ECAI 2020
G.D. Giacomo et al. (Eds.)

doi:10.3233/FAIA200394

Rapidly Finding the Best Arm Using Variance

Marco Faella[1] and **Alberto Finzi**[2] and **Luigi Sauro**[3]

Abstract. We address the problem of identifying the best arm in a pure-exploration multi-armed bandit problem. In this setting, the agent repeatedly pulls arms in order to identify the one associated with the maximum expected reward. We focus on the fixed-budget version of the problem in which the agent tries to find the best arm given a fixed number of arm pulls. We propose a novel sequential elimination method exploiting the empirical variance of the arms. We detail and analyse the overall approach providing theoretical and empirical results. The experimental evaluation shows the advantage of our variance-based rejection method in heterogeneous test settings, considering both identification accuracy and execution time.

1 Introduction

In this paper, we consider the problem of selecting the best expected value among a finite set of random variables (RVs), assuming unknown distributions. Such a problem can be formulated as a stochastic best arm identification problem in multi-arm bandits (MAB). In this setting, an agent (or *forecaster*) repeatedly chooses an action (or arm) and observes a reward, drawn from an unknown, but fixed probability distribution associated with each arm. The aim of the forecaster is to identify the arm maximizing the expected reward, based on the observations. Notice that the classical problem in MAB is to maximize the *cumulative* reward by effectively balancing the trade-off between the exploration and the exploitation of the arms [18, 13, 3]. We focus instead on the pure exploration version of the problem [4, 1, 8, 5, 19], in which the forecaster ultimately outputs a recommended arm, and the objective function is the expected reward of that arm.

The best arm identification problem is relevant for various applications. For instance, in channel allocation for mobile phones, an exploration period before the communication start is needed to identify the best channel to operate [1]. Another example is provided by preference elicitation applications, where a system is tasked with discovering the preferences of a user. Efficient elicitation is based on identifying the query (i.e., the arm) with the highest expected value of information, a pure exploration problem [17, 14]. Best arm identification problems are also relevant to sequential decision making under uncertainty to support policy selection methods [12, 6].

Since we are interested in rapidly estimating the best arm given limited resources, in this work we focus on the *fixed-budget best arm identification* problem (FBBAI) [1, 5, 19], in which the agent is given a maximum number of arm pulls to find the best arm.

The main contribution of this paper is a novel sequential elimination approach to FBBAI, called *variance-based rejects* (VBR) algorithm, that exploits variance estimation for pull allocation and arm rejection. Sequential elimination approaches to this problem have been proposed in the literature [1, 11, 19], but these methods base their decisions only on the estimated mean values. Moreover, their effectiveness has been assessed, both theoretically and empirically, considering probability distributions in which the variance is either limited due to a bounded support (e.g. Rademacher) or univocally determined by the mean value (e.g. Bernoulli). In contrast, when the arms are associated with more general distributions, higher central moments provide further information on the shape of the unknown distributions and may hence improve the pull attribution strategy and the accuracy of the final arm selection.

In the literature, the only variance-based approach to the fixed-budget best arm identification problem is proposed by [9] in a multi-bandit scenario. However, that method does not leverage incremental rejection and is limited by design to bounded distributions. Similar issues can be found in other methods, where empirical Bernstein bounds are exploited to address different, but related problems, like fixed-confidence multi-arm bandit [2] or single-arm stopping [16]. An incremental rejection algorithm based on confidence bounds is proposed by [15] in the context of the model selection problem; however, those bounds are computed in a data-independent way based on Hoeffding's inequality, and assuming a known and bounded range of possible values.

In this paper, we tackle these limitations by proposing the VBR algorithm, in which both the empirical means and variances are exploited by a sequential elimination method to rapidly find the best arm, assuming arbitrary distributions. Following this approach, the initial arms are incrementally pruned until only one is left or the overall budget is consumed. At each filtering cycle, a budget is allocated to refine the empirical mean and variance of the remaing arms; the arms whose upper confidence bound is lower than the current maximal lower bound are dismissed. We provide both theoretical and empirical evidence about the effectiveness of our approach.

From a theoretical point of view, we introduce the problem complexity measure H_σ, which refines the measure H_2 proposed in [1]. We then prove an upper bound for the accuracy (that is, probability of misidentification) of VBR in terms of H_σ. Up to our knowledge, this is the first upper bound for a class of unbounded distributions, under the rather general assumption that the distribution is sub-Gaussian. Interestingly, the experiments also show that H_σ is strongly correlated with the actual accuracy of both VBR and its competitors.

Finally, we provide an experimental evaluation that compares VBR with respect to the main FBBAI algorithms proposed in the literature. In particular, we assess the algorithms' accuracy in heterogeneous experimental settings, considering various distribution types, both discrete and continuous, and different parameters. To support on-line applications, where actual execution time is critical, we also measure the algorithms' running times for a given budget, observing substantial performance differences between different algorithms.

[1] University of Naples Federico II, Italy, email: m.faella@unina.it
[2] University of Naples Federico II, Italy, email: alberto.finzi@unina.it
[3] University of Naples Federico II, Italy, email: luigi.sauro@unina.it

The collected results show the advantage of our variance-based approach in terms of both time performance and accuracy, on a wide variety of input distributions. More specifically, we show that VBR dominates the other methods in most of the proposed settings, while providing competitive results in the remaining test cases. We also show that the only other variance-based method proposed in the literature [9], when comparable, is dominated in accuracy and significantly slower than VBR.

2 Preliminaries

Problem setup

Consider a set of K arms, enumerated by $[K] = \{1, \dots, K\}$. Each arm $i \in [K]$ is associated with a reward which is a random variable X_i with expectation $\mu_i = \mathbb{E}[X_i]$ and variance $\sigma_i^2 = \mathbb{E}[(X_i - \mu_i)^2]$. Initially, a forecaster ignores the probability distribution associated to each arm, however she can iteratively choose an arm and observe an independent sample of its reward. After a fixed number T of rounds, the forecaster is supposed to return the arm with maximal expectation. The *fixed-budget best arm identification* problem (FBBAI) concerns the design of an allocation strategy minimizing the probability of misidentification.

For convenience we assume that arms are ordered by their expected values and that there is only one optimal arm, i.e. $\mu_1 > \mu_2 \geq \cdots \geq \mu_K$. By $\Delta_i = \mu_1 - \mu_i$ we denote the sub-optimality gap of arm $i = 2, \dots, K$. In particular, we set $\Delta = \Delta_2$ as the minimal gap.

Assume that at round $t = 1, \dots, T$ an arm i has been chosen $s \leq t$ times and that $x_{i,1}, \dots, x_{i,s}$ are the observed rewards. Then, $\hat{\mu}_{i,s} = \frac{1}{s} \sum_{j=1}^{s} x_{i,j}$ is the empirical mean of arm i observed after s pulls and

$$\hat{\sigma}_{i,s} = \frac{1}{s-1} \sqrt{\sum_{j=1}^{s} (x_{i,j} - \hat{\mu}_{i,s})^2}$$

is the corresponding empirical standard deviation.

Due to the central limit theorem (CLT), as the number of samples s tends to infinity, the random variable (RV) $\hat{\mu}_{i,s}$ tends to a normal distribution $\mathcal{N}(\mu_i, \epsilon_{i,s}^2)$, where $\epsilon_{i,s} = \frac{\sigma_i}{\sqrt{s}}$. Then, given a real $\gamma > 0$, we define the confidence upper and lower bounds $UB_{i,s}[\gamma]$ and $LB_{i,s}[\gamma]$ as:

$$UB_{i,s}[\gamma] = \hat{\mu}_{i,s} + \gamma\, \hat{\epsilon}_{i,s} \quad \text{and} \quad LB_{i,s}[\gamma] = \hat{\mu}_{i,s} - \gamma\, \hat{\epsilon}_{i,s}\,.$$

where $\hat{\epsilon}_{i,s} = \frac{\hat{\sigma}_{i,s}}{\sqrt{s}}$ is the estimated standard error. For $\gamma = 1.96$, $UB_{i,s}[\gamma]$ is called the upper 95% confidence limit and, symmetrically, $LB_{i,s}[\gamma]$ is the lower 95% confidence limit. Intuitively, due to CLT, the probability that μ_i is included in the interval from $LB_{i,s}[1.96]$ to $UB_{i,s}[1.96]$ is about 0.95. Thereafter, when clear from the context, we omit the subscript s and implicitly consider all the pulls of a given arm up to a certain round.

In previous works [1, 11], the hardness of a best arm identification problem as been measured through the value

$$H_2 = \max_{i \in \{2,\dots,K\}} \frac{i}{\Delta_i^2}\,.$$

Here, we introduce the refined measure

$$H_\sigma = \max_{i \in \{2,\dots,K\}} \frac{\sigma_1^2 + \sigma_i^2}{\Delta_i^2}$$

that explicitly depends also on the variance of the arms. In the next section, we show that the probability of misidentification of our approach is bounded above by a function of H_σ. Moreover, the experiments reported later in the paper show that H_σ can be more accurate than H_2 in predicting the hardness of a problem instance, for both our approach and its competitors.

Previous algorithms

Here, we briefly describe the main algorithms for FBBAI proposed in the literature.

SR: Successive rejects [1]. In SR, the initial budget is split in $K - 1$ arm elimination phases according to the following definition. First, let $\overline{\log}(K) = \frac{1}{2} + \sum_{i=2}^{K} \frac{1}{i}$, let $n_0 = 0$, and

$$n_j = \left\lceil \frac{1}{\overline{\log}(K)} \frac{T - K}{K + 1 - j} \right\rceil\,, \tag{1}$$

with $1 \leq j \leq K - 1$. In each phase $l \in \{0, \dots, K - 2\}$, all the surviving arms are pulled $n_{l+1} - n_l$ times each and the corresponding empirical means are updated accordingly. Then, the arm whose mean is minimal is rejected. After $K - 1$ phases the only surviving arm is returned.

SH: Sequential halving [11]. Analogously to SR, SH progressively rejects the candidate arms until a single one is left. The initial budget is split evenly across $\lceil \log_2 T \rceil$ phases. The budget for each phase is uniformly distributed over the remaining arms and the empirical mean values are updated. At the end of a phase, the worst *half* of the arms (in terms of empirical mean) are ruled out.

UCBE: Adaptive upper confidence bound exploration [1]. The initial budget is split in $K - 1$ phases as in SR. At the beginning of each phase l, the empirical gaps $\hat{\Delta}_i = (\max_{1 \leq j \leq K} \hat{\mu}_j) - \hat{\mu}_i$ are computed and $\hat{H}_{1,l}$ is set to the empirical value of H_2 among the worst l arms:

$$\hat{H}_{1,l} = \max_{K-l+1 \leq i \leq K} \frac{i}{\hat{\Delta}^2_{\langle i \rangle}}\,,$$

where $\langle i \rangle$ is an ordering of the arms such that $\hat{\Delta}_{\langle 2 \rangle} \leq \cdots \leq \hat{\Delta}_{\langle K \rangle}$. Then, at each round t of phase l, UCBE pulls the arm i with highest upper confidence bound

$$\hat{\mu}_i + \sqrt{\frac{T}{\hat{H}_{1,l} \cdot s_i(t-1)}}\,,$$

where $s_i(t-1)$ is the number of samples arm i has accumulated up to round $t - 1$.

GapEV: Gap-based exploration with variance [9]. This is the only algorithm that exploits empirical variances to distribute pulls[4]. At each round, the algorithm pulls the arm i that maximizes the quantity

$$-\hat{\Delta}_i + \sqrt{\frac{2a\hat{\sigma}_i^2}{s_i(t-1)}} + \frac{7ab}{3(s_i(t-1) - 1)}\,, \tag{2}$$

where $\hat{\sigma}_i^2$ is the empirical variance of arm i, a is an exploration parameter, and b is an upper bound to the value of the rewards. Hence, GapEV is designed to work on RVs with bounded and known support, whereas our proposal obviates both assumptions.

[4] GapE-V is closely related to the algorithm UCB-V [2]. We have focused on GapE-V because it has been specifically designed for the best arm identification problem, whereas UCB-V refers to the classical multi-armed bandit problem, where exploitation may interfere with mere identification.

Notice that we are presenting GapEV in a single-bandit context, whereas it was originally aimed at the more general multi-bandit setting, where the objective is to identify the best arm in each bandit.

Roughly speaking, SR and SH can be considered two extremes in a spectrum: SR cautiously rejects arms one by one, whereas the number of candidate arms decay exponentially in SH. However, both SR and SH are inflexible algorithms in that the number of arms dismissed at each phase is predetermined and does not depend on the observations.

UCBE and GapEV are more flexible because they select which arm has to be pulled one round at a time. On the other hand, the fact that the budget is carefully distributed at round scale has a notable impact on the execution time and makes these methods less appropriate to on-line applications (see also Section 4).

When selecting the next arm to pull, UCBE favours those arms with greater empirical mean, and the arms that have received fewer pulls so far (an exploration term). In GapEV, the exploration term depends on the empirical variance of each arm: arms having exhibited a larger variance are more likely to be pulled, as they require more data to be accurately assessed.

Our main idea is then to design an allocation strategy that is as fast as SR and SH, and yet it can flexibly modulate its aggressiveness, i.e. how many candidates to reject at the end of each phase, based on the empirical estimation of means and variances.

3 Variance-Based Rejects Algorithm

In this section we describe the VBR algorithm (Algorithm 1). The algorithm proceeds in at most $K-1$ phases which iteratively prune the initial set of arms until only one is left and returned.

The variables Q and budget hold the set of surviving arms and the residual budget, respectively. Initially, all the arms are possible candidates (line 1) and all the input budget is available (line 2). The variable elim maintains the total number of eliminated arms up to the previous phase whereas $\mathrm{arm\_bdg}$ holds the per-arm budget for the current phase. For the first phase, the latter is initialized to $n_1 - n_0$, where n_j is defined as in SR (see (1)).

During a phase each surviving arm i is sampled $\mathrm{arm\_bdg}$ times and the corresponding empirical mean $\hat{\mu}_i$ and standard deviation $\hat{\sigma}_i$ are updated (lines 6-9). Then, all the arms i whose upper bound $UB_i[\gamma]$ is lower than the currently maximal lower bound MaxLB are dismissed (lines 10–11, and 15). If no arm satisfies the previous condition, then the same policy as SR is applied. That is, the arm with the lowest empirical mean is dismissed (lines 12–15).

Next, if there is more than one surviving arm, then the set of candidates, the per-arm budget and the number of eliminated arms are updated for the next phase (lines 18–20). According to line 19, it is straightforward to see that if VBR rejects k arms during a phase, then in the next phase it will consume the budget that SR would consume to reject the same number of arms.

Otherwise, a special catch-up phase uses the residual budget to improve the estimates and perform the final selection (lines 21–27). Indeed, differently from SR and SH, where the number of arms that are dismissed during a phase is predetermined, in VBR this number depends on the stochastic values $UB_i[\gamma]$ and $LB_i[\gamma]$ and hence it cannot be known beforehand. Consequently, a phase might end up with only one arm left without having used the entire input budget. If so, in order not to waste budget, the last dismissed arms are recovered and the remaining budget is equally distributed. Finally, the arm with maximal empirical mean is returned.

input : A set $[K]$ of arms, a budget $T \in \mathbb{N}$, a confidence parameter $\gamma > 0$.
output: An arm in $[K]$.

1 $Q \leftarrow [K]$
2 $\mathrm{budget} \leftarrow T$
3 $\mathrm{elim} \leftarrow 0$
4 $\mathrm{arm\_bdg} \leftarrow n_1 - n_0$
5 **while** $|Q| > 1$ **do**
6 **foreach** $i \in Q$ **do**
7 sample $\mathrm{arm\_bdg}$ times arm i
8 update $\hat{\mu}_i$ and $\hat{\sigma}_i$
9 **end**
10 $\mathrm{MaxLB} \leftarrow \max_{i \in Q} LB_i[\gamma]$
11 $\mathrm{Reject} \leftarrow \{i \in Q \mid UB_i[\gamma] < \mathrm{MaxLB}\}$
12 **if** $\mathrm{Reject} = \emptyset$ **then** `// mimic SR`
13 $\mathrm{Reject} \leftarrow \{\mathrm{argmin}_{i \in Q} \hat{\mu}_i\}$
14 **end**
15 $Q' \leftarrow Q \setminus \mathrm{Reject}$
16 $\mathrm{budget} \leftarrow \mathrm{budget} - |Q| \cdot \mathrm{arm\_bdg}$
17 **if** $|Q'| > 1$ **then**
18 $Q \leftarrow Q'$
19 $\mathrm{arm\_bdg} \leftarrow \frac{1}{|Q|} \sum_{j=\mathrm{elim}+1}^{K-|Q|} (n_{j+1} - n_j)(K-j)$
20 $\mathrm{elim} \leftarrow K - |Q|$
21 **else** `// catch-up phase`
22 **foreach** $i \in Q$ **do**
23 sample $\left\lfloor \frac{\mathrm{budget}}{|Q|} \right\rfloor$ times arm i
24 update $\hat{\mu}_i$
25 **end**
26 $i_M \leftarrow \mathrm{argmax}_{i \in Q} \hat{\mu}_i$
27 $Q \leftarrow \{i_M\}$
28 **end**
29 **end**
30 **return** *i s.t.* $Q = \{i\}$

Algorithm 1: VBR: Variance-Based Rejects algorithm.

We now provide an upper bound to the probability of misidentification (that is, the probability that VBR returns an arm different from 1) under the hypothesis that each arm i is associated to a sub-Gaussian random variable X_i with parameter σ_i. Recall that a RV X with expected value μ is said to be sub-Gaussian with parameter $\lambda > 0$ if for all $t \in \mathbb{R}$ it holds that

$$\mathbb{E}[e^{t \cdot (X-\mu)}] \leq e^{-\frac{\lambda^2 t^2}{2}} .$$

Clearly, if a random variable is sub-Gaussian with parameter λ, then it is also sub-Gaussian with any positive parameter λ' smaller than λ.

The following result proves that the probability of misidentification decreases exponentially with the budget and is connected to the shape of the RVs X_i via the measure H_σ.

Theorem 1 *Let $[K]$ be a set of arms where each RV X_i is sub-Gaussian with parameter σ_i. The probability of misidentification of VBR, denoted by $Pr(err)$, is at most*

$$\psi(K) \cdot \exp\left(-\frac{\phi(T,K)}{2H_\sigma}\right) ,$$

where:

$$\psi(K) = \frac{(K-1)(K+2)}{2} \;\; \textit{and} \;\; \phi(T,K) = \frac{T-K}{\overline{\log}(K) \cdot (K+1)} .$$

Proof. Let $[K]$ be a set of sub-Gaussian arms and assume that, for some budget T and confidence parameter γ, VBR takes m phases to select an arm as output. Moreover, let $Q(r)$ be the set of arms that are still viable candidates at the beginning of phase $r \leq m$. The probability of misidentification $Pr(err)$ is the probability that the true best arm 1 is dismissed either at some phase $r < m$ (we denote this event by err_r) or in the last catch-up phase m, denoted by err_c:

$$Pr(err) = Pr\left(\bigcup_{r=1}^{m-1} err_r \cup err_c\right). \tag{3}$$

Then, by a union bound, we have that

$$Pr(err) \leq \sum_{r=1}^{m-1} Pr(err_r) + Pr(err_c). \tag{4}$$

The event err_r occurs when $1 \in Q(r)$ and one of the following conditions holds: either $UB_1[\gamma] < \text{MaxLB}$ (i.e., $UB_1[\gamma] < LB_i[\gamma]$, for some $i \in Q(r)$) or Reject is empty at line 12 and $\hat{\mu}_1$ is smaller than all the other means $\hat{\mu}_j$ with $j \in Q(r) \setminus \{1\}$. Since we are trying to bound the probability of err_r, we can assume that the conditions $1 \in Q(r)$ and $\text{Reject} = \emptyset$ have probability equal to 1 and hence we neglect these conjuncts. Formally, we have that

$$\begin{aligned} Pr(err_r) &\leq Pr(err_r^1) + Pr(err_r^2) \\ err_r^1 &= \bigcup_{i \in Q(r)} (UB_1[\gamma] < LB_i[\gamma]) \\ err_r^2 &= \bigcap_{i \in Q(r)\setminus\{1\}} (\hat{\mu}_1 < \hat{\mu}_i). \end{aligned}$$

Then, by a union bound, $Pr(err_r)$ is at most

$$\sum_{i \in Q(r)} Pr\left(UB_1[\gamma] < LB_i[\gamma]\right) + Pr\left(\hat{\mu}_1 < \hat{\mu}_k\right), \tag{5}$$

where k is any arm in $Q(r) \setminus \{1\}$.

Let s_r be the number of pulls performed on each arm in $Q(r)$ up to and including phase r. Notice that s_r is greater than or equal to n_r as defined in (1). Moreover, being X_k and X_1 sub-Gaussians with parameters σ_k and σ_1, respectively, the RV $\hat{\mu}_k - \hat{\mu}_1$ is a sub-Gaussian with expectation $-\Delta_k$ and parameter $\tau_k = \sqrt{\frac{\sigma_1^2+\sigma_k^2}{s_r}}$. By applying the concentration inequality of sub-Gaussian distributions [20], we have that

$$\begin{aligned} Pr\left(\hat{\mu}_1 < \hat{\mu}_k\right) &= Pr\left(\hat{\mu}_k - \hat{\mu}_1 + \Delta_k > \Delta_k\right) \\ &\leq \exp\left(-\frac{\Delta_k^2}{2\tau_k^2}\right). \end{aligned}$$

Then, we have that

$$\frac{\Delta_k^2}{2\tau_k^2} = \frac{\Delta_k^2 \cdot s_r}{2(\sigma_1^2+\sigma_k^2)} \geq \frac{\Delta_k^2 \cdot n_r}{2(\sigma_1^2+\sigma_k^2)} \geq \frac{\phi(T,K)}{2H_\sigma \cdot}, \tag{6}$$

where

$$\phi(T,K) = \frac{T-K}{\overline{\log}(K) \cdot (K+1)}.$$

Consequently, we obtain

$$Pr\left(\hat{\mu}_1 < \hat{\mu}_k\right) \leq \exp\Big(-\frac{\phi(T,K)}{2H_\sigma}\Big).$$

Similarly, $Pr\left(UB_1[\gamma] < LB_i[\gamma]\right)$ is equal to

$$Pr\left(\hat{\mu}_i - \hat{\mu}_1 + \Delta_i > \gamma \cdot (\hat{\epsilon}_1 + \hat{\epsilon}_i) + \Delta_i\right).$$

By using again the concentration inequality of sub-Gaussian distributions, we have that

$$\begin{aligned} Pr\left(UB_1[\gamma] < LB_i[\gamma]\right) &\leq \exp\Big(-\frac{(\gamma \cdot (\hat{\epsilon}_1 + \hat{\epsilon}_i) + \Delta_i)^2}{2\tau_i^2}\Big) \\ &\leq \exp\Big(-\frac{\Delta_i^2}{2\tau_i^2}\Big) \\ &\leq \exp\Big(-\frac{\phi(T,K)}{2H_\sigma}\Big) \text{ by (6)}. \end{aligned}$$

We can use the above bounds and the fact that $|Q(r)| \leq K - r$ to obtain that (5) is smaller than $(K-r+1) \cdot \exp\big(-\frac{\phi(T,K)}{2H_\sigma}\big)$.

Finally, since $m \leq K-1$, we have that

$$\sum_{r=1}^{m-1} Pr(err_r) \leq \frac{K \cdot (K-1)}{2} \cdot \exp\Big(-\frac{\phi(T,K)}{2H_\sigma}\Big). \tag{7}$$

The event err_c that arm 1 is dismissed in the catch-up phase m occurs when $1 \in Q(m)$ and $\hat{\mu}_1 < \hat{\mu}_i$, for some $i \in Q(m) \setminus \{1\}$, as we omit the first condition and obtain

$$\begin{aligned} Pr(err_c) &\leq Pr\left(\bigcup_{i \in Q(m)\setminus\{1\}} \hat{\mu}_1 < \hat{\mu}_i\right) \\ &\leq \sum_{i \in Q(m)\setminus\{1\}} Pr(\hat{\mu}_1 < \hat{\mu}_i) \\ &\leq (K-1) \cdot \exp\Big(-\frac{\phi(T,K)}{2H_\sigma}\Big). \end{aligned} \tag{8}$$

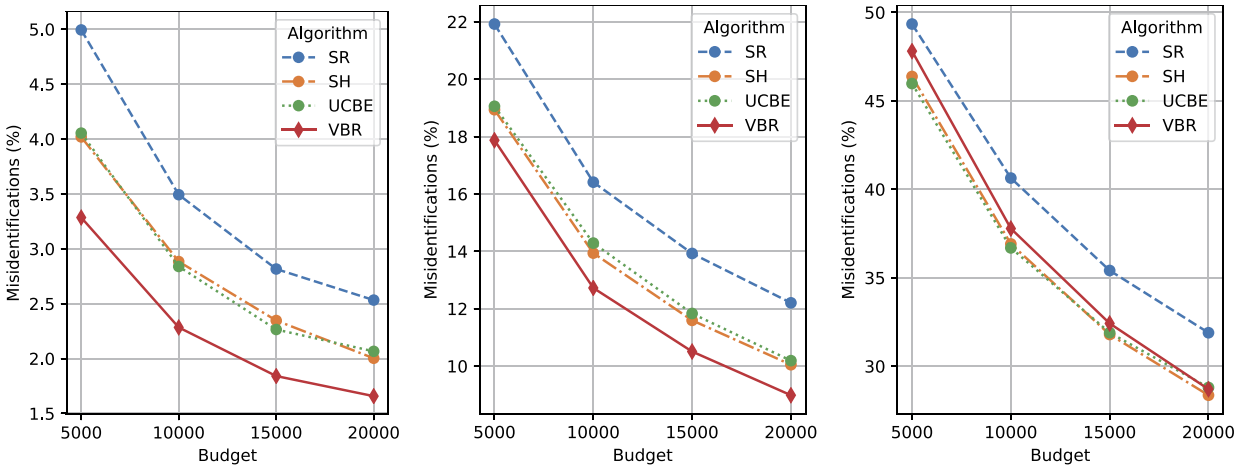
The thesis follows from equations (4), (7), and (8). □

The upper bound in Theorem 1 shares some evident similarities with the upper bounds of competing algorithms. In all approaches the probability of error exponentially decays with the budget T. Regarding the multiplicative term $\psi(K)$ in Theorem 1, the analogous term in SR is quadratic in K as in our bound, whereas in SH that term is only logarithmic in K. In GapEV, the multiplicative term is proportional to the product $K \cdot T$ and hence, since T is typically much greater than K, our approach shows a better behaviour in this respect.

Nevertheless, it is worth to note that the upper bound provided in Theorem 1 is more general than the analogous results shown for its competitors. More specifically, the other upper bounds assume that the RVs are bounded in some interval $[a,b]$. Bounded RVs are just a special case of sub-Gaussian RVs with parameter σ, which clearly include also unbounded distributions (first and foremost Gaussian distributions). Moreover, our bound depends on the variances via the problem complexity measure H_σ, whereas previous bounds generally use a variance-independent measure H_2.[5] The experiments in the next section will show that H_σ is significantly more accurate than H_2 in predicting the misidentification rate in case of unbounded RVs (see Figure 4).

In [1] the authors provide a lower bound for the best arm identification problem, but it is limited to Bernoulli RVs only. Again, since Bernoulli RVs are bounded, this result depends on the complexity measure H_2. Other than that, the gap between our upper bound and such lower bound essentially derives from the multiplicative factor $\psi(K)$, which is replaced by a constant in the lower bound. Extending the lower bound to sub-Gaussian distributions is left to future works.

[5] Only GapEV is equipped with a complexity measure that takes variances into account, but this measure is tailored to RVs with bounded and known support.



(a) *Normal-S* (b) *Normal-M* (c) *Normal-L*

Figure 1: Percentage of misidentification on normal distributions and $K = 40$.

4 Experiments

In this section, we report on a set of experiments based on a prototype implementation of Algorithm 1 developed in Java. We compare its performance and accuracy with the algorithms presented in Section 2. Summarizing, we compare the following algorithms: Successive rejects (**SR**), Sequential halving (**SH**) Adaptive upper confidence bound exploration (**UCBE**), Adaptive gap-based exploration with variance (**GapEV**), and Variance-based rejects (**VBR**).

As a quick preview, the experiments reveal that VBR frequently outperforms all other algorithms on a wide range of input scenarios, both in error rate and in execution time.

All experiments were run on an AMD Ryzen 2700X clocked at 3.7Ghz.

Input distributions. A *problem instance* is defined by K RVs $X_1, \ldots, X_K$ and a budget $T > 0$. As distributions for the RVs, we consider the following kinds, identified by a label:

***Normal-X*:** Normal distributions with expected values μ_i uniformly distributed in $[0, 1]$ and standard deviations σ_i distributed as follows:

X=S. Uniformly in $[0.01, 0.1]$

X=M. Uniformly in $[0.1, 0.5]$

X=L. Uniformly in $[0.5, 2]$

***Normal-H2*:** Normal distributions with $\mu_0 = 1$ and $\mu_i = 0$ for all $i \neq 0$. Standard deviations σ_i distributed uniformly in $[2, 10]$ (including σ_0). Notice that for this family of inputs it holds $H_2 = K$.

***Rademacher*:** Rademacher distributions with parameters x, y uniformly distributed in $[0, 1]$. This is a discrete distribution with two equally likely outcomes x and y.

***Bernoulli*:** Bernoulli distributions with parameter p uniformly distributed in $[0, 1]$.

Parameters. In all experiments the following parameters are fixed:

- VBR confidence coefficient γ. Heuristically fixed to 2 (more information below).
- UCBE exploration rate c. Fixed to 1, since this is the best setting in the experiments of [1].
- GapEV exploration parameter a. It adapts dynamically during the execution according to the estimation algorithm discussed in [9].
- Each experiment is run on 100k problem instances.

We let the following parameters vary in the experiments:

- Budget T. Varying in $\{5k, 10k, 15k, 20k\}$.
- Number of arms K. In most experiments we set $K = 40$, to obtain significant misidentification rates. In one experiment we let K vary in $\{40, 80, 160, 320\}$.

Percentage of misidentification

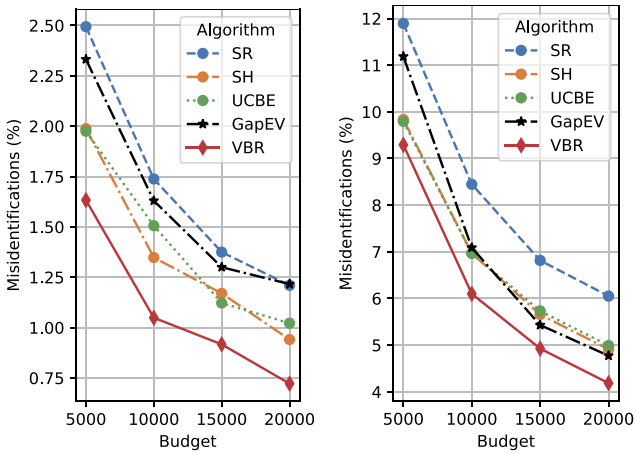
As customary in the fixed-budget setting, we measure the percentage of inputs on which a given algorithm fails to identify the true best arm. Figure 1 reports the misidentification percentages on experiments *Normal-S*, *Normal-M*, and *Normal-L*. Data shows that VBR outperforms all other algorithms on small and medium variances, whereas for large variances it sits in the middle of the group.

Figure 2 reports the misidentification percentages for the discrete distributions *Rademacher* and *Bernoulli*. Once again, VBR achieves lower error rate than the previous algorithms across the range of budgets.

Table 1 summarizes the average percentage of misidentification across all classes of inputs and all budgets, normalized w.r.t. VBR. On 4 classes, VBR displays the lowest error rate. Compared to the second-best algorithm, VBR performs from 8% to 29% better in these scenarios. On the class *Normal-L*, where the standard deviation can be 80 times the expected value of Δ, VBR is outperformed by UCBE and SH by a 3% margin. As for the other algorithms, UCBE performs very similarly to SH, but the latter can boast a faster execution time (see next section).

Execution time

The fixed-budget model is a useful abstraction for comparing the performance of different algorithms on equal footing. In some multi-armed bandit applications, requesting a sample is the most expensive operation that the estimation process performs. In other applications, such as preference elicitation, requesting a sample is a mildly expensive operation that may involve querying an MCMC sampler [14].



(a) *Rademacher* (b) *Bernoulli*

Figure 2: Percentage of misidentification on *Rademacher* and *Bernoulli* with $K = 40$.

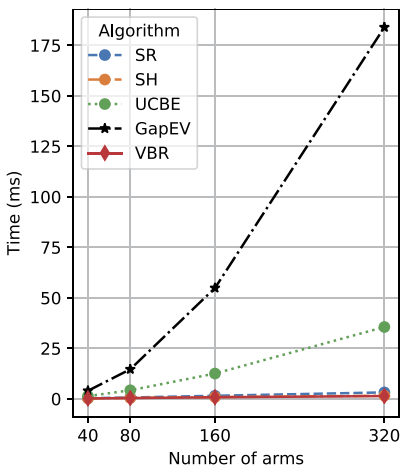


Figure 3: Execution time on *Bernoulli* with varying K and $T = 100 \cdot K$ (100 samples per arm on average).

The *other* calculations performed by the estimation algorithm may end up being at least as expensive as collecting the samples from the arms. Therefore, in this section we compare the actual execution time of the algorithms.

Figure 3 shows the time performance of all algorithms on *Bernoulli*. The results show that all algorithms perform in very similar time, except UCBE and GapEV, which are significantly slower. That can be explained by the fact that those algorithms assign samples to arms *one at a time*, whereas all other algorithms assign blocks of samples. GapEV is further encumbered by the necessity to update the adaptive a parameter after each new sample. Notice that our

	VBR	GapEV	UCBE	SH	SR
Normal-S	**1**	-	1.23	1.28	1.51
Normal-M	**1**	-	1.10	1.08	1.28
Normal-L	1	-	**0.97**	**0.97**	1.04
Rademacher	**1**	1.51	1.30	1.29	1.54
Bernoulli	**1**	1.14	1.13	1.11	1.36

Table 1: Normalized misidentification rates for different classes of inputs. Values are averaged over all values of the budget and then normalized w.r.t. the performance of VBR. The lowest value for each row is emphasized.

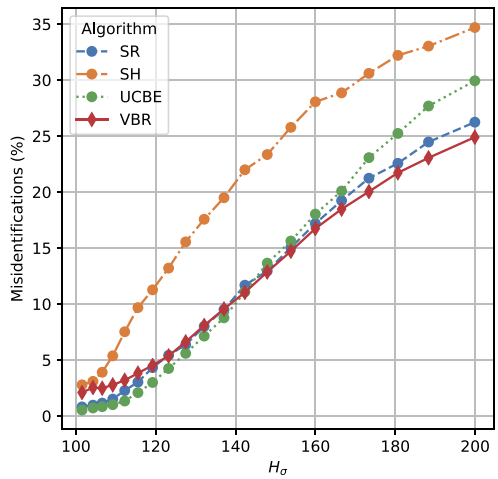


Figure 4: Percentage of misidentification on *Normal-H2* (random normals with fixed H_2 and varying values of H_σ).

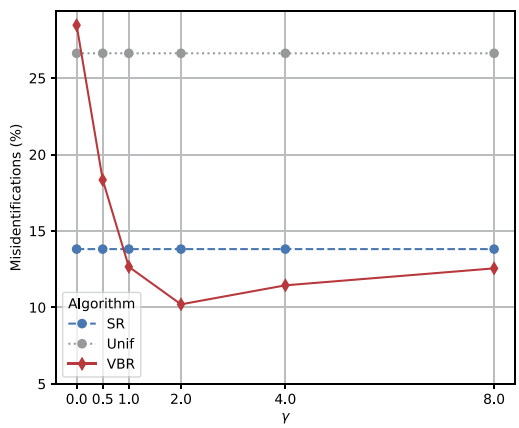


Figure 5: Percentage of misidentification for different values of γ on *Normal-M*.

UCBE implementation attempts to optimize time efficiency by employing appropriate data structures (balanced trees).

Estimating the problem hardness

Parameter H_2 has been proved to be connected to the hardness of the problem via a lower bound [1] and upper bounds that apply to the algorithms SR, SH, and UCBE. However, H_2 depends only on the expected values of the arms, and does not take variances into account. For this reason, we introduced the refined measure H_σ. To show that H_σ can be a more accurate measure of problem hardness, we report the result of running a set of experiments on a family of normal distributions with *fixed* H_2.

Figure 4 plots the percentage of misidentification on random normals grouped by intervals of H_σ. More precisely, for each problem instance we computed its H_σ value and we ran all the algorithms. Let $\underline{H}_\sigma$ and $\overline{H}_\sigma$ denote the minimum and maximum value of H_σ occurring in the experiment, we divided the interval $[\underline{H}_\sigma, \overline{H}_\sigma]$ into 20 sub-intervals, so that each sub-interval contains the same number of problem instances. The x-coordinates of the points in Figure 4 are

the right endpoints of the 20 sub-intervals. The y-coordinates report the average percentage of misidentification on that class of instances.

The plot shows that, with constant H_2, H_σ is very accurate in predicting the remaining variability in problem hardness (in terms of misidentifications).

On the choice of γ

The rejection confidence parameter γ allows us to tune the behavior of VBR between two extremes. For values close to 0, VBR behaves more and more like the uniform allocation strategy[6], because all arms except one are discarded during the first phase; then, all arms are re-evaluated in the catch-up phase and the final decision is made. For large values of γ, no arm is ever discarded based on its upper confidence bound and the algorithm falls back on behaving like SR.

Figure 5 plots the accuracy of VBR for different values of γ, and compares it to SR and the uniform allocation algorithm (labeled *Unif*). The results confirm the above analysis and show that the performance of VBR on random Gaussians is best for γ close to 2.

5 Conclusions

We addressed the problem of stochastic best arm identification in multi-arm bandits considering the fixed-budget setting. We proposed the VBR method, a novel sequential rejection approach exploiting variance for pull allocation and arm rejection.

Alternative methods in the literature are based on the estimated mean values only, with the exception of GapEV [9], where variance exploitation has been applied to the multi-bandit setting, but the approach is limited by design to bounded distributions and does not leveradge incremental rejection. Analogously to UCBE, GapEV assigns samples to arms one at a time, which predictably slows down the execution, as shown by our experiments (see Figure 3).

We introduced and detailed the novel approach, providing both theoretical and empirical results. On the one hand, we obtain a theoretical upper bound for the accuracy of VBR which uses a novel variance-based measure H_σ of problem complexity. On the other hand, we empirically compared our approach with respect to alternative methods in the literature, addressing both identification accuracy and time performance. In contrast with assessments based on a few specific cases of input distributions [1, 19], we evaluated the approaches with respect to a variety of randomly generated input distributions, both bounded and unbounded.

The collected results show that the VBR method dominates the others in the majority of the proposed settings, with up to 30% improved accuracy. In the few remaining test cases, VBR lags behind the best algorithm by a small 3% accuracy margin. This assessment shows that the present approach exhibits an effective balance between performance and adaptiveness to heterogeneous settings, including both discrete and continuous distributions and different parameter setups.

As future work, we aim at extending the proposed method considering not only a budget of arm pulls, but also a target confidence level [7, 8, 10]. The goal would then be to minimize the trials needed to find the best arm while aiming at the prescribed confidence and respecting the maximum budget. Finally, as an application, we are currently developing a preference elicitation framework that exploits VBR to rapidly identify the queries with the highest expected value of information.

[6] The uniform allocation strategy samples each arm the same number of times (that is, T/K) and picks the arm with the highest sample mean.

REFERENCES

[1] Jean-Yves Audibert and Sébastien Bubeck, 'Best arm identification in multi-armed bandits', in *COLT 2010 - 23th Conference on Learning Theory*, pp. 41–53, (2010).

[2] Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári, 'Tuning bandit algorithms in stochastic environments', in *Proceedings of the 18th international conference on Algorithmic Learning Theory (ALT 2007)*, eds., Marcus Hutter, Rocco A. Servedio, and Eiji Takimoto, volume 4754 of *Lecture Notes in Computer Science*, pp. 150–165, Berlin/Heidelberg, Germany, (2007). Springer.

[3] Sébastien Bubeck and Nicolò Cesa-Bianchi, 'Regret analysis of stochastic and nonstochastic multi-armed bandit problems', *Foundations and Trends in Machine Learning*, **5**(1), 1–122, (2012).

[4] Sébastien Bubeck, Rémi Munos, and Gilles Stoltz, 'Pure exploration in multi-armed bandits problems', in *Proceedings of the 20th International Conference on Algorithmic Learning Theory*, ALT'09, pp. 23–37, Berlin, Heidelberg, (2009). Springer-Verlag.

[5] Alexandra Carpentier and Andrea Locatelli, 'Tight (lower) bounds for the fixed budget best arm identification bandit problem', in *COLT 2016 - 29th Conference on Learning Theory*, pp. 590–604, (2016).

[6] Christos Dimitrakakis and Michail G. Lagoudakis, 'Rollout sampling approximate policy iteration', *Mach. Learn.*, **72**(3), 157–171, (September 2008).

[7] Eyal Even-Dar, Shie Mannor, and Yishay Mansour, 'Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems', *J. Mach. Learn. Res.*, **7**, 1079–1105, (2006).

[8] Victor Gabillon, Mohammad Ghavamzadeh, and Alessandro Lazaric, 'Best arm identification: A unified approach to fixed budget and fixed confidence.', in *NIPS 2012*, pp. 3221–3229, (2012).

[9] Victor Gabillon, Mohammad Ghavamzadeh, Alessandro Lazaric, and Sébastien Bubeck, 'Multi-bandit best arm identification', in *Advances in Neural Information Processing Systems 24*, eds., J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, 2222–2230, Curran Associates, Inc., (2011).

[10] Kevin G. Jamieson, Matthew Malloy, Robert D. Nowak, and Sébastien Bubeck, 'lil' UCB: An optimal exploration algorithm for multi-armed bandits.', in *COLT*, eds., Maria-Florina Balcan, Vitaly Feldman, and Csaba Szepesvári, volume 35 of *JMLR Workshop and Conference Proceedings*, pp. 423–439. JMLR.org, (2014).

[11] Zohar Karnin, Tomer Koren, and Oren Somekh, 'Almost optimal exploration in multi-armed bandits', in *International Conference on Machine Learning*, pp. 1238–1246, (2013).

[12] Levente Kocsis and Csaba Szepesvári, 'Bandit based Monte-Carlo planning', in *Proceedings of the 17th European Conference on Machine Learning*, ECML'06, pp. 282–293. Springer-Verlag, (2006).

[13] T.L Lai and Herbert Robbins, 'Asymptotically efficient adaptive allocation rules', *Adv. Appl. Math.*, **6**(1), 4–22, (March 1985).

[14] John R. Lepird, Michael P. Owen, and Mykel J. Kochenderfer, 'Bayesian preference elicitation for multiobjective engineering design optimization', *J. Aerospace Inf. Sys.*, **12**(10), 634–645, (2015).

[15] Oded Maron and Andrew W. Moore, 'Hoeffding races: Accelerating model selection search for classification and function approximation', in *NIPS*, eds., Jack D. Cowan, Gerald Tesauro, and Joshua Alspector, pp. 59–66. Morgan Kaufmann, (1993).

[16] Volodymyr Mnih, Csaba Szepesvári, and Jean-Yves Audibert, 'Empirical Bernstein stopping', in *Proceedings of the Twenty-fifth International Conference on Machine Learning (ICML 2008)*, pp. 672–679, New York, NY, USA, (2008). ACM.

[17] Kevin Regan and Craig Boutilier, 'Regret-based reward elicitation for Markov decision processes', in *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09, pp. 444–451. AUAI Press, (2009).

[18] Herbert Robbins, 'Some aspects of the sequential design of experiments', *Bulletin of the American Mathematical Society*, **58**(5), 527–535, (1952).

[19] Shahin Shahrampour, Mohammad Noshad, and Vahid Tarokh, 'On sequential elimination algorithms for best-arm identification in multi-armed bandits', *IEEE Trans. Signal Processing*, **65**(16), 4281–4292, (2017).

[20] Martin J. Wainwright, *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, 2019.