Fifty-Fifth Annual Allerton Conference
Allerton House, UIUC, Illinois, USA
October 3-6, 2017

Nonlinear Sequential Accepts and Rejects for Identification of Top Arms in Stochastic Bandits

Shahin Shahrampour and Vahid Tarokh

Abstract— **We address the M-best-arm identification problem in multi-armed bandits. A player has a limited budget to explore K arms ($M < K$), and once pulled, each arm yields a reward drawn (independently) from a fixed, unknown distribution. The goal is to find the top M arms in the sense of expected reward. We develop an algorithm which proceeds in rounds to deactivate arms iteratively. At each round, the budget is divided by a nonlinear function of remaining arms, and the arms are pulled correspondingly. Based on a decision rule, the deactivated arm at each round may be accepted or rejected. The algorithm outputs the accepted arms that should ideally be the top M arms. We characterize the decay rate of the misidentification probability and establish that the nonlinear budget allocation proves to be useful for different problem environments (described by the number of competitive arms). We provide comprehensive numerical experiments showing that our algorithm outperforms the state-of-the-art using suitable nonlinearity.**

I. Introduction

Multi-Armed Bandits (MAB) is a sequential decision-making framework for the exploration-exploitation dilemma [1], [2]. In MAB, a player explores a finite set of arms, and pulling each arm reveals a *reward* to the player. In the stochastic MAB, the rewards for each arm are independent samples from an *unknown*, fixed distribution. The player aims to exploit the arm with the largest expected reward as often as possible to maximize the gain. This framework has been formulated in terms of the *cumulative* regret, a comparison measure between the player's performance versus a clairvoyant knowing the best arm *a priori*. Early studies on MAB dates back to several decades ago, but the problem has attracted a lot of renewed interest due to its modern applications, such as web search and advertising, wireless cognitive radios, and multi-channel communication systems (see e.g. [3]–[7] and references therein).

More recently, many researchers have examined MAB in a pure-exploration framework where the player aims to minimize the *simple* regret. This task is closely related to (probability of) finding the best arm in the pool [8]. As a result, the best-arm identification problem has received a considerable attention in the literature of machine learning [8]–[14]. It is well-known that algorithms developed to minimize the cumulative regret (exploration-exploitation) perform poorly for the simple-regret minimization (pure-exploration). Consequently, one must adopt different strategies for optimal best-arm recommendation [12]. To motivate the pure-exploration setting, consider 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 [8].

In this paper, we consider the M-best-arm identification problem in the *fixed-budget* setting [15]. Given a fixed number of arm pulls, the player attempts to maximize the probability of correctly identifying the top M arms (in the sense of the expected reward). Note that this setting differs from the *fixed-confidence* setting, in which the objective is to minimize the number of trials to find the top M arms with a certain confidence [16], [17]. Recently, for best-arm identification ($M = 1$) in the fixed-budget setting, the authors of [18] proposed an efficient algorithm based on nonlinear sequential elimination. The idea is to discard the suboptimal arms sequentially and divide the budget according to a *nonlinear function* of remaining arms at each round. With a suitable nonlinearity, the nonlinear budget allocation was proven to improve upon Successive Rejects [8] (its linear counterpart) as well as Sequential Halving [13].

Inspired by the success of nonlinear budget allocation for best-arm identification [18], in this work, we extend the Successive Accepts and Rejects (`SAR`) algorithm in [15] to nonlinear budget allocation for M-best-arm identification. Our algorithm, called Nonlinear Sequential Accepts and Rejects (`NSAR`), proceeds in rounds. At each round, the arms are pulled strategically and their empirical rewards are calculated. Then, one arm is deactivated, and according to a decision rule the arm may be accepted or rejected. Unlike `SAR` that divides the budget by a linear function of remaining arms, `NSAR` (our algorithm) does so in a nonlinear fashion. For two general reward regimes, we prove theoretically that our algorithm achieves a lower sample complexity compared to `SAR`, which improves the decay rate of the misidentification probability. We also provide various numerical experiments to support our theoretical results, and moreover, we compare `NSAR` to the fixed-budget version of `AT-LUCB` in [19].

A. Related Work

Pure-exploration in the PAC-learning setup was examined in [9], where Successive Elimination for finding an ϵ-optimal

This work was supported by DARPA under grant number N66001-15-C-4028.

S. Shahrampour and V. Tarokh are with the John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, 02138 USA. (e-mail: {shahin,vahid}@seas.harvard.edu).

978-1-5386-3266-6/17/$31.00 ©2017 IEEE

arm with probability $1-\delta$ (fixed-confidence setting) was developed. The matching lower bounds for the problem were provided in [10], [20]. Many algorithms for pure-exploration are inspired by the celebrated UCB1 algorithm for exploration-exploitation [2]. As an example, Audibert et al. [8] proposed UCB-E, which modifies UCB1 for pure-exploration. In addition, Jamieson et al. [21] proposed an optimal algorithm for the fixed-confidence setting, inspired by the law of the iterated logarithm. Gabillon et al. [14] presented a unifying approach for fixed-budget and fixed-confidence settings. For identification of multiple top arms (or M-best-arm identification), Kalyanakrishnan et al. [16] developed the HALVING algorithm in the fixed-confidence setting, which is later improved by the LUCB algorithm in [17]. For the fixed-confidence setting, more recent progress can be found in [22]–[24]. In [25], the M-best-arm identification problem was posed using a notion of aggregate regret, and it was applied to crowdsourcing. Furthermore, Kaufmann et al. [26] studied the identification of multiple top arms using KL-divergence-based confidence intervals. The authors of [27] investigated both settings to show that the complexity of the fixed-budget setting may be smaller than that of the fixed-confidence setting.

II. Preliminaries

Notation: For integer K, we define $[K] := \{1, \dots, K\}$ to represent the set of positive integers smaller than or equal to K. We use $|S|$ to denote the cardinality of the set S, and $\lceil \cdot \rceil$ to denote the ceiling function, respectively. We use the notation $f(x) = \mathcal{O}(g(x))$ when there exists a positive constant $L > 0$ and a point x_0 such that $|f(x)| \le L\,|g(x)|$ for $x \ge x_0$. Throughout, the random variables are denoted in bold letters.

A. Problem Statement

In the stochastic Multi-armed Bandit (MAB) problem, a player explores a finite set of K arms. When the player samples an arm, the corresponding *reward* of that arm is observed. The rewards of each arm $i \in [K]$ are drawn independently from an *unknown, fixed* distribution with the expected value μ_i. The support of the distribution is the unit interval $[0,1]$, and the rewards are generated independently across the arms. For simplicity, we have the following assumption on the order of arms

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

where the strict inequalities guarantee that there is no ambiguity over the top M arms $[M]$. Let $\Delta_i := \mu_1 - \mu_i$ denote the *gap* between arm i and arm 1, measuring the sub-optimality of arm i, and $\widehat{\boldsymbol{\mu}}_{i,n}$ the (empirical) average reward obtained by pulling arm i for n times.

In this work, we address the M-best-arm identification setup, a pure-exploration problem in which the player aims to find the top M arms $[M]$ with a high probability. The two well-known settings for this problem are the fixed-confidence and the fixed-budget. In the former, the objective is to minimize the number of arm pulls needed to identify the top M arms with a certain confidence. In the latter, which is the focus of this work, the problem is posed formally as:

Problem 1: Given a total budget of T arm pulls, an M-best-arm identification algorithm outputs the arms $\{\mathbf{J}_1, \dots, \mathbf{J}_M\}$. Find the decay rate of misidentification probability, i.e., the decay rate of $\mathbb{P}(\{\mathbf{J}_1, \dots, \mathbf{J}_M\} \neq [M])$.

For the case that $M=1$, known as best-arm identification, it is proven that classical MAB techniques in the exploration-exploitation setting (e.g. UCB1) are not optimal. In particular, Bubeck et al. [12] have showed that upper bounds on the cumulative regret results in lower bounds on the simple regret, i.e., the smaller the cumulative regret, the larger the simple regret. The underlying intuition is that in the exploration-exploitation setting, we aim to find the best arm *as quickly as possible* to exploit it, and in this case, playing even the second-best arm for a long time yields an unacceptable cumulative regret. On the other hand, in the best-arm identification problem, there is no need to minimize an intermediate cost, and the player only recommends the best arm at the end. Therefore, exploring the suboptimal arms *strategically* during the game helps the player to make a better final decision. In other words, the performance is only measured by the final output, regardless of the number of pulls for the suboptimal arms.

B. Previous Performance Guarantees and Our Result

Though the focus of this work is M-best-arm identification, we start by reviewing some of the results for the case of $M = 1$ (best-arm identification). Any (single) best-arm identification algorithm samples the arms based on some strategy and outputs a single arm as the best. In order to characterize the misidentification probability of these algorithms, we need to define a few quantities. The decay rate of misidentification probability for two of the state-of-the-art algorithms, Successive Rejects [8] and Sequential Halving [13], relies on the complexity measure H_2, defined as

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

which is equal to H_1 up to logarithmic factors in K [8]. In Successive Rejects, at round r, the $K-r+1$ remaining arms are played proportional to the whole budget divided by $K-r+1$ (a linear function of r). As the linear function is not necessarily the best sampling rule, the authors of [18] extended Successive Rejects to Nonlinear Sequential Elimination which divides the budget at round r by the nonlinear function $(K-r+1)^p$, based on an input parameter $p \in (0,2]$ ($p=1$ recovers Successive Rejects). The performance of the algorithm depends on the following quantities

$$H(p) := \max_{i\neq 1} \frac{i^p}{\Delta_i^2} \qquad \text{and} \qquad C_p := 2^{-p} + \sum_{r=2}^{K} r^{-p}. \tag{3}$$

For each of the three algorithms, the bound on the misidentification probability can be written in the form of $\beta \exp(-T/\alpha)$, where α and β are provided in Table I ($\overline{\log}\ K = 0.5 + \sum_{i=2}^{K} i^{-1}$). It was shown in [18] that

TABLE I

THE PARAMETERS α AND β FOR THE ALGORITHMS PROPOSED FOR (SINGLE) BEST-ARM IDENTIFICATION. THE MISIDENTIFICATION PROBABILITY FOR EACH ALGORITHM DECAYS IN THE FORM OF $\beta \exp(-T/\alpha)$. THE QUANTITIES USED IN THE TABLE ARE DEFINED IN (2) AND (3).

Algorithm	Successive Rejects	Sequential Halving	Nonlinear Sequential Elimination
α	$H_2\overline{\log}K$	$8H_2 \log_2 K$	$H(p)C_p$
β	$0.5K(K-1)\exp\Big(K/(H_2\overline{\log}K)\Big)$	$3\log_2 K$	$(K-1)\exp(K/H(p)C_p)$

TABLE II

THE SAMPLING COMPLEXITY OF ALGORITHMS PROPOSED FOR M-BEST-ARM IDENTIFICATION. IT IDENTIFIES THE SMALLEST T FOR WHICH EACH ALGORITHM RECOMMENDS THE TOP M ARMS WITH PROBABILITY AT LEAST $1-\delta$. THE QUANTITIES USED IN THE TABLE ARE DEFINED IN (3) AND (4).

Algorithm	SAR	AT-LUCB	NSAR (our algorithm)
Sampling complexity order	$H_2^{\langle M\rangle}\overline{\log}K \log\frac{K}{\delta}$	$H_1^{\langle M\rangle} \log \frac{H_1^{\langle M\rangle}}{\delta}$	$H^{\langle M\rangle}(p)C_p \log\frac{K}{\delta}$

in many regimes for the arm gaps, $p \neq 1$ provides better results (theoretical and practical), and Nonlinear Sequential Elimination outperforms the other two algorithms. The value of p must be tuned, but the tuning is more qualitative rather than quantitative, i.e., the algorithm performs reasonably well as long as p is either in $(0,1)$ or $(1,2)$, and thus, the value of p needs not be specific.

In this work, our goal is to extend this idea to M-best-arm identification. For convenience, we discuss the performance of these algorithms in terms of the *sample complexity*, defined as the smallest budget T needed to achieve the confidence level δ for misidentification probability, i.e., the smallest T for which $\mathbb{P}(\{\mathbf{J}_1,\ldots,\mathbf{J}_M\} \neq [M]) \leq \delta$. For M-best-arm identification, we need to define a new set of quantities and complexity measures as

$$\begin{aligned}
\Delta_i^{\langle M\rangle} &= \begin{cases} \mu_i - \mu_{M+1}, & \text{if } i \leq M \\ \mu_M - \mu_i, & \text{otherwise} \end{cases} \\
H_1^{\langle M\rangle} &= \sum_{i=1}^{K} \left(\Delta_i^{\langle M\rangle}\right)^{-2} \\
H_2^{\langle M\rangle} &= \max_{i\neq 1}\left\{ i\left(\Delta_{(i)}^{\langle M\rangle}\right)^{-2}\right\} \\
H^{\langle M\rangle}(p) &= \max_{i\neq 1}\left\{ i^p\left(\Delta_{(i)}^{\langle M\rangle}\right)^{-2}\right\}, \qquad (4)
\end{aligned}$$

where $\Delta_{(i)}^{\langle M\rangle}$ for each $(i) \in [K]$ is such that

$$\Delta_{(1)}^{\langle M\rangle} \leq \Delta_{(2)}^{\langle M\rangle} \leq \cdots \leq \Delta_{(K)}^{\langle M\rangle}.$$

Based on the definitions above,

$$H^{\langle M\rangle}(1) = H_2^{\langle M\rangle} \neq H_1^{\langle M\rangle}.$$

Table II tabulates the sample complexities of three algorithms for M-best-arm identification: SAR [15], AT-LUCB [19], and NSAR proposed in this paper. It follows immediately from (4) that for $p \in (0,1)$, $H^{\langle M\rangle}(p) \leq H_2^{\langle M\rangle}$, and for $p \in (1,2]$, $H^{\langle M\rangle}(p) \geq H_2^{\langle M\rangle}$. Also, in view of (3), $C_p > \log K$ for $p \in (0,1)$ and $C_p < \log K$ for $p \in (1,2]$. Therefore, the comparison of $H_2^{\langle M\rangle}\overline{\log}\,K$ and $H^{\langle M\rangle}(p)C_p$, the sample complexities of SAR and NSAR, is not obvious. As in the case of single best-arm identification, we will show that in many regimes for rewards, NSAR can outperform SAR.

Note that AT-LUCB [19] is an anytime algorithm, i.e., it does not require a pre-assigned budget. In that sense, AT-LUCB is more powerful compared to algorithms designed specifically for the fixed-budget setting, but since it can also be used in this framework, we include it in the table as a benchmark and will compare our results with this algorithm in the numerical experiments.

III. Nonlinear Sequential Accepts and Rejects

In this section, we propose the Nonlinear Sequential Accepts and Rejects (NSAR) algorithm for M-best-arm identification in the fixed budget setting. The algorithm follows the steps of SAR [15], except for the fact that the budget allocation at each round is a nonlinear function of arms. The details of NSAR is given in Figure 1. The algorithm is given a budget T of arm pulls. At any round $r \in [K-1]$, it maintains an active set of arms $\mathbf{A}_r$, initialized by $\mathbf{A}_1 = [K]$. The algorithm proceeds for $K-1$ rounds to deactivate the arms sequentially (one arm at each round) until a single arm is left. Based on an input value $p \in (0,2]$, the constant C_p and the sequence $\{n_r\}_{r=1}^{K-1}$ are calculated for any $r \in [K-1]$. At round r, the algorithm samples the $K+1-r$ active arms for $n_r - n_{r-1}$ times and computes the empirical average of rewards for each arm. Then, it orders the empirical rewards and calculates the empirical version of gaps, where the true gaps $\Delta_i^{\langle M\rangle}$ for $i \in [K]$ are defined in the first line of (4). The arm with the highest empirical gap is deactivated: if its empirical reward is within the top M arms, it is accepted; otherwise, it is rejected. At the end, the algorithm outputs M accepted arms as the top M arms.

Note that our algorithm with the choice of $p = 1$ amounts to SAR. We will show 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 IV. The following proposition encapsulates the theoretical guarantee of the algorithm (the proof is given in the appendix).

```
Nonlinear Sequential Accepts and Rejects

Input: budget $T$, parameter $p > 0$.
Initialize: $\mathbf{A}_1 = [K]$, $n_0 = 0$, $\mathbf{m}_1 = M$.
Let
    $C_p = 2^{-p} + \sum_{r=2}^{K} r^{-p}$
    $n_r = \left\lceil \frac{T-K}{C_p(K-r+1)^p} \right\rceil$ for $r \in [K-1]$
At round $r = 1, \dots, K-1$:
    (1) Sample each arm in $\mathbf{A}_r$ for $n_r - n_{r-1}$ times.
    (2) Let $\sigma_r : [K+1-r] \to \mathbf{A}_r$ be a permutation that orders the empirical means such that
            $\widehat{\boldsymbol{\mu}}_{\sigma_r(1),n_r} \ge \widehat{\boldsymbol{\mu}}_{\sigma_r(2),n_r} \ge \cdots \ge \widehat{\boldsymbol{\mu}}_{\sigma_r(K+1-r),n_r}$.
        Then, for any $\ell \in [K+1-r]$, define the following empirical gaps
            $\widehat{\boldsymbol{\Delta}}_{\sigma_r(\ell),n_r} = \begin{cases} \widehat{\boldsymbol{\mu}}_{\sigma_r(\ell),n_r} - \widehat{\boldsymbol{\mu}}_{\sigma_r(\mathbf{m}_r+1),n_r}, & \text{if } \ell \le \mathbf{m}_r \\ \widehat{\boldsymbol{\mu}}_{\sigma_r(\mathbf{m}_r),n_r} - \widehat{\boldsymbol{\mu}}_{\sigma_r(\ell),n_r}, & \text{otherwise} \end{cases}$
    (3) Identify $\mathbf{index} := \mathrm{argmax}_{i \in \mathbf{A}_r} \widehat{\boldsymbol{\Delta}}_{i,n_r}$, set $\mathbf{U}_r := \{\mathbf{index}\}$ and $\mathbf{A}_{r+1} = \mathbf{A}_r \setminus \mathbf{U}_r$, i.e., discard the arm $\mathbf{index}$.
    (4) If $\widehat{\boldsymbol{\mu}}_{\mathbf{index},n_r} > \widehat{\boldsymbol{\mu}}_{\sigma_r(\mathbf{m}_r+1),n_r}$, accept the arm $\mathbf{index}$, set $\mathbf{m}_{r+1} = \mathbf{m}_r - 1$ and $\mathbf{J}_{M-\mathbf{m}_{r+1}} = \mathbf{index}$.
    (5) After finishing $r = K-1$, the survived arm is accepted, if we have accepted $M-1$ arms at the beginning of $r = K-1$; otherwise, the survived arm is rejected.
Output: $\{\mathbf{J}_1, \dots, \mathbf{J}_M\}$.
```

Fig. 1. The NSAR algorithm for identification of the best-M arms.

Proposition 2: Let the Nonlinear Sequential Accepts and Rejects algorithm in Figure 1 run for a given $p \in (0, 2]$, *and let* C_p *and* $H^{\langle M \rangle}(p)$ *be defined as in* (3) *and* (4). *Then, the misidentification probability satisfies the bound,*

$$\mathbb{P}\left(\{\mathbf{J}_1, \dots, \mathbf{J}_M\} \neq [M]\right) \le 2K^2 \exp\left(-\frac{T-K}{8C_p H^{\langle M \rangle}(p)}\right).$$

The performance of NSAR relies on the input parameter p, but this choice is more qualitative rather than quantitative. In particular, larger values for p increase $H^{\langle M \rangle}(p)$ and decrease C_p, and hence, there is a trade-off in selecting p. According to Table II, to compare NSAR with SAR and AT-LUCB , we have to evaluate the corresponding sample complexities. Fair theoretical comparisons with AT-LUCB is delicate, since $H_1^{\langle M \rangle}$ is in essence slightly different from $H_2^{\langle M \rangle}$ and $H^{\langle M \rangle}(p)$. However, we will provide comprehensive simulations in Section IV to compare all algorithms. We consider two instances for sub-optimality of arms in this section to compare NSAR with SAR:

1 **A large group of competitive arms:** The top M arms are roughly similar such that $\mu_1 \approx \mu_M$, $\mu_M - \mu_{M+1} = \delta_1$ is non-negligible, and the other arms are just as competitive as each other, i.e., $\mu_{M+1} \approx \mu_K$.

2 **A small group of competitive arms:** The top M arms are roughly similar such that $\mu_1 \approx \mu_M$. $\mu_M - \mu_{M'} = \delta_1$ for a small number of arms ($M' = \mathcal{O}(1)$ with respect to K) and $\mu_{M+1} \approx \mu_{M'}$, $\mu_{M'} - \mu_{M'+1} = \delta_2$, and $\mu_{M'+1} \approx \mu_K$. We also have $\delta_1 \ll \delta_2$.

The subsequent corollary follows from Proposition 2. Note that the orders are expressed with respect to K.

Corollary 3: Consider the Nonlinear Sequential Accepts and Rejects algorithm in Figure 1. Let constants p *and* q *be chosen such that* $1 < p \le 2$ *and* $0 < q < 1$. *Then, for the two settings given above, the bound on the misidentification probability presented in Proposition 2 satisfies*

Regime 1	**Regime 2**
$C_q H^{\langle M \rangle}(q) = \mathcal{O}(K)$	$C_p H^{\langle M \rangle}(p) = \mathcal{O}(1)$

Now let us compare NSAR and SAR using the result of Corollary 3. Returning to Table II and calculating $H_2^{\langle M \rangle}$ for Regimes 1 and 2, we can derive the following table, which shows that with a proper tuning for p, we can save a $\mathcal{O}(\log K)$ factor in the sampling complexity. Though we do not have prior information on gaps to categorize them specifically, the choice of the input parameter p is more qualitative rather than quantitative, i.e., once the sub-optimal arms are almost the same $0 < p < 1$ performs better than

TABLE III
THE SAMPLING COMPLEXITY FOR NSAR (OUR ALGORITHM) AND SAR. FOR REGIME 1, WE SET $0 < q < 1$, AND FOR REGIME 2, WE USE $1 < p \le 2$. THE ORDER DOES NOT INCLUDE THE $\log \frac{K}{\delta}$ TERM AS IT IS IN COMMON BETWEEN THE TWO ALGORITHMS.

Algorithm	SAR	NSAR
Regime 1	$\mathcal{O}(K \log K)$	$\mathcal{O}(K)$
Regime 2	$\mathcal{O}(\log K)$	$\mathcal{O}(1)$

$1 < p \leq 2$, and when there are a few real competitive arms, $1 < p \leq 2$ outperforms $0 < p < 1$. Next, we will show in the numerical experiments that a wide range of values for p can potentially result in efficient algorithms with small misidentification error.

IV. Numerical Experiments

A. Academic Example

We now empirically evaluate our proposed algorithm on a few settings studied in [15]. More specifically, we compare NSAR with SAR, AT-LUCB, as well as uniform allocation (UNI), where in the UNI algorithm, we simply divide the budget uniformly across the arms. We remark that AT-LUCB in [19] is an anytime algorithm, i.e., it does not require a pre-assigned budget; however, since it can also be used for the fixed-budget setting, we include it in our numerical experiments as a benchmark. We consider $K = 50$ arms and assume Bernoulli distribution on the rewards. For the following setups, we examine two values for top arms $M \in \{2, 4\}$ (we use the notation $x\!:\!y$ to denote integers in $[x, y]$):

1 **One group of suboptimal arms:** $\mu_{1:M} = 0.7$ and $\mu_{M+1:K} = 0.5$.
2 **Two groups of suboptimal arms:** $\mu_{1:M} = 0.7$, $\mu_{M+1:2M} = 0.66$, and $\mu_{2M+1:K} = 0.5$.
3 **Three groups of suboptimal arms:** $\mu_{1:M} = 0.7$, $\mu_{M+1:2M} = 0.66$, $\mu_{2M+1:3M} = 0.62$, and $\mu_{3M+1:K} = 0.5$.
4 **Arithmetic Progression:** $\Delta_i = \frac{0.6(i-1)}{K-1}$ for $i = 2 : K$.
5 **Beta(5,5):** The expected values of Bernoulli distributions are generated according to a beta distribution with shape parameters 5 and 5.
6 **One real competitive arm:** $\mu_{1:M} = 0.7$, $\mu_{M+1} = 0.68$ and $\mu_{M+2:K} = 0.5$.

We run 4000 experiments for each setup with a specific value of M, and we calculate the misidentification probability by averaging out over the error in experiment runs. We set the budget T in each setup equal to $\left\lceil H_1^{\langle M \rangle} \right\rceil$ in the corresponding setup as suggested in [15], and we also choose the parameters of AT-LUCB as instructed in [19].

We illustrate the overall performance of the algorithms in Figure 2 for different setups. The height of each bar shows the misidentification probability, and the index guideline is as follows: **(i)** indices 1-5: NSAR with parameter $p \in \{0.7, 0.85, 1.1, 1.2, 1.3\}$. **(ii)** index 6: SAR. **(iii)** index 7: AT-LUCB. **(iv)** index 8: UNI. The legends are the same for all of the plots, and hence, they are omitted in most of the plots.

The results are consistent with Corollary 3, and the following comments are in order:

- Setup 1 corresponds to Regime 1 in Corollary 3. As expected, with any choice of $0 < p < 1$, NSAR should outperform SAR, and we observe that this happens when $p \in \{0.7, 0.85\}$. However, in this regime, our algorithm is inferior compared to AT-LUCB.
- Setups 2-3-6 are considered close to Regime 2 in Corollary 3 as we have a small number of arms competitive to the top M arms. Thus, we should choose $1 < p \leq 2$. We observe that in these setups, at least for two choices out of $p \in \{1.1, 1.2, 1.3\}$, NSAR is as competitive as SAR and AT-LUCB (outperforming at least one of them). One should observe that the improvement in Corollary 3 is $\mathcal{O}(\log K)$ which increases slowly with K. Since we only have $K = 50$ numbers, using larger values for p is not suitable in these setups, because the increase in $H^{\langle M \rangle}(p)$ worsens the performance overall. Though for larger values of K, the improvement must be more visible, we avoid that due to prohibitive time-complexity of Monte Carlo simulations.
- In Setup 4, again our algorithm for $p \in \{1.1, 1.2, 1.3\}$ outperforms SAR and is as competitive as AT-LUCB.
- In Setup 5, we choose the expected values of Bernoulli rewards randomly and concentrate them around 0.5. For all three choices of $p \in \{1.1, 1.2, 1.3\}$, our algorithm outperforms SAR and AT-LUCB.
- In all setups, the naive UNI algorithm is outperformed by the other methods.

Overall, the performance of algorithms depends on the problem environment. If we have prior knowledge of the environment, we can select the suitable algorithm. The notable feature of NSAR is incorporation of this prior knowledge in tuning of p without changing the foundation of the algorithm.

B. Application in Cartoon Caption Contest

Every week, the New Yorker magazine holds a cartoon caption contest[1] where readers can write a funny caption for a specific cartoon. Then, the staffs search through the pool of captions (which potentially large) to find the funniest ones. These captions are rates as “not funny”, “somewhat funny”, or “funny” by a number of volunteers using the crowdsourcing system NEXT[2] [28]. If the crowdsourcing system presents the captions to voters uniformly at random (non-adaptively), vast number of samples are needed before the funny captions are captured. Thus, the system should choose the captions adaptively and strategically to present them to voters. In this context, the arms are associated to the captions, and the rewards are associated to “not funny”, “somewhat funny”, or “funny”. It is more plausible to use an anytime algorithm (e.g. AT-LUCB) for this application, but we compare our algorithm versus AT-LUCB and SAR for various number of budgets.

Let us consider the dataset #559 in the cartoon caption contest[3] with $K = 138$ captions rated 62742 times. We use a categorical distribution to translate “not funny” to 0, “somewhat funny” to 0.5, and “funny” to 1, respectively, and estimate the probability of $\{0, 0.5, 1\}$ using the passive dataset. We then construct the corresponding MAB problem with stochastic rewards according to the resulted distribution.

[1] https://contest.newyorker.com/.

[2] https://amplab.cs.berkeley.edu/the-new-yorker-uses-next-to-crowd-source-the-next-caption-contest-winner/.

[3] Dataset #559 in https://github.com/nextml/caption-contest-data.

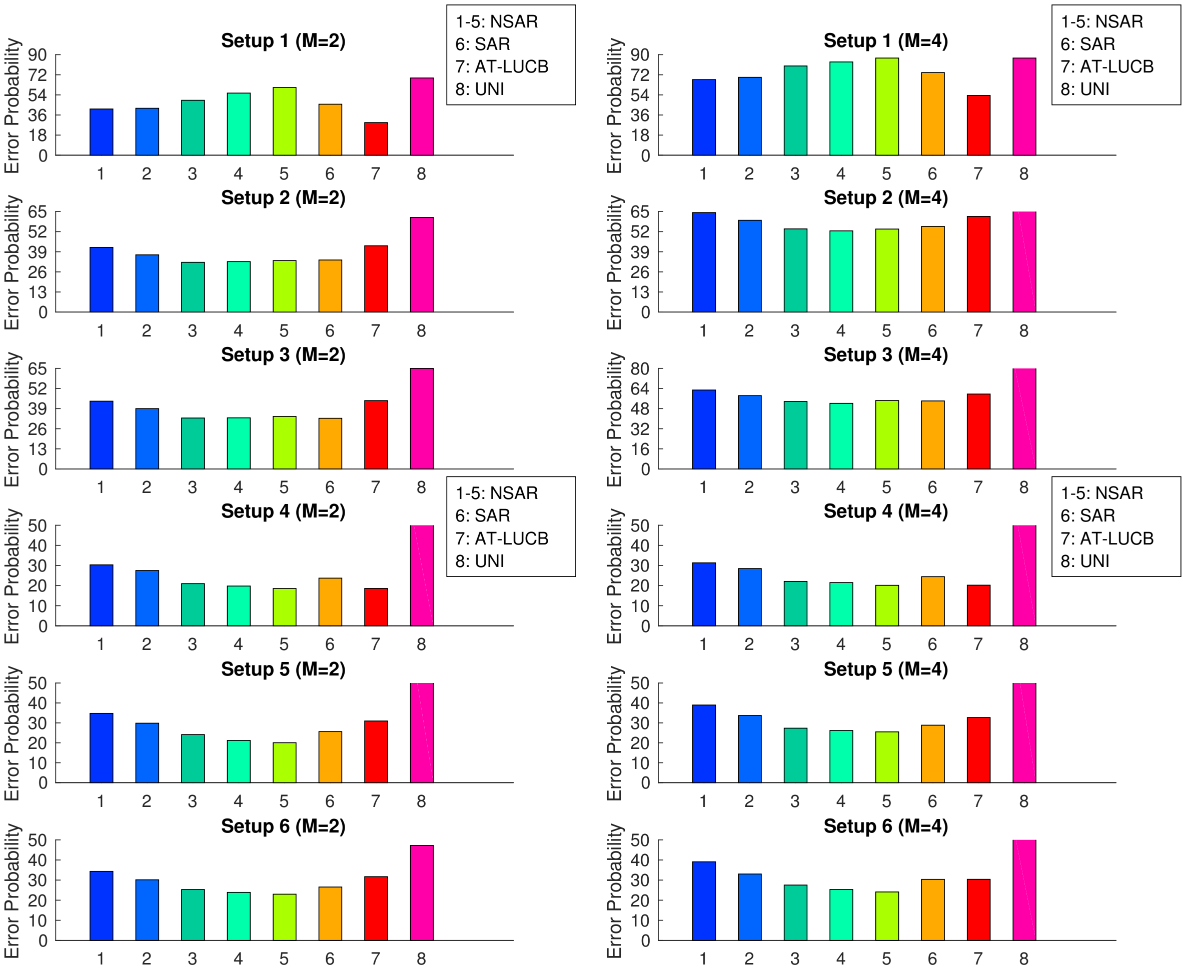

Fig. 2. The figure shows the misidentification probability for NSAR, SAR, AT−LUCB, and UNI algorithms in six different setups. The six plots on the left relate to the case $M = 2$, and the six plots on the right are associated with $M = 4$. The height of each bar represents the misidentification probability, and each index (or color) represents one algorithm tuned with a specific parameter.

Our goal is to find the top $M = 2$ captions, and we estimate the misidentification probability for each algorithm by averaging its performance over 1000 experiments. The decay of misidentification probability with the budget increase is depicted in Figure 3 for all algorithms. Since UNI is outperformed by orders of magnitude compared to all algorithms, we do not include it in the plot. We can observe that our algorithm with $p = 0.85$ outperforms both SAR and AT−LUCB, while the choice of $p = 1.1$ is not good enough to beat SAR in this environment. We finally remark that the parameters of AT−LUCB are set according to [19]; it is possible that other hyper-parameter configurations yield better results.

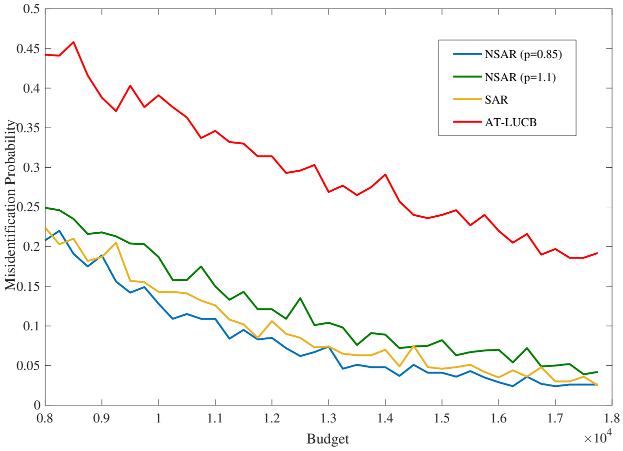

Fig. 3. Comparison of the results on cartoon caption contest data.

V. Conclusion

We considered M-best-arm identification in stochastic multi-armed bandits, where the objective is to find the top M arms in the sense of the expected reward. We presented an algorithm working based on sequential deactivation of arms in rounds. The key is to allocate the budget of arm pulls in a nonlinear fashion at each round. We proved theoretically and empirically that we can gain from the nonlinear budget allocation in several problem environments, compared to the state-of-the-art methods. An important future direction is to propose a method that adaptively fine-tunes the nonlinearity according to the problem environment.

VI. Appendix

Fact 1: (Hoeffding's inequality) *Let $W_1, \dots, W_n$ 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 Proposition 2

Recall that $\widehat{\boldsymbol{\mu}}_{i,n}$ denotes the average reward of pulling arm i for n times. Now consider the following event

$$\mathcal{E} := \left\{\forall i \in [K], \forall r \in [K-1] : \left|\widehat{\boldsymbol{\mu}}_{i,n_k} - \mu_i\right| \leq \frac{1}{4}\Delta^{\langle M\rangle}_{(K+1-r)}\right\}.$$

Using Hoeffding's inequality (Fact 1), we get

$$\begin{aligned}\mathbb{P}\left(\mathcal{E}^C\right) &\leq \sum_{i=1}^{K}\sum_{r=1}^{K-1} \mathbb{P}\left(\left|\widehat{\boldsymbol{\mu}}_{i,n_k} - \mu_i\right| > \frac{1}{4}\Delta^{\langle M\rangle}_{(K+1-r)}\right) \\ &\leq \sum_{i=1}^{K}\sum_{r=1}^{K-1} 2\exp\left(-2n_r\left(\frac{1}{4}\Delta^{\langle M\rangle}_{(K+1-r)}\right)^2\right).\end{aligned}$$

Noting the fact that $n_r = \left\lceil \frac{T-K}{C_p(K+1-r)^p} \right\rceil \geq \frac{T-K}{C_p(K+1-r)^p}$, we can use above to conclude that

$$\begin{aligned}\mathbb{P}\left(\mathcal{E}^C\right) &\leq 2K^2 \max_{r\in[K-1]}\left\{\exp\left(-\frac{T-K}{8}\frac{\left(\Delta^{\langle M\rangle}_{(K+1-r)}\right)^2}{C_p(K+1-r)^p}\right)\right\} \\ &= 2K^2\exp\left(-\frac{T-K}{8}\min_{r\in[K-1]}\left\{\frac{\left(\Delta^{\langle M\rangle}_{(K+1-r)}\right)^2}{C_p(K+1-r)^p}\right\}\right) \\ &= 2K^2\exp\left(-\frac{T-K}{8C_pH^{\langle M\rangle}(p)}\right).\end{aligned}$$

The rest of the proof is to show that the event $\mathcal{E}$ warrants that the algorithm does not make erroneous decision. This part follows precisely by the induction argument given in [15] (see page 4-5). □

Proof of Corollary 3

First, let us analyze the order of C_p defined as

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

For any $p > 1$, C_p is a convergent sum when $K \to \infty$. Thus, 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

$$H^{\langle M\rangle}(p) = \max_{i\neq 1}\left\{i^p\left(\Delta^{\langle M\rangle}_{(i)}\right)^{-2}\right\}.$$

For Regime 1, $q \in (0,1)$ and we have

$$\max_{i\neq 1}\left\{i^q\left(\Delta^{\langle M\rangle}_{(i)}\right)^{-2}\right\} \approx \frac{K^q}{\delta_1^2}$$

Combining with C_q, the product $C_qH^{\langle M\rangle}(q) = \mathcal{O}(K)$. For Regime 2, $p \in (1,2]$ and we have

$$\max_{i\neq 1}\left\{i^p\left(\Delta^{\langle M\rangle}_{(i)}\right)^{-2}\right\} \approx \frac{M'^p}{\delta_1^2} = \mathcal{O}(1),$$

since $\delta_1 \ll \delta_2$. Therefore, combining with $C_p = \mathcal{O}(1)$, the product $C_pH^{\langle M\rangle}(p) = \mathcal{O}(1)$. □

References

[1] T. L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Advances in applied mathematics*, 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," *Machine learning*, vol. 47, no. 2-3, pp. 235–256, 2002.

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

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

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

[6] S. Vakili, K. Liu, and Q. Zhao, "Deterministic sequencing of exploration and exploitation for multi-armed bandit problems," *IEEE Journal of Selected Topics in Signal Processing*, vol. 7, no. 5, pp. 759–767, 2013.

[7] D. Kalathil, N. Nayyar, and R. Jain, "Decentralized learning for multiplayer multiarmed bandits," *IEEE Transactions on Information Theory*, vol. 60, no. 4, pp. 2331–2345, 2014.

[8] J.-Y. Audibert and S. Bubeck, "Best arm identification in multi-armed bandits," in *COLT-23th Conference on Learning Theory-2010*, 2010, pp. 13–p.

[9] E. Even-Dar, S. Mannor, and Y. Mansour, "PAC bounds for multi-armed bandit and markov decision processes," in *Computational Learning Theory*. Springer, 2002, pp. 255–270.

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

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

[12] ——, "Pure exploration in finitely-armed and continuous-armed bandits," *Theoretical Computer Science*, vol. 412, no. 19, pp. 1832–1852, 2011.

[13] Z. Karnin, T. Koren, and O. Somekh, "Almost optimal exploration in multi-armed bandits," in *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, 2013, pp. 1238–1246.

[14] V. Gabillon, M. Ghavamzadeh, and A. Lazaric, "Best arm identification: A unified approach to fixed budget and fixed confidence," in *Advances in Neural Information Processing Systems*, 2012, pp. 3212–3220.

[15] S. Bubeck, T. Wang, and N. Viswanathan, "Multiple identifications in multi-armed bandits," in *Proceedings of The 30th International Conference on Machine Learning (ICML)*, 2013, pp. 258–265.

[16] S. Kalyanakrishnan and P. Stone, "Efficient selection of multiple bandit arms: Theory and practice," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 2010, pp. 511–518.

[17] S. Kalyanakrishnan, A. Tewari, P. Auer, and P. Stone, "PAC subset selection in stochastic multi-armed bandits," in *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, 2012, pp. 655–662.

[18] S. Shahrampour, M. Noshad, and V. Tarokh, "On sequential elimination algorithms for best-arm identification in multi-armed bandits," *IEEE Transactions on Signal Processing*, vol. 65, no. 16, pp. 4281–4292, Aug 2017.

[19] K.-S. Jun and R. D. Nowak, "Anytime exploration for multi-armed bandits using confidence information." in *International Conference on Machine Learning (ICML)*, 2016, pp. 974–982.

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

[21] K. Jamieson, M. Malloy, R. Nowak, and S. Bubeck, “lil’ucb: An optimal exploration algorithm for multi-armed bandits,” in *Proceedings of The 27th Conference on Learning Theory*, 2014, pp. 423–439.

[22] L. Chen, J. Li, and M. Qiao, “Nearly instance optimal sample complexity bounds for top-k arm selection,” *arXiv preprint arXiv:1702.03605*, 2017.

[23] H. Jiang, J. Li, and M. Qiao, “Practical algorithms for best-k identification in multi-armed bandits,” *arXiv preprint arXiv:1705.06894*, 2017.

[24] J. Chen, X. Chen, Q. Zhang, and Y. Zhou, “Adaptive multiple-arm identification,” *arXiv preprint arXiv:1706.01026*, 2017.

[25] Y. Zhou, X. Chen, and J. Li, “Optimal PAC multiple arm identification with applications to crowdsourcing,” in *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, 2014, pp. 217–225.

[26] E. Kaufmann and S. Kalyanakrishnan, “Information complexity in bandit subset selection,” in *Conference on Learning 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,” *Journal of Machine Learning Research*, vol. 17, no. 1, pp. 1–42, 2016.

[28] K. G. Jamieson, L. Jain, C. Fernandez, N. J. Glattard, and R. Nowak, “Next: A system for real-world development, evaluation, and application of active learning,” in *Advances in Neural Information Processing Systems*, 2015, pp. 2656–2664.