm $i \in [K]$ corresponds to i.i.d samples of an *unknown* distribution whose expected value is μ_i. We assume that the distribution is supported on the unit interval $[0, 1]$, and the rewards are generated independently across the arms. Without loss of generality, we further assume that the expected value of the rewards are ordered as

$$\mu_1 > \mu_2 \geq \cdots \geq \mu_K, \tag{1}$$

and therefore, arm 1 is the unique best arm. We let $\Delta_i := \mu_1 - \mu_i$ denote the *gap* between arm i and arm 1, measuring the sub-optimality of arm i. We also represent by $\bar{\mathbf{x}}_{i,n}$ the average reward obtained from pulling arm i for n times.

In this paper, we address the best-arm identification setup, a pure-exploration problem in which the player aims to find the arm with the largest expected value with a high confidence. There are two well-known scenarios for which the problem has been studied: fixed confidence and fixed budget. In the fixed-confidence setting, the objective is to minimize the number of trials needed to achieve a fixed confidence. However, in this work, we restrict our attention to the fixed-budget, formally described as follows:

Problem 1: Given a total budget of T arm pulls, minimize the probability of misidentifying the best arm.

It is well-known that classical MAB techniques in the exploration-exploitation setting, such as `UCB1`, are not efficient for the identification of the best arm. In fact, Bubeck *et al.* have proved in [17] that upper bounds on the cumulative regret yield lower bounds on the simple regret, i.e., the smaller the cumulative regret, the larger the simple regret. In particular, for all Bernoulli distributions on the rewards, a constant $L > 0$, and a function $f(\cdot)$, they have proved if an algorithm satisfies

$$\text{Expected Cumulative Regret} \leq Lf(T),$$

after T rounds, we have

$$\text{Misidentification Probability} \geq$$

$$\text{Expected Simple Regret} \geq D_1 \exp(-D_2 f(T)),$$

for two positive constants D_1 and D_2. Given that for optimal algorithms in the exploration-exploitation setting, we have $f(T) = \mathcal{O}(\log T)$, these algorithms decay polynomially fast for best-arm identification. However, a carefully designed best-arm identification algorithm achieves an exponentially fast decay rate (see e.g. [12], [18]).

The underlying intuition is that in the exploration-exploitation setting, only playing the best arm matters. For instance, playing the second best arm for a long time can result in a dramatically large cumulative regret. Therefore, the player needs to minimize the exploration time to focus only on the best arm. On the contrary, in the best-arm identification setting, player must recommend the best arm at the end of the game. Hence, exploring the suboptimal arms "strategically" during the game helps the player to make a better recommendation. In other words, the performance is measured by the final recommendation regardless of the time spent on the suboptimal arms. We focus on sequential elimination algorithms for the best-arm identification in the next section.

B. Previous Performance Guarantees and Our Result

In this work, we examine sequential-elimination type algorithms in the fixed budget setting. We propose a general algorithm that unifies sequential elimination methods, including celebrated `Succ-Rej` [12] and `Seq-Halv` [18]. We then use a special case of this general framework to develop an algorithm called Nonlinear Sequential Elimination. We show that this al-

gorithm is more efficient than `Succ-Rej` and `Seq-Halv` in several problem scenarios.

Any sequential elimination algorithm samples the arms based on some strategy. It then discards a few arms at each round and stops when it is only left by one arm. In order to integrate sequential elimination algorithms proceeding in R rounds, we use the following key observation: let any such algorithm play each (remaining) arm for n_r times (in total) by the end of round $r \in [R]$. If the algorithm needs to discard b_r arms at round r, it must satisfy the following budget constraint

$$b_1 n_1 + b_2 n_2 + \cdots + (b_R + 1)n_R \leq T, \qquad (2)$$

since the b_r arms eliminated at round r have been played n_r times, and the surviving arm has been played n_R times. Alternatively, letting $g_r := \sum_{i=r}^{R} b_i + 1$ denote the number of remaining arms at the start of round r, one can pose the budget constraint as

$$g_1 n_1 + g_2(n_2 - n_1) + \cdots + g_R(n_R - n_{R-1}) \leq T. \qquad (3)$$

Our first contribution is to derive a generic performance measure for such algorithm (Theorem 2), relating the algorithm efficiency to $\{b_r\}_{r=1}^R$ as well as the sampling scheme determining $\{n_r\}_{r=1}^R$. Note that `Succ-Rej` satisfies $b_r = 1$ for $R = K - 1$ rounds, whereas `Seq-Halv` is characterized via $g_{r+1} = \lceil g_r/2 \rceil$ with $g_1 = K$ for $R = \lceil \log_2 K \rceil$. While it is shown in [14] that in many settings for MAB, the quantity H_1 in the following plays a key role,

$$H_1 := \sum_{i=2}^{K} \frac{1}{\Delta_i^2} \qquad \text{and} \qquad H_2 := \max_{i \neq 1} \frac{i}{\Delta_i^2}, \qquad (4)$$

the performance of both `Succ-Rej` and `Seq-Halv` is described via the complexity measure H_2, which is equal to H_1 up to logarithmic factors in K [12]. In particular, for each algorithm the bound on the probability of misidentification can be written in the form of $\beta \exp(-T/\alpha)$, where α and β are given in Table I in which $\overline{\log} K = 0.5 + \sum_{i=2}^{K} i^{-1}$.

In `Succ-Rej`, at round r, the $K - r + 1$ remaining arms are played proportional to the whole budget divided by $K - r + 1$ which is a linear function of r. Motivated by the fact that this linear function is not necessarily the best sampling rule, as our second contribution, we specialize our general framework to an algorithm which can be cast as a nonlinear extension of `Succ-Rej`. This algorithm is called Nonlinear Sequential Elimination (`N-Seq-El`), where the term "nonlinear" refers to the fact that at round r, the algorithm divides the budget by the nonlinear function $(K - r + 1)^p$ for a positive real $p > 0$ (an input value). We prove (Proposition 3) that the performance of our algorithm depends on the following quantities

$$H(p) := \max_{i \neq 1} \frac{i^p}{\Delta_i^2} \qquad \text{and} \qquad C_p := 2^{-p} + \sum_{r=2}^{K} r^{-p}, \qquad (5)$$

as described in Table I. We do not use p as a subscript in the definition of $H(p)$ to avoid confusion over the fact that $H(1) = H_2$ due to definition (4). Indeed, `N-Seq-El` with $p = 1$ recovers `Succ-Rej`, but we show that in many regimes for arm gaps, $p \neq 1$ provides better theoretical results (Corollary 4). We

TABLE I
THE PARAMETERS α AND β FOR THE ALGORITHMS DISCUSSED IN THIS SECTION, WHERE THE MISIDENTIFICATION PROBABILITY FOR EACH OF THEM DECAYS IN THE FORM OF $\beta \exp(-T/\alpha)$. THE RELEVANT COMPLEXITY MEASURES USED IN THIS TABLE ARE DEFINED IN (4) AND (5)

	Succ-Rej	Seq-Halv	Our algorithm
α	$H_2 \overline{\log} K$	$8 H_2 \log_2 K$	$H(p) C_p$
β	$0.5 K(K-1) \exp\left(K/(H_2 \overline{\log} K)\right)$	$3 \log_2 K$	$(K-1) \exp\left(K/H(p)C_p\right)$

also illustrate this improvement in the numerical experiments in Section VI, where we observe that $p \neq 1$ can outperform Succ-Rej and Seq-Halv in many settings considered in [12], [18]. Since the value of p is received as an input, we remark that our algorithm needs tuning to perform well; however, the tuning is more qualitative rather than quantitative, i.e., the algorithm maintains a reasonable performance as long as p is in a certain interval, and therefore, the value of p needs not be specific. We will discuss this in the next section in more details.

III. NONLINEAR SEQUENTIAL ELIMINATION

In this section, we propose our generic sequential elimination method and analyze its performance in the fixed budget setting. Consider a sequential elimination algorithm given budget T of arm pulls. The algorithm maintains an active set initialized by the K arms, and it proceeds for R rounds to discard the arms sequentially until it is left by a single arm. Let us use $\lceil \cdot \rceil$ to denote the ceiling function. Then, for a constant C and a decreasing, positive sequence $\{z_r\}_{r=1}^R$, set $n_r = \lceil (T-K)/(Cz_r) \rceil$ at round $r \in [R]$, let the algorithm sample the remaining arms for $n_r - n_{r-1}$ times, and calculate the empirical average of rewards for each arm. If the algorithm dismisses b_r arms with lowest average rewards, and we impose the constraint $\sum_{r=1}^R b_r = K - 1$ on the sequence $\{b_r\}_{r=1}^R$, the algorithm outputs a single arm, the one that it hopes to be the best arm. This general point of view is summarized in Fig. 1 on the left side.

The choice of C in the algorithm (see Fig. 1) must warrant that the budget constraint (3) holds. When we substitute n_r into (3), we get

$$
\begin{aligned}
n_R + \sum_{r=1}^R b_r n_r &= \left\lceil \frac{T-K}{Cz_R} \right\rceil + \sum_{r=1}^R b_r \left\lceil \frac{T-K}{Cz_r} \right\rceil \\
&\leq \frac{T-K}{Cz_R} + 1 + \sum_{r=1}^R b_r + \sum_{r=1}^R b_r \frac{T-K}{Cz_r} \\
&= K + \frac{T-K}{C}\left(\frac{1}{z_R} + \sum_{r=1}^R \frac{b_r}{z_r} \right) = T,
\end{aligned}
$$

where in the last line we used the condition $\sum_{r=1}^R b_r = K - 1$. The following theorem provides an upper bound on the error probability of the algorithm.

Theorem 2: Consider the General Sequential Elimination algorithm outlined in Fig. 1. For any $r \in [R]$, let b_r denote the number of arms that the algorithm eliminates at round r. Let also $g_r := |\mathbf{G}_r| = \sum_{i=r}^R b_i + 1$ be the number of remaining arms at the start of round r. Given a fixed budget T of arm pulls and the

input sequence $\{z_r\}_{r=1}^R$, setting C and $\{n_r\}_{r=1}^R$ as described in the algorithm, the misidentification probability satisfies the bound,

$$
\mathbb{P}\left(\mathbf{G}_{R+1} \neq \{1\}\right) \leq
$$

$$
R \max_{r \in [R]} \{b_r\} \exp\left(-\frac{T-K}{C} \min_{r \in [R]} \left\{ \frac{2\Delta_{g_{r+1}+1}^2}{z_r} \right\} \right).
$$

It is already quite well-known that the sub-optimality Δ_i of each arm i plays a key role in the identification quality; however, an important subsequent of Theorem 2 is that the performance of any sequential elimination algorithm also relies on the choice of z_r which governs the constant C. In [12], Succ-Rej employs $z_r = K - r + 1$, i.e., at each round the remaining arms are played equally often in total. This results in C being of order $\log K$.

We now use the abstract form of the generic algorithm to specialize it to Nonlinear Sequential Elimination delineated in Fig. 1. The algorithm works with $b_r = 1$ and $z_r = (K-r+1)^p$ for a given $p > 0$, and it is called "nonlinear" since p is not necessarily equal to one. The choice of $p = 1$ reduces the algorithm to Succ-Rej. In Section IV, we prove that in many regimes for arm gaps, $p \neq 1$ provides better theoretical results, and we further exhibit the efficiency in the numerical experiments in Section VI. The following proposition encapsulates the theoretical guarantee of the algorithm.

Proposition 3: Let the Nonlinear Sequential Elimination algorithm in Fig. 1 run for a given $p > 0$, and let C_p and $H(p)$ be as defined in (5). Then, the misidentification probability satisfies the bound,

$$
\mathbb{P}\left(\mathbf{G}_K \neq \{1\}\right) \leq (K-1) \exp\left(-2\frac{T-K}{C_p H(p)} \right).
$$

The performance of the algorithm does depend on the parameter p, but the choice is more qualitative rather than quantitative. For instance, if the sub-optimal arms are almost the same, i.e., $\Delta_i \approx \Delta$ for $i \in [K]$, noting the definition of C_p and $H(p)$ in (5), we observe that $0 < p < 1$ performs better than $p > 1$. In general, larger values for p increase $H(p)$ and decrease $C(p)$. Therefore, there is a trade-off in choosing p. We elaborate on this issue in Sections IV and VI, where we observe that a wide range of values for p can be used for tuning, and the trade-off can be addressed using either $0 < p < 1$ or $1 < p \leq 2$.

We remark that in the proof of Theorem 2, the constant behind the exponent can be improved if one can avoid the union bounds. To do so, one needs to assume some structure in the sequence $\{g_r\}_{r=1}^R$. For instance, the authors in [18] leverage the fact that

General Sequential Elimination Algorithm	**Nonlinear Sequential Elimination Algorithm**
Input: budget T, sequence $\{z_r, b_r\}_{r=1}^{R}$. **Initialize:** $\mathbf{G}_1 = [K]$, $n_0 = 0$. Let $$C = \frac{1}{z_R} + \sum_{r=1}^{R} \frac{b_r}{z_r}$$ $$n_r = \left\lceil \frac{T-K}{Cz_r} \right\rceil \text{ for } r \in [R]$$ At round $r = 1, \dots, R$: (1) Sample each arm in $\mathbf{G}_r$ for $n_r - n_{r-1}$ times. (2) Let $\mathbf{B}_r$ be the set of b_r arms with smallest average rewards. (3) Let $\mathbf{G}_{r+1} = \mathbf{G}_r \setminus \mathbf{B}_r$, i.e., discard the set worst b_r arms. **Output:** $\mathbf{G}_{R+1}$.	**Input:** budget T, parameter $p > 0$. **Initialize:** $\mathbf{G}_1 = [K]$, $n_0 = 0$. Let $$C_p = 2^{-p} + \sum_{r=2}^{K} r^{-p}$$ $$n_r = \left\lceil \frac{T-K}{C_p(K-r+1)^p} \right\rceil \text{ for } r \in [K-1]$$ At round $r = 1, \dots, K-1$: (1) Sample each arm in $\mathbf{G}_r$ for $n_r - n_{r-1}$ times. (2) Let $\mathbf{B}_r$ be the set containing the arm with smallest average reward. (3) Let $\mathbf{G}_{r+1} = \mathbf{G}_r \setminus \mathbf{B}_r$, i.e., discard the worst single arm. **Output:** $\mathbf{G}_K$.

Fig. 1. The algorithm on the left represents a general recipe for sequential elimination, whereas the one on the right is a special case of the left hand side, which extends `Succ-Rej` to nonlinear budget allocation.

$g_{r+1} = \lceil g_r/2 \rceil$ to avoid union bounds in `Seq-Halv`. The idea can be extended to when the ratio of g_{r+1}/g_r is a constant independent of r.

Finally, while the result of Proposition 3 provides a general performance bound with respect to reward regimes, in the next section, we provide two important corollaries of the proposition to study several regimes for rewards, where we can simplify the bound and compare our result with other algorithms.

IV. PERFORMANCE IN SEVERAL SUB-OPTIMALITY REGIMES

Inspired by numerical experiments carried out in the previous works [12], [18], we consider a few instances for sub-optimality of arms in this section, and we demonstrate how `N-Seq-El` fares in these cases. We would like to distinguish three general regimes that encompass interesting settings for sub-optimality and determine the values of p for which we expect `N-Seq-El` to achieve faster identification rates:

1) *Arithmetic progression of gaps:* In this case, $\Delta_i = (i-1)\Delta_0$ for $i > 1$, where Δ_0 is a constant.
2) *A large group of competitive, suboptimal arms:* We recognize this case as when there exists a constant[1] $\varepsilon \geq 0$ such that $\Delta_i/\Delta_2 \leq 1+\varepsilon$ for arms $i \in S$, where $|S|$ grows linearly as a function of K, and for $i \notin S$, $\Delta_i/\Delta_2 \geq i$.
3) *A small group of competitive, suboptimal arms:* This case occurs when there exists a constant $\varepsilon \geq 0$ such that $\Delta_i/\Delta_2 \leq 1+\varepsilon$ for $i \in S$, where $|S|$ is of constant order with respect to K, and for $i \notin S$, $\Delta_i/\Delta_2 \geq i$.

We now state the following corollary of Proposition 3, which proves to be useful for our numerical evaluations. Note that the orders are expressed with respect to K.

---

[1]The choice of ε must be constant with respect to i and K.

Corollary 4: Consider the Nonlinear Sequential Elimination algorithm described in Fig. 1. Let constants p and q be chosen such that $1 < p \leq 2$ and $0 < q < 1$. Then, for the three settings given above, the bound on the misidentification probability presented in Proposition 3 satisfies

Regime 1	Regime 2	Regime 3
$H(p)C_p = \mathcal{O}(1)$	$H(q)C_q = \mathcal{O}(K)$	$H(p)C_p = \mathcal{O}(1)$

We can now compare our algorithm with `Succ-Rej` and `Seq-Halv` using the result of Corollary 4. Returning to Table I and calculating H_2 for Regimes 1 to 3, we can observe the following table, which shows that with a right choice of p for `N-Seq-El`, we can save a $\mathcal{O}(\log K)$ factor in the exponential rate comparing to other methods. Though we do not have a prior information on gaps to categorize them into one of the Regimes 1 to 3 (and then choose p), the choice of p is more qualitative rather than quantitative. Roughly speaking: if the sub-optimal arms are almost the same $0 < p < 1$ performs better than $p > 1$, and if there are a few real competitive arms, $p > 1$ outperforms $0 < p < 1$. Therefore, the result of Corollary 4 is of practical interest, and we will show using numerical experiments (Section VI) that a wide range of values for p can potentially result in efficient algorithms.

One should observe that the number of competitive arms is a key factor in tuning p. We now provide another corollary of Proposition 3, which presents the suitable range of parameter p given the growth rate of competitive arms as follows.

Corollary 5: Let the number of competitive arms be an arbitrary function f_K of total arms, i.e., there exists a constant $\varepsilon \geq 0$ such that $\Delta_i/\Delta_2 \leq 1+\varepsilon$ for $i \in S$, where $|S| = f_K$, and for $i \notin S$, $\Delta_i/\Delta_2 \geq i$. Then, there exists a suitable choice of p for which `N-Seq-El` outperforms `Succ-Rej` and `Seq-Halv` in

TABLE II
THE MISIDENTIFICATION PROBABILITY FOR ALL THREE ALGORITHMS
DECAYS IN THE FORM OF $\beta \exp(-T/\alpha)$. THE TABLE REPRESENTS THE
PARAMETER α FOR EACH ALGORITHM IN REGIMES 1 TO 3. FOR REGIME 2,
WE SET $0 < p < 1$, AND FOR REGIMES 1 AND 3, WE USE $1 < p \leq 2$

	Succ-Rej	Seq-Halv	Our algorithm
Regime 1	$\mathcal{O}(\log K)$	$\mathcal{O}(\log K)$	$\mathcal{O}(1)$
Regime 2	$\mathcal{O}(K \log K)$	$\mathcal{O}(K \log K)$	$\mathcal{O}(K)$
Regime 3	$\mathcal{O}(\log K)$	$\mathcal{O}(\log K)$	$\mathcal{O}(1)$

TABLE III
GIVEN THAT THE MISIDENTIFICATION PROBABILITY DECAYS IN THE FORM OF
$\beta \exp(-T/\alpha)$, THE TABLE REPRESENTS THE CHOICE OF p FOR WHICH
THE PARAMETER α IN N-Seq-El IS SMALLER THAN THOSE
OF Succ-Rej AND Seq-Halv. NOTE THAT WE USE
$l_K := \max\{0, 1 - \log(\log K)/\log(K/f_K)\}$ AND
$u_K := \min\{2, 1 + \log(\log K)/\log(f_K)\}$

Condition on f_K	Suitable Range of Parameter p
$1 \leq f_K \leq \log K$	$1 < p \leq 2$
$\log K < f_K < \frac{K}{\log K}$	$\min\{l_K, 1\} < p < \max\{u_K, 1\}$
$\frac{K}{\log K} \leq f_K \leq K - 1$	$0 < p < 1$

TABLE IV
THE TABLE SHOWS THE SUITABLE INTERVAL OF p FOR
$K \in \{40, 120, 5 \times 10^2, 5 \times 10^4, 5 \times 10^6, 5 \times 10^8\}$ GIVEN THE GROWTH
RATE OF COMPETITIVE ARMS AS DERIVED IN COROLLARY 5

K \ f_K	$K^{0.3}$	$K^{0.5}$	$K^{0.7}$
40	$(0.5, 2)$	$(0.3, 1.7)$	$(0, 1.5)$
120	$(0.53, 2)$	$(0.35, 1.65)$	$(0, 1.47)$
$5 * 10^2$	$(0.58, 1.97)$	$(0.42, 1.58)$	$(0.03, 1.42)$
$5 * 10^4$	$(0.69, 1.73)$	$(0.56, 1.44)$	$(0.27, 1.31)$
$5 * 10^6$	$(0.75, 1.59)$	$(0.65, 1.35)$	$(0.41, 1.25)$
$5 * 10^8$	$(0.79, 1.5)$	$(0.7, 1.3)$	$(0.5, 1.21)$

the sense that the misidentification probability decays with a faster rate, i.e., we have $\mathcal{O}(H(p)C_p) \leq \mathcal{O}(H_2 \log K)$. For different conditions on the growth of f_K, the suitable choice of parameter p is presented in Table III.

The corollary above indicates that a side information in the form of the number of competitive arms (in the order) can help us tune the input parameter p. The corollary exhibits a smooth interpolation between when the competitive arms are small versus when they are large. Perhaps, the most interesting regime is the middle row, where the choice of p is given as a function of the number of arms K. Consider the following example where we calculate the choice of p for $f_K = K^\gamma$, where $\gamma \in \{0.3, 0.5, 0.7\}$.

As we can see in Table IV, since the condition on p depends in logarithmic orders on K and f_K, we have flexibility to tune p even for very large K. However, we only use $K \in \{40, 120\}$ for the numerical experiments in Section VI, since the time-

complexity of Monte Carlo simulations is prohibitive on large number of arms.

V. SIDE OBSERVATIONS

In the previous sections, we considered a scenario in which pulling an arm yields only the reward of the chosen arm. However, there exist applications where pulling an arm can additionally result in some side observations. For a motivative example, consider the problem of web advertising, where an ad placer offers an ad to a user and receives a reward only if the user clicks on the ad. In this example, if the user clicks on a vacation ad, the ad placer receives the side information that the user could have also clicked on ads for rental cars. The value of side observations in the stochastic MAB was studied in [22] for exploration-exploitation setting. The side-observation model is described via an undirected graph that captures the relationship between the arms. Once an arm is pulled, the player observes the reward of the arm as well as its neighboring arms. In exploration-exploitation settings, the analysis of MAB with side observations relies on the cliques of the graph [21], [22]. In this section, we would like to consider the impact in the pure-exploration setting.

In the pure-exploration, we minimize the simple regret rather than the cumulative regret, and therefore, the player's best bets are the most connected arms resulting in more observations. Now consider a partition of the set $[K]$ into M blocks $\{V_i\}_{i=1}^M$ such that each V_i contains a star graph, i.e., there exists an arm in V_i connected to all other arms. Given such partition, the player can follow a simple rule to leverage side observations and eliminate blocks one by one. The idea is to sample the central arm in each block, which reveals the rewards for all the arms in that block. At round r, sample the remaining blocks (the central arms of each block) for $n_r - n_{r-1}$ times, and find the arm with largest average reward in each block. With the best arms of all blocks at hand, remove the block whose best arm is the worst comparing to other blocks. Continue the sampling until only one block is left, and output the arm with largest average reward in the final block as the best arm. The algorithm, called Sequential Block Elimination, is summarized in Fig. 2, and its performance is characterized in the following theorem.

Theorem 6: Let the Sequential Block Elimination algorithm in Fig. 2 run given the input sequence $\{z_r\}_{r=1}^M$ and blocks $\{V_i\}_{i=1}^M$. Define $V := \max_{i \in [M]}\{|V_i|\}$ to be the maximum cardinality among the blocks. Then, the misidentification probability satisfies the bound

$$\mathbb{P}(\{\mathbf{J}\} \neq \{1\}) \leq VM \exp\left(-\frac{T-M}{C} 2 \min_{r \in [M]} \left\{\frac{\Delta_{M+1-r}^2}{z_r}\right\}\right).$$

In particular, the choice of $z_r = (M + 1 - r)^p$ for a given $p > 0$, yields

$$\mathbb{P}(\{\mathbf{J}\} \neq \{1\}) \leq VM \exp\left(-2\frac{T-M}{CH_{M,p}}\right),$$

where $H(M, p) := \max_{i \in [M] \setminus \{1\}}\{\frac{i^p}{\Delta_i^2}\}$.

The theorem indicates that once we can partition the arms into M blocks, the complexity measure $H(p)$ need not be maximized

---

Sequential Block Elimination Algorithm

Input: budget T, sequence $\{z_r\}_{r=1}^{M}$, blocks $\{V_i\}_{i=1}^{M}$.

Initialize: $\mathbf{G}_1 = \{V_1, \ldots, V_M\}$, $n_0 = 0$.

Let $C = \sum_{r=1}^{M} z_r^{-1}$ and $n_r = \left\lceil \frac{T-M}{C z_r} \right\rceil$ for $r \in [M]$

At round $r = 1, \ldots, M-1$:

(1) Sample each block in $\mathbf{G}_r$ for $n_r - n_{r-1}$ times.
(2) For any block $V_m \in \mathbf{G}_r$, let $\mathbf{y}_m \leftarrow \max_{i \in V_m}\{\bar{\mathbf{x}}_{i,n_r}\}$.
(3) Calculate $\mathbf{j} \leftarrow \operatorname{argmin}_m\{\mathbf{y}_m\}$ to locate the block with smallest maximum average.
(4) Let $\mathbf{B}_r = \{V_{\mathbf{j}}\}$, and eliminate the block $V_{\mathbf{j}}$, i.e. $\mathbf{G}_{r+1} = \mathbf{G}_r \setminus \mathbf{B}_r$.

If $r = M$: Stop at (2), and let $\mathbf{J}$ be the arm index with the largest empirical average in the only remaining block.

Output: $\{\mathbf{J}\}$

Fig. 2. Sequential Block Elimination discards the blocks one by one and selects the best arm in the last block.

over K arms. Instead, it is maximized over the top M arms. The improvement must be more visible in the settings that arms are roughly as competitive as each other, since we can slow down the linear growth of K by M.

As a final note, $M = K$ recovers the fully bandit (no side observation) setting with $V = 1$, and we observe that in such case, we recover the result of Proposition 3 from Theorem 6 in the rate of exponential decay. However, the constant behind the exponential would change from $K - 1$ in Proposition 3 to K in Theorem 6. This is the artifact of an extra term contributing to the upper bound, which can be removed for the case $M = K$.

VI. NUMERICAL EXPERIMENTS

In this section, we empirically evaluate our algorithm on the settings previously studied in [12], [18]. In these experiments, we compare N-Seq-El with Succ-Rej, Seq-Halv, and UCB-E. Though we include UCB-E proposed in [12] as a benchmark, we remark that the algorithm requires a prior knowledge of a parameter that depends on H_1 defined in (4). The adaptive version of UCB-E was also developed in [12]. The algorithm (called AUCB-E) does not need prior knowledge of H_1, and it calculates the parameter online. The experiments in [18] suggest that for $T \approx H_1$, Succ-Rej, Seq-Halv, and UCB-E outperform AUCB-E, and it is not surprising that the prior knowledge of H_1 must give UCB-E an advantage over AUCB-E.

We consider Bernoulli distribution on the rewards, assuming that the expected value of Bernoulli distribution for the best arm is $\mu_1 = 0.7$. In what follows, we use the notation $x : y$ to denote integers in $[x, y]$. We examine the following setups for two values of arm numbers $K \in \{40, 120\}$:

1) *One group of suboptimal arms:* $\mu_{2:K} = 0.6$.
2) *Two groups of suboptimal arms:* $\mu_{2:m} = 0.7 - \frac{2}{K}$, $\mu_{m+1:K} = 0.4$, and $m = \left\lceil \log \frac{K}{2} + 1 \right\rceil$.
3) *Three groups of suboptimal arms:* $\mu_{2:m} = 0.7 - \frac{2}{K}$, $\mu_{m+1:2m} = 0.7 - \frac{4}{K}$, $\mu_{2m+1:K} = 0.4$, and $m = \left\lceil \log \frac{K}{2} + 1 \right\rceil$.
4) *Arithmetic progression:* $\Delta_i = \frac{0.6(i-1)}{K-1}$ for $i = 2 : K$.
5) *Geometric progression:* $\Delta_i = 0.01(1 + \frac{4}{K})^{i-2}$ for $i = 2 : K$.

6) *One real competitive arm:* $\mu_2 = 0.7 - \frac{1}{2K}$ and $\mu_{3:K} = 0.2$.

We run 4000 experiments for each setup with specific K, and the misidentification probability is averaged out over the experiment runs. The budget T considered in each setup is equal to $\lceil H_1 \rceil$ in the corresponding setup following [12], [18]. Fig. 3 illustrates the performance of each algorithm in different setups. The height of each bar depicts the misidentification probability, and the index guideline is as follows: **(i)** indices 1-4: N-Seq-El with parameter $p = 0.75, 1.35, 1.7, 2$. **(ii)** index 5: Succ-Rej. **(iii)** index 6: Seq-Halv. **(iv)** indices 7-9: UCB-E with parameter $a = cT/H_1$, for $c = 1, 2, 4$. The legends are the same for all the plots, and we remove some of them to avoid clutter.

The results are consistent with Corollary 4, and the following comments are in order:

- Setup 1 perfectly relates to Regime 2 in Corollary 4 ($\varepsilon = 0$). While any choice of $0 < p < 1$ gives an $\mathcal{O}(K)$ rate, the rate deteriorates to $\mathcal{O}(K^p)$ by choosing $1 < p \leq 2$. Therefore, only the choice of $p = 0.75$ is appropriate for N-Seq-El. This choice joined with Succ-Rej (which amounts to our algorithm with $p = 1$) outperform others even UCB-E.

- Setup 4 defines Regime 1 in Corollary 4. Accordingly, choice of $1 < p \leq 2$ outperforms Succ-Rej and Seq-Halv, and $p = 1.7, 2$ prove to be competitive to UCB-E.

- In Setups 2-3, the number of competitive arms grows with $\mathcal{O}(\log K)$, but they can be considered close to Regime 3 in Corollary 4 as the growth is sub-linear. Also, Setup 6 is the ideal match for Regime 3 in Corollary 4. Therefore, any choice of $1 < p \leq 2$ should be suitable in these cases. We observe that for $p = 1.35, 1.7, 2$, N-Seq-El outperforms Succ-Rej and Seq-Halv, while being quite competitive to UCB-E in Setups 2-3 for $K = 40$.

- In Setup 5, we were expecting to observe good performance for $1 < p \leq 2$. While we do see that for $K = 40$, N-Seq-El outperforms Succ-Rej and Seq-Halv with $p = 1.35, 1.7$, the performance is not quite satisfactory for $K = 120$ (only $p = 1.35$ performs well). A poten-
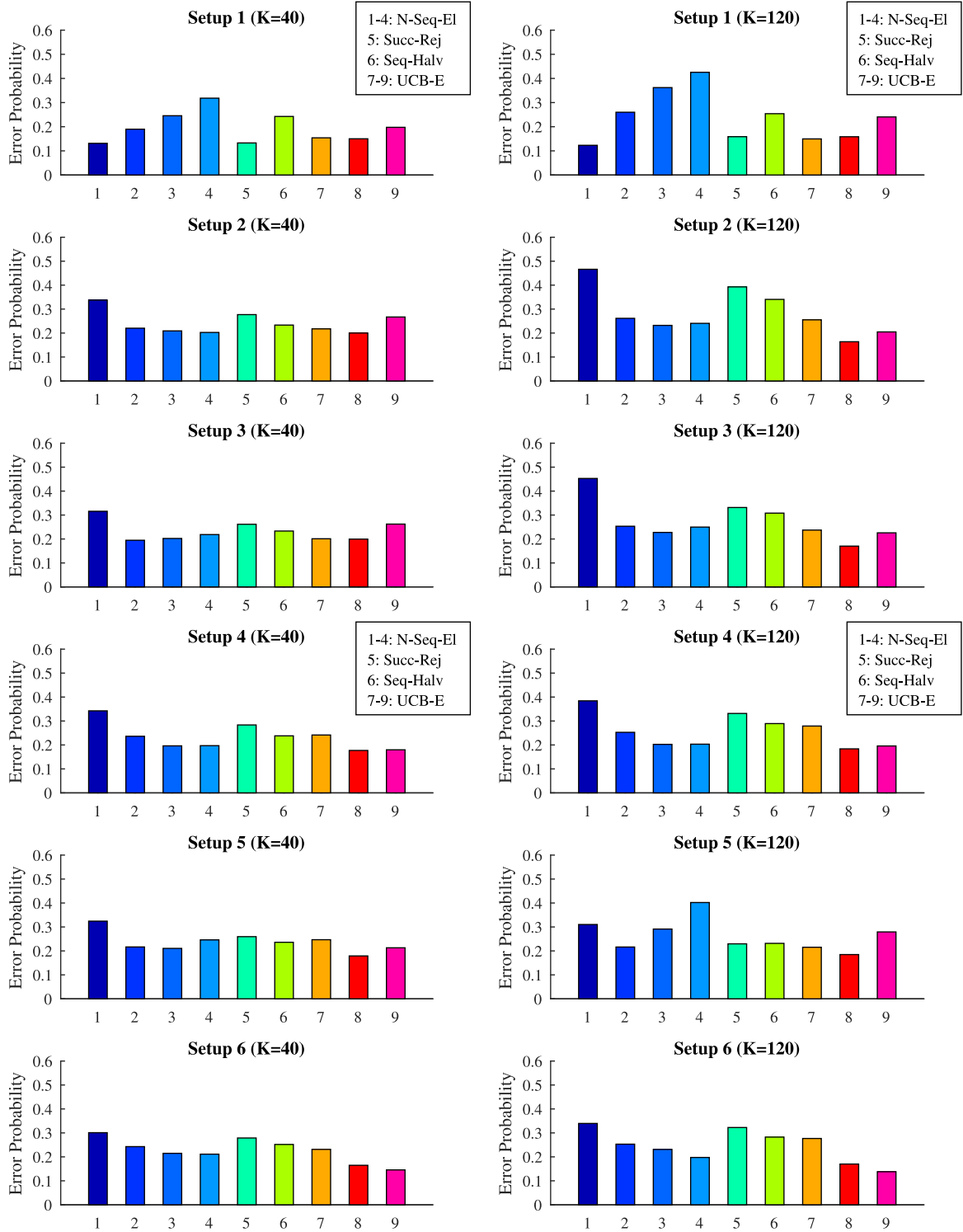
Fig. 3. The figure shows the misidentification probability for N-Seq-El, Succ-Rej, Seq-Halv, and UCB-E algorithms in six different setups. The six plots on the left hand side relate to the case $K = 40$, and the six plots on the right hand side are associated with $K = 120$. The height of each bar depicts the misidentification probability, and each index (or color) represents one algorithm tuned with a specific parameter in case the algorithm is parameter dependent.

tial reason is that we want to keep the rewards bounded in $[0, 1]$, thereby choosing the geometric rate for Δ_i so slow that it is dominated by i^p rate in (5). To support our argument, we present a complementary evaluation of this case for a small K, where we can use faster geometric growth.

- As a final remark, note that in consistent with Table II, for relevant cases we observe an improvement of performance when K increases. For instance, in Setup 1 with $p = 0.75$, the ratio of misidentification probability of N-Seq-El to Seq-Halv increases from 1.84 ($K = 40$) to 2.06
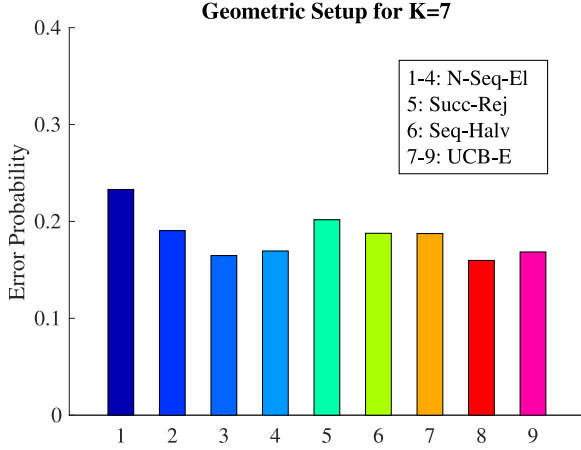
Fig. 4. The figure shows the misidentification probability for `N-Seq-El`, `Succ-Rej`, `Seq-Halv`, and `UCB-E` algorithms for an experiment with $K = 7$. The height of each bar depicts the misidentification probability, and each index (or color) represents one algorithm tuned with a specific parameter in case the algorithm is parameter dependent.

$(K = 120)$, or in Setup 4 with $p = 2$, the ratio of misidentification probability of `N-Seq-El` to `Succ-Rej` increases from 1.43 $(K = 40)$ to 1.63 $(K = 120)$.

A. Complementary Numerical Experiment for Geometric Sub-Optimality Setup

As we remarked in the very last comment of numerical experiments, choosing small number of arms, we can space the expected value of arm rewards such that in (5) the sub-optimality term Δ_i dominates i^p for $1 < p \leq 2$. We propose a simple experiment for $K = 7$ arms, where $\mu_1 = 0.7$ and $\Delta_i = (0.6)^{8-i}$ in Fig. 4. As expected, `N-Seq-El` tuned with $p > 1$ is competitive, and it achieves its best performance with $p = 1.7$, winning the competition against others, while almost equalizing with `UCB-E` for $c = 2$.

VII. CONCLUSION

We considered best-arm identification in the stochastic multi-armed bandits, where a player is given a certain budget to explore a finite number of arms. The player's objective is to detect the arm with largest expected reward. We contribute to the literature of best-arm identification by **(i)** unifying sequential elimination algorithms under a general framework, which introduces a novel performance metric for this class of algorithms, **(ii)** developing a nonlinear sequential elimination algorithm with provable theoretical and practical guarantees, **(iii)** and establishing a theoretical result on the value of side observations in the pure-exploration setting.

Having established that we gain advantage from nonlinear budget allocation, an important future direction is to propose a method that starts with a specific nonlinear rate and finetunes the rate according to the problem environment. The main challenge is that the quantity C should perhaps be time-varying, and its value needs to be cautiously controlled by the algorithm, so that the algorithm does not overspend the budget.

APPENDIX

Fact 1: (Hoeffding's inequality) Let $W_1, \ldots, W_n$ be independent random variables with support on the unit interval with probability one. If $S_n = \sum_{i=1}^n W_i$, then for all $a > 0$, it holds that

$$\mathbb{P}\left(S_n - \mathbb{E}[S_n] \geq a\right) \leq \exp\left(\frac{-2a^2}{n}\right).$$

Proof of Theorem 2: Recall that $\mathbf{G}_r$ denotes the set of arms not eliminated by the start of round r with $g_r = |\mathbf{G}_r|$ being its cardinality. Also, $\mathbf{B}_r$ represents the set of arms that we decide to discard after playing round r with $b_r = |\mathbf{B}_r|$ denoting its cardinality. It evidently holds that $\mathbf{G}_{r+1} = \mathbf{G}_r \setminus \mathbf{B}_r$ for $r \in [R]$. Therefore,

$$\mathbb{P}\left(\mathbf{G}_{R+1} \neq \{1\}\right) = \mathbb{P}\left(1 \in \cup_{r=1}^R \mathbf{B}_r\right) = \sum_{r=1}^R \mathbb{P}\left(1 \in \mathbf{B}_r\right), \tag{6}$$

since the sets of removed arms at each round are disjoint, i.e. $\mathbf{B}_i \cap \mathbf{B}_j = \emptyset$ for $i \neq j$. We can then write

$$\mathbb{P}\left(1 \in \mathbf{B}_r\right) = \sum_{G_r} \mathbb{P}\left(1 \in \mathbf{B}_r \mid \mathbf{G}_r = G_r\right) \mathbb{P}\left(\mathbf{G}_r = G_r\right). \tag{7}$$

Now for any particular G_r, consider the worst b_r arms, i.e., the bottom b_r arms when arms are ordered in terms of (true) expected value. If the best arm (arm 1) is set to be eliminated at the end of round r, its empirical average must be less than at least one of these b_r arms. In the case that $G_r = \{1, 2, \ldots, g_r\}$, the bottom b_r arms would be $\{g_r - b_r + 1, \ldots, g_r\}$. Therefore, recalling that $\bar{\mathbf{x}}_{i,n}$ denotes the average reward of pulling arm i for n times, and using Hoeffding's inequality (Fact 1), we get

$$\mathbb{P}\left(1 \in \mathbf{B}_r \mid \mathbf{G}_r = \{1, 2, \ldots, g_r\}\right)$$

$$\leq \sum_{i=g_r - b_r + 1}^{g_r} \mathbb{P}\left(\bar{\mathbf{x}}_{1,n_r} \leq \bar{\mathbf{x}}_{i,n_r}\right)$$

$$\leq b_r \exp\left(-2n_r \Delta_{g_r - b_r + 1}^2\right)$$

$$= b_r \exp\left(-2n_r \Delta_{g_{r+1}+1}^2\right), \tag{8}$$

where the last step is due to the fact that $g_{r+1} = g_r - b_r$. In any other case for G_r the best of the worst b_r arms cannot be better than arm $g_r - b_r + 1$. As a result, combining (7) and (8), we obtain

$$\mathbb{P}\left(1 \in \mathbf{B}_r\right) \leq \sum_{G_r} b_r \exp\left(-2n_r \Delta_{g_{r+1}+1}^2\right) \mathbb{P}\left(\mathbf{G}_r = G_r\right)$$

$$= b_r \exp\left(-2n_r \Delta_{g_{r+1}+1}^2\right).$$

Then, in view of (6) we derive

$$\mathbb{P}\left(\mathbf{G}_{R+1} \neq \{1\}\right) \leq \sum_{r=1}^R b_r \exp\left(-2n_r \Delta_{g_{r+1}+1}^2\right)$$

$$\leq R \max_{r \in [R]} \{b_r\} \max_{r \in [R]} \left\{\exp\left(-2n_r \Delta_{g_{r+1}+1}^2\right)\right\}.$$

Noting the fact that $n_r = \left\lceil \frac{T-K}{C z_r} \right\rceil \geq \frac{T-K}{C z_r}$, we can use above to conclude that

$$\mathbb{P}\left(\mathbf{G}_{R+1} \neq \{1\}\right)$$
$$\leq R \max_{r \in [R]}\{b_r\} \max_{r \in [R]} \left\{ \exp\left(-\frac{T-K}{C z_r} 2\Delta_{g_{r+1}+1}^2 \right) \right\},$$

which completes the proof. ∎

Proof of Proposition 3: We point out that the algorithm is a special case of General Sequential Elimination where $R = K - 1$, $b_r = 1$, $g_r = K - r + 1$, and $z_r = (K - r + 1)^p$. The proof then follows immediately from the result of Theorem 2. ∎

Proof of Corollary 4: The proof follows by substituting each case in (5). We need to understand the order of C_p and $H(p)$ for different regimes of p. Let us start by

$$C_p = 2^{-p} + \sum_{r=2}^{K} r^{-p},$$

and noting that for any $p > 1$, C_p is a convergent sum when $K \to \infty$. Therefore, for the regime $p > 1$, the sum is a constant, i.e., $C_p = \mathcal{O}(1)$. On the other hand, consider $q \in (0,1)$, and note that the sum is divergent, and for large K we have $C_q = \mathcal{O}(K^{1-q})$. Now, let us analyze

$$H(p) = \max_{i \neq 1} \frac{i^p}{\Delta_i^2}.$$

For Regime 1, since $\Delta_i = (i-1)\Delta_0$, for $p \in (1, 2)$, we have

$$H(p) = \max_{i \neq 1} \left(1 + \frac{1}{i-1} \right)^p \frac{(i-1)^{p-2}}{\Delta_0^2}$$
$$\leq \max_{i \neq 1} \left(1 + \frac{1}{i-1} \right)^p \frac{1}{\Delta_0^2} = \frac{1.5^p}{\Delta_0^2},$$

which is of constant order with respect to K. Therefore, the product $C_p H(p) = \mathcal{O}(1)$.

For Regime 2, we have

$$H(q) = \max_{i \neq 1} \frac{i^q}{\Delta_i^2} \leq \max_{i \neq 1} \frac{i^q}{\Delta_2^2} \leq \frac{K^q}{\Delta_2^2},$$

and the maximum order can be achieved as the number of arms close to the second best arm grows linearly in K. Combining with $C_q = \mathcal{O}(K^{1-q})$, the product $C_q H(q) = \mathcal{O}(K)$.

For Regime 3, if $i \in S$, we have

$$\max_{i \in S} \frac{i^p}{\Delta_i^2} = \mathcal{O}(1),$$

since the cardinality of S is of constant order with respect to K. On the other hand, since $1 < p \leq 2$, we have

$$\max_{i \notin S} \frac{i^p}{\Delta_i^2} \leq \max_{i \notin S} \frac{i^p}{i^2 \Delta_2^2} = \max_{i \notin S} \frac{i^{p-2}}{\Delta_2^2} = \mathcal{O}(1).$$

Therefore, $H(p)$ is of constant order, and combining with $C_p = \mathcal{O}(1)$, the product $C_p H(p) = \mathcal{O}(1)$. ∎

Proof of Corollary 5: According to calculations in the previous proof, we have that $C_p = \mathcal{O}(K^{1-p})$ for $0 < p < 1$, and $C_p = \mathcal{O}(1)$ for $1 < p \leq 2$. In order to find out the order of $H(p)$ and H_2, we should note that

$$\mathcal{O}(H(p)) = \mathcal{O}\left(\max_{i \in S} \frac{i^p}{\Delta_i^2} \right) = \mathcal{O}\left(\max_{i \in S} \frac{i^p}{\Delta_2^2} \right) = f_K^p,$$

and we simply have $\mathcal{O}(H_2) = \mathcal{O}(H(1)) = f_K$. Now returning to Table I, we must compare $\mathcal{O}(H(p)C_p) = f_K^p \mathcal{O}(C_p)$ and $\mathcal{O}(H_2 \log K) = f_K \log K$ for all three the rows given in Table III.

Row 1: In the case that $1 \leq f_K \leq \log K$, for any choice of $1 < p \leq 2$, since $C_p = \mathcal{O}(1)$, we always have

$$f_K^p \leq f_K \log K \Leftrightarrow f_K^{p-1} \leq \log K.$$

Row 3: In the case that $\frac{K}{\log K} \leq f_K \leq K - 1$, for any choice of $0 < p < 1$, since $C_p = \mathcal{O}(K^{1-p})$, we have

$$f_K^p \mathcal{O}(C_p) = f_K^p K^{1-p} = f_K \left(\frac{K}{f_K} \right)^{1-p} \leq f_K \log K.$$

Row 2: In this case, we have that $\log K < f_K < \frac{K}{\log K}$. To prove the claim in the Table III, we have to break down the analysis into two cases:

Row 2 – Case 1: First, consider the case $1 - \frac{\log \log K}{\log\left(\frac{K}{f_K}\right)} < p < 1$. Since $C_p = \mathcal{O}(K^{1-p})$, we have that

$$f_K^p \mathcal{O}(C_p) = f_K^p K^{1-p} = f_K \left(\frac{K}{f_K} \right)^{1-p}$$
$$\leq f_K \left(\frac{K}{f_K} \right)^{\frac{\log \log K}{\log\left(\frac{K}{f_K}\right)}} = f_K \log K.$$

Row 2 – Case 2: Second, consider the case $1 < p < 1 + \frac{\log \log K}{\log f_K}$. Since $C_p = \mathcal{O}(1)$, we have that

$$f_K^p \mathcal{O}(C_p) = f_K^p \leq f_K^{1 + \frac{\log \log K}{\log f_K}} = f_K f_K^{\frac{\log \log K}{\log f_K}} = f_K \log K.$$

Therefore, for all of the conditions on f_K, we showed proper choice of p which guarantees $f_K^p \mathcal{O}(C_p) \leq f_K \log K$. ∎

Proof of Theorem 6: Recall that the elements of $\mathbf{G}_r$ are the set of $M + 1 - r$ blocks not eliminated by the start of round $r \in [M]$, and $\mathbf{B}_r$ only contains the block we decide to discard after playing round r. Also, recall that arm 1 is the best arm, located in the block V_1, and $V := \max_{i \in [M]}\{|V_i|\}$ denotes the maximum cardinality of the blocks.

If the algorithm does not output the best arm, the arm is either eliminated with block V_1 in one the rounds in $[M-1]$, or not selected at the final round M. Therefore,

$$\mathbb{P}\left(\mathbf{J} \neq 1\right) \leq \mathbb{P}\left(\mathrm{argmax}_{i \in V_1}\{\bar{\mathbf{x}}_{i,n_M}\} \neq 1 \mid \mathbf{G}_M = \{V_1\} \right)$$
$$+ \sum_{r=1}^{M-1} \mathbb{P}\left(\mathbf{B}_r = \{V_1\}\right). \tag{9}$$

The first term can be bounded simply by Hoeffding's inequality

$$\mathbb{P}\left(\operatorname{argmax}_{i \in V_1}\{\bar{\mathbf{x}}_{i,n_M}\} \neq 1 \mid \mathbf{G}_M = \{V_1\}\right)$$

$$\leq |V_1| \exp\left(-2n_M\Delta_2^2\right) = |V_1| \exp\left(-2n_M\Delta_1^2\right), \quad (10)$$

using the convention $\Delta_1 := \Delta_2$. On the other hand, for any $r \in [M-1]$

$$\mathbb{P}\left(\mathbf{B}_r = \{V_1\}\right)$$

$$= \sum_{G_r} \mathbb{P}\left(\mathbf{B}_r = \{V_1\} \mid \mathbf{G}_r = G_r\right) \mathbb{P}\left(\mathbf{G}_r = G_r\right). \quad (11)$$

Note that, without loss of generality, we sort the blocks based on their best arm as

$$\mu_1 = \max_{i \in V_1} \mu_i > \max_{i \in V_2} \mu_i \geq \cdots \geq \max_{i \in V_M} \mu_i,$$

so the best possible arm in V_m cannot be better than arm m in view of above and (1). We remove block V_1 after execution of round r only if it is the worst among all other candidates. Therefore, consider the particular case that $G_r = \{V_1, V_2, \ldots, V_{M+1-r}\}$ contains the best possible $M + 1 - r$ blocks that one can keep until the start of round r. In such case,

$$\mathbb{P}\left(\mathbf{B}_r = \{V_1\} \mid \mathbf{G}_r = \{V_1, V_2, \ldots, V_{M+1-r}\}\right)$$

$$\leq V \exp\left(-2n_r\Delta_{M+1-r}^2\right).$$

In any other case for G_r the best possible arm in the worst block cannot be better than arm $M + 1 - r$. Therefore, combining above with (11), we obtain

$$\mathbb{P}\left(\mathbf{B}_r = \{V_1\}\right) \leq V \sum_{G_r} \exp\left(-2n_r\Delta_{M+1-r}^2\right) \mathbb{P}\left(\mathbf{G}_r = G_r\right)$$

$$= V \exp\left(-2n_r\Delta_{M+1-r}^2\right).$$

Incorporating above and (10) into (9), we derive

$$\mathbb{P}\left(\mathbf{J} \neq 1\right) \leq |V_1| \exp\left(-2n_M\Delta_1^2\right)$$

$$+ \sum_{r=1}^{M-1} V \exp\left(-2n_r\Delta_{M+1-r}^2\right)$$

$$\leq V \sum_{r=1}^{M} \exp\left(-2n_r\Delta_{M+1-r}^2\right)$$

$$\leq VM \exp\left(-\frac{T-M}{C}2 \min_{r \in [M]}\left\{\frac{\Delta_{M+1-r}^2}{z_r}\right\}\right),$$

noticing the choice of n_r in the algorithm (Fig. 2). Having finished the first part of the proof, when we set $z_r = (M+1-r)^p$ for $r \in [M-1]$ and $z_M = 2^p$, we have

$$\mathbb{P}\left(\mathbf{J} \neq 1\right) \leq VM \exp\left(-2\frac{T-M}{CH_{M,p}}\right),$$

where $H(M,p) := \max_{i \in [M] \setminus \{1\}}\left\{\frac{i^p}{\Delta_i^2}\right\}$. ∎

References

[1] T. L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Adv. Appl. Math.*, vol. 6, no. 1, pp. 4–22, 1985.

[2] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Mach. Learn.*, vol. 47, no. 2–3, pp. 235–256, 2002.

[3] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire, "The non-stochastic multiarmed bandit problem," *SIAM J. Comput.*, vol. 32, no. 1, pp. 48–77, 2002.

[4] S. Bubeck and N. Cesa-Bianchi, "Regret analysis of stochastic and non-stochastic multi-armed bandit problems," *Found. Trends Mach. Learn.*, vol. 5, no. 1, pp. 1–122, 2012.

[5] A. Mahajan and D. Teneketzis, "Multi-armed bandit problems," in *Foundations and Applications of Sensor Management*. Berlin, Germany: Springer-Verlag, pp. 121–151, 2008.

[6] K. Liu and Q. Zhao, "Distributed learning in multi-armed bandit with multiple players," *IEEE Trans. Signal Process.*, vol. 58, no. 11, pp. 5667–5681, Nov. 2010.

[7] K. Wang and L. Chen, "On optimality of myopic policy for restless multi-armed bandit problem: An axiomatic approach," *IEEE Trans. Signal Process.*, vol. 60, no. 1, pp. 300–309, Jan. 2012.

[8] S. Vakili, K. Liu, and Q. Zhao, "Deterministic sequencing of exploration and exploitation for multi-armed bandit problems," *IEEE J. Sel. Top. Signal Process.*, vol. 7, no. 5, pp. 759–767, Oct. 2013.

[9] D. Kalathil, N. Nayyar, and R. Jain, "Decentralized learning for multiplayer multiarmed bandits," *IEEE Trans. Inf. Theory*, vol. 60, no. 4, pp. 2331–2345, Apr. 2014.

[10] S. Bagheri and A. Scaglione, "The restless multi-armed bandit formulation of the cognitive compressive sensing problem," *IEEE Trans. Signal Process.*, vol. 63, no. 5, pp. 1183–1198, Mar. 2015.

[11] K. Kanoun, C. Tekin, D. Atienza, and M. Van Der Schaar, "Big-data streaming applications scheduling based on staged multi-armed bandits," *IEEE Trans. Comput.*, vol. 65, no. 12, pp. 3591–3605, Dec. 2016.

[12] J.-Y. Audibert and S. Bubeck, "Best arm identification in multi-armed bandits," in *Proc. 23rd Conf. Learn. Theory*, 2010.

[13] E. Even-Dar, S. Mannor, and Y. Mansour, "PAC bounds for multi-armed bandit and Markov decision processes," in *Proc. Comput. Learn. Theory*. Berlin, Germany: Springer-Verlag, 2002, pp. 255–270.

[14] S. Mannor and J. N. Tsitsiklis, "The sample complexity of exploration in the multi-armed bandit problem," *J. Mach. Learn. Res.*, vol. 5, pp. 623–648, 2004.

[15] E. Even-Dar, S. Mannor, and Y. Mansour, "Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems," *J. Mach. Learn. Res.*, vol. 7, pp. 1079–1105, 2006.

[16] S. Bubeck, R. Munos, and G. Stoltz, "Pure exploration in multi-armed bandits problems," in *Algorithmic Learning Theory*. Berlin, Germany: Springer-Verlag, 2009, pp. 23–37.

[17] S. Bubeck, R. Munos, and G. Stoltz, "Pure exploration in finitely-armed and continuous-armed bandits," *Theor. Comput. Sci.*, vol. 412, no. 19, pp. 1832–1852, 2011.

[18] Z. Karnin, T. Koren, and O. Somekh, "Almost optimal exploration in multi-armed bandits," in *Proc. 30th Int. Conf. Mach. Learn.*, 2013, pp. 1238–1246.

[19] V. Gabillon, M. Ghavamzadeh, and A. Lazaric, "Best arm identification: A unified approach to fixed budget and fixed confidence," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 3212–3220.

[20] A. Carpentier and A. Locatelli, "Tight (lower) bounds for the fixed budget best arm identification bandit problem," in *Proc. 29th Annu. Conf. Learn. Theory*, 2016, pp. 590–604.

[21] S. Mannor and O. Shamir, "From bandits to experts: On the value of side-observations," in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, pp. 684–692.

[22] S. Caron, B. Kveton, M. Lelarge, and S. Bhagat, "Leveraging side observations in stochastic bandits," in *Proc. 28th Conf. Uncertainty Artif. Intell.*, 2012, pp. 142–151.

[23] S. Buccapatnam, A. Eryilmaz, and N. B. Shroff, "Stochastic bandits with side observations on networks," in *Proc. ACM Int. Conf. Meas. Model. Comput. Syst.*, 2014, pp. 289–300.

[24] K. Jamieson, M. Malloy, R. Nowak, and S. Bubeck, "lil' UCB: An optimal exploration algorithm for multi-armed bandits," in *Proc. 27th Conf. Learn. Theory*, 2014, pp. 423–439.

[25] K. Jamieson and R. Nowak, "Best-arm identification algorithms for multi-armed bandits in the fixed confidence setting," in *Proc. 2014 48th Annu. Conf. Inf. Sci. Syst.*, 2014, pp. 1–6.

[26] E. Kaufmann and S. Kalyanakrishnan, "Information complexity in bandit subset selection," in *Proc. Conf. Learn. Theory*, 2013, pp. 228–251.

[27] E. Kaufmann, O. Cappé, and A. Garivier, "On the complexity of best-arm identification in multi-armed bandit models," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 1–42, 2016.

[28] D. Russo, "Simple Bayesian algorithms for best arm identification," in *Proc. 29th Annu. Conf. Learn. Theory*, 2016.

[29] S. Kalyanakrishnan and P. Stone, "Efficient selection of multiple bandit arms: Theory and practice," in *Proc. 27th Int. Conf. Mach. Learn.*, 2010, pp. 511–518.

[30] S. Kalyanakrishnan, A. Tewari, P. Auer, and P. Stone, "PAC subset selection in stochastic multi-armed bandits," in *Proc. 29th Int. Conf. Mach. Learn.*, 2012, pp. 655–662.

[31] S. Bubeck, T. Wang, and N. Viswanathan, "Multiple identifications in multi-armed bandits," in *Proc. 30th Int. Conf. Mach. Learn.*, 2013, pp. 258–265.

[32] Y. Zhou, X. Chen, and J. Li, "Optimal PAC multiple arm identification with applications to crowdsourcing," in *Proc. 31st Int. Conf. Mach. Learn.*, 2014, pp. 217–225.

[33] T. Pepels, T. Cazenave, M. H. Winands, and M. Lanctot, "Minimizing simple and cumulative regret in Monte-Carlo tree search," in *Computer Games*. Berlin, Germany: Springer-Verlag, 2014, pp. 1–15.

[34] Y.-C. Liu and Y. Tsuruoka, "Regulation of exploration for simple regret minimization in Monte-Carlo tree search," in *Proc. IEEE Conf. Comput. Intell. Games*, 2015, pp. 35–42.

Shahin Shahrampour received the B.Sc. degree from Sharif University of Technology, Tehran, Iran, in 2009 and the M.S.E. degree in electrical engineering, the M.A. degree in statistics, and the Ph.D. degree in electrical and systems engineering from the University of Pennsylvania, PA, USA, in 2012, 2014, and 2015, respectively. He is currently a Postdoctoral Fellow in electrical engineering at Harvard University, Cambridge, MA, USA. His research interests include machine learning, online learning, optimization, and distributed algorithms.

Mohammad Noshad received the Ph.D. degree in electrical engineering from the University of Virginia, Charlottesville, VA, USA, in 2013. He was a Postdoctoral researcher at Harvard University. He is currently the Chief Technology Officer at VLNComm, Charlottesville, VA, USA. His research interests include visible light communications, free-space optical communications, time-series analysis, and machine learning. He received the Best Paper Award at the IEEE GLOBECOM 2012.

Vahid Tarokh is currently a Professor of applied mathematics, a Senior Research Fellow in electrical engineering, and a member of Center for Mathematics and Applications at Harvard University, Cambridge, MA, USA. In the past, he has published on Lie groups and algebras, lattices, graphical models, pseudorandomness, information theory, random matrix theory, coding, communications, pricing and scheduling, statistical signal processing (including adaptive algorithms) and applications (e.g., MRI and EEG signal processing), radar, and distributed astronomy. His main research interests include applied and computational mathematics. He has been selected as a Guggenheim Fellow in applied mathematics (for his work on random and pseudorandom matrices), and has received four honorary degrees.