Variance-Dependent Best Arm Identification

Pinyan Lu[1] **Chao Tao**[2] **Xiaojin Zhang**[3]

[1]ITCS, Shanghai University of Finance and Economics
[2]Department of Computer Science, Indiana University Bloomington
[3]Department of Computer Science and Engineering, The Chinese University of Hong Kong

Abstract

We study the problem of identifying the best arm in a stochastic multi-armed bandit game. Given a set of n arms indexed from 1 to n, each arm i is associated with an unknown reward distribution supported on $[0,1]$ with mean θ_i and variance σ_i^2. Assume $\theta_1 > \theta_2 \geqslant \cdots \geqslant \theta_n$. We propose an adaptive algorithm which explores the gaps and variances of the rewards of the arms and makes future decisions based on the gathered information using a novel approach called *grouped median elimination*. The proposed algorithm guarantees to output the best arm with probability $(1-\delta)$ and uses at most $O\left(\sum_{i=1}^{n}\left(\frac{\sigma_i^2}{\Delta_i^2}+\frac{1}{\Delta_i}\right)(\ln\delta^{-1}+\ln\ln\Delta_i^{-1})\right)$ samples, where Δ_i $(i \geqslant 2)$ denotes the reward gap between arm i and the best arm and we define $\Delta_1 = \Delta_2$. This achieves a significant advantage over the variance-independent algorithms in some favorable scenarios and is the first result that removes the extra $\ln n$ factor on the best arm compared with the state-of-the-art. We further show that $\Omega\left(\sum_{i=1}^{n}\left(\frac{\sigma_i^2}{\Delta_i^2}+\frac{1}{\Delta_i}\right)\ln\delta^{-1}\right)$ samples are necessary for an algorithm to achieve the same goal, thereby illustrating that our algorithm is optimal up to doubly logarithmic terms.

1 INTRODUCTION

The stochastic multi-armed bandit (MAB) is a famous framework that captures well the trade-off between exploration and exploitation. In the MAB game, a player faces a set of n $(n \geqslant 2)$ arms indexed from 1 to n. When arm i is sampled, the player observes an instant reward which is *i.i.d.* generated from an unknown distribution $\mathcal{D}_i$ supported on $[0,1]$ with mean θ_i and variance σ_i^2. In the *pure exploration* setting of a MAB game, by making a sequence of samples, the player identifies one (or a set of) desired arm(s). This framework is motivated by many application domains such as medical trials Robbins [1952], communication networks Audibert and Bubeck [2010], simulation optimization Chen and Lee [2011], recommendation systems Kohli et al. [2013], and crowdsourcing Zhou et al. [2014].

In this paper, we focus on the *best arm identification* problem. The *best arm* is the one with the maximum expected reward. Without loss of generality, we assume $\theta_1 > \theta_2 \geqslant \cdots \geqslant \theta_n$ which is however not known beforehand to the player. We say an algorithm is δ-*correct* if it returns the best arm with probability at least $(1-\delta)$. The goal of the best arm identification problem is to design an algorithm equipped by the player to δ-correctly identify the best arm, with as few samples as possible. Previously, the confidence intervals were mainly constructed utilizing the mean rewards of the arms, e.g., Even-Dar et al. [2002], Audibert and Bubeck [2010], Gabillon et al. [2012], Karnin et al. [2013], Jamieson et al. [2014], Chen and Li [2015]. It is worth noting that the variance of the rewards also embodies important information. The variance of rewards could be employed to provide significant advantages over the pure mean-based algorithms. We design an efficient algorithm to solve the problem of best arm identification by exploiting the variance of the rewards, which requires significantly fewer samples in many favorable cases. We further provide a lower bound which illustrates that our algorithm is optimal up to doubly logarithmic terms.

1.1 RELATED WORKS

In the seminal work of Even-Dar et al. [2002], the authors showed that if $\theta_1 - \theta_2 \geqslant \Delta$, then their Median Elimination algorithm uses at most $O(\frac{n}{\Delta^2}\ln\delta^{-1})$ samples [1]. In the same paper, they also showed that for every δ-correct algorithm,

[1]In fact, the algorithm provides the following stronger (PAC) guarantee – if there are multiple arms with mean rewards at least $(\theta_1 - \Delta)$, then the algorithm returns an arbitrary one among these

Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence (UAI 2021), PMLR 161:1120–1129.

the worst-case sample complexity among all instances such that $\theta_1 - \theta_2 \geqslant \Delta$ is at least $\Omega(\frac{n}{\Delta^2}\ln\delta^{-1})$. The $\Theta(\frac{n}{\Delta^2}\ln\delta^{-1})$ bound can be improved when the input data is easy, which is measured via the *reward gaps* between every sub-optimal arm and the best arm. Formally, let $\Delta_i = \theta_1 - \theta_i$ for $i \geqslant 2$ and $\Delta_1 = \Delta_2$ denote the reward gaps. Intuitively, less samples are required if many reward gaps are significantly larger than $\Delta = \Delta_1$. With this intuition, Even-Dar et al. [2002] showed the first *gap-dependent* algorithm called Successive Elimination, which achieves δ-correctness using $O(\sum_{i=2}^n \Delta_i^{-2}(\ln\delta^{-1} + \ln n + \ln\ln\Delta_i^{-1}))$ samples. Since then, the gap-dependent algorithms for the best arm identification problem have been extensively studied, e.g., Gabillon et al. [2012], Karnin et al. [2013], Jamieson et al. [2014], Chen and Li [2015], Chen et al. [2017]. Both the Exponential Gap Elimination algorithm in Karnin et al. [2013] and the lil'UCB algorithm in Jamieson et al. [2014] achieve δ-correctness with sample complexity [2]

$$O\left(\sum_{i=2}^n \Delta_i^{-2}(\ln\delta^{-1} + \ln\ln\Delta_i^{-1})\right). \tag{1}$$

Chen et al. [2017] further showed a δ-correct algorithm with sample complexity

$$O\Bigg(\sum_{i=2}^n \frac{1}{\Delta_i^2}(\ln\delta^{-1} + \mathrm{Ent}(\Delta_2,\ldots,\Delta_n)) + \frac{1}{\Delta_2^2}\ln\ln\frac{1}{\Delta_2}\cdot\mathrm{polylog}(n,\delta^{-1})\Bigg), \tag{2}$$

where $\mathrm{Ent}(\Delta_2,\ldots,\Delta_n)$ is an entropy-like function. This bound improves the result of Karnin et al. [2013] and Jamieson et al. [2014] when the second additive term is dominated by the first term (which is the usual case).

On the lower bound side, Mannor and Tsitsiklis [2004], Kaufmann et al. [2016] showed that every gap-dependent δ-correct algorithm uses at least $\Omega(\sum_{i=2}^n \Delta_i^{-2}\ln\delta^{-1})$ samples in expectation; and this lower bound holds for all possible gap parameters. Based on the results in Farrell [1964], Jamieson et al. [2014] showed that even when there are only two arms, for every 0.1-correct algorithm, there exists an input instance where $\Omega(\Delta^{-2}\ln\ln\Delta^{-1})$ samples are needed. Therefore the sample complexity in (1) matches the lower bound up to $\ln\ln\Delta_i^{-1}$ terms for $i \geqslant 3$. The first mentioned lower bound was further improved by Chen et al. [2017] to $\Omega(\sum_{i=2}^n \Delta_i^{-2}(\ln\delta^{-1} + \mathrm{Ent}(\Delta_2,\ldots,\Delta_n))$.

To further improve the sample complexity, another line of research tries to leverage information beyond reward gaps i.e., variance Gabillon et al. [2012] and Kullback–Leibler (KL) divergence Maillard et al. [2011], Garivier and Cappé [2011], Kaufmann and Kalyanakrishnan [2013], Tanczos et al. [2017] to construct a more refined confidence interval. Let $\mathrm{KL}(X,Y)$ denote the KL-divergence between two random variables X and Y. The state-of-the-art algorithm lil-KLUCB proposed in Tanczos et al. [2017] utilizes Chernoff information, derived from the KL divergence and achieves a high-probability sample complexity upper bound scaling as

$$\inf_{\widetilde{\theta}_2,\ldots,\widetilde{\theta}_n} \frac{1}{\mathrm{D}^*(\theta_1,\widetilde{\theta})}\left(\ln(n/\delta) + \ln\ln\frac{1}{\mathrm{D}^*(\theta_1,\widetilde{\theta})}\right) + \sum_{i\geqslant 2}\frac{1}{\mathrm{D}^*(\theta_i,\widetilde{\theta}_i)}\left(\ln\delta^{-1} + \ln\ln\frac{1}{\mathrm{D}^*(\theta_i,\widetilde{\theta}_i)}\right),$$

where $\widetilde{\theta}_i \in (\theta_i,\theta_1)$, $\widetilde{\theta} = \max_{i\geqslant 2}\widetilde{\theta}_i$, and $\mathrm{D}^*(x,y) = \max_{z\in[x,y]}\min\{\mathrm{KL}(\mathrm{Ber}(z),\mathrm{Ber}(x)),\mathrm{KL}(\mathrm{Ber}(z),\mathrm{Ber}(y))\}$ denotes the Chernoff information. However, there is still a $\ln n$ factor appearing in the term corresponding to the number of samples on the best arm.

1.2 OUR RESULTS

Theorem 1.1 (Restatement of Theorem 5.1) *We propose an algorithm called* VD-BESTARMID(n,δ) *which, with probability at least* $(1-\delta)$*, outputs the best arm and uses at most*

$$O\left(\sum_{i=1}^n\left(\frac{\sigma_i^2}{\Delta_i^2} + \frac{1}{\Delta_i}\right)(\ln\delta^{-1} + \ln\ln\Delta_i^{-1})\right) \tag{3}$$

samples.

Since the expected sample complexity of VD-BESTARMID is not guaranteed to be bounded, using the trick developed in Chen et al. [2017], we are also able to transform VD-BESTARMID to an algorithm whose expected sample complexity is bounded.

Theorem 1.2 *We can construct an algorithm* VD-BESTARMID*(n,δ) $(\delta \leqslant .1)$ to return the best arm with probability at least $(1-\delta)$, while the expected sample complexity is $O\left(\sum_{i=1}^n\left(\frac{\sigma_i^2}{\Delta_i^2} + \frac{1}{\Delta_i}\right)(\ln\delta^{-1} + \ln\ln\Delta_i^{-1})\right)$.*

For completeness of the paper, we present the proof of Theorem 1.2 in Appendix 2.

Note that the square term scales with the variance instead of a constant, which could lead to significant improvement in some cases. We present a specific example that VD-BESTARMID(n,δ) achieves better performance than other mean-based algorithms as follows.

arms.

[2]Here for simplicity we assume Δ_i is sufficiently small, and the same applies to the rest of this paper. When Δ_i approaches 1, the doubly logarithmic term should be $\ln(e + \ln\Delta_i^{-1})$ to avoid negative evaluations.

Example 1 *Suppose we are given n Bernoulli arms (i.e., the reward of each arm is either 0 or 1), the mean reward of arm i is $\theta_i = 1 - \frac{i}{n}$ for $i = 1, 2, \ldots, n$. Our variance-dependent algorithm achieves δ-correctness with $O(n \ln n(\ln \delta^{-1} + \ln \ln n))$ samples. In contrast, the expressions in the big-O notations in both* (1) *and* (2) *are $\Omega(n^2 \ln \delta^{-1})$. We show the detailed calculation in Appendix 3.*

Let $[n] = \{1, 2, \ldots, n\}$. In the following theorem, we present a lower bound for algorithms aiming to identify the best arm. Therefore, our algorithmic bound (3) matches the lower bound up to doubly logarithmic terms.

Theorem 1.3 (Restatement of Theorem H.1) *For any $\sigma_i^2 < 0.1, i \in [n]$ and $0 < \Delta_i < 0.1, i = 2, \ldots, n$, there exists an input instance with matching parameters (gaps and variances) such that any δ-correct algorithm ($\delta < 0.1$) needs at least*

$$\Omega\left(\sum_{i=1}^{n}\left(\frac{\sigma_i^2}{\Delta_i^2} + \frac{1}{\Delta_i}\right)\ln \delta^{-1}\right) \tag{4}$$

samples.

1.3 ORGANIZATION AND PROOF OUTLINE

In Section 2, we first describe and analyze a few procedures to estimate the variance of the rewards of a given arm, and the arm's mean reward based on the variance estimation. In Section 3, we present a straightforward way to use these procedures to identify the best arm, with the sub-optimal sample complexity $O(\sum_{i=1}^{n}(\frac{\sigma_i^2}{\Delta_i^2} + \frac{1}{\Delta_i})(\ln \delta^{-1} + \ln \ln \Delta_i^{-1} + \ln n))$ (note the extra $\ln n$ term comparing with our desired bound (3)). Then we develop our main variance-dependent algorithm for best-arm identification in Section 4 and Section 5.

In Section 4, we present a key technical component, procedure BESTARMEST, to estimate the best-arm's mean reward up to ϵ precision with probability $(1 - \delta)$ and uses at most $O(\sum_{i=1}^{n}(\frac{\sigma_i^2}{\epsilon^2} + \frac{1}{\epsilon})(\ln \delta^{-1} + \ln \ln \epsilon^{-1}))$ samples. Note that this bound is similar to that of the median elimination algorithm proposed in Even-Dar et al. [2002] in the sense that both are independent on the reward gap parameters. However, our BESTARMEST procedure does explore the variance information and forms its strategy accordingly. To achieve this goal, BESTARMEST uses the idea *grouped median elimination* and iteratively performs the following procedure: first estimate each arm's reward variance and divide the arms into groups, so that arms in the same group have similar reward variance estimations; then perform variance-dependent mean estimation and median elimination within each group. If the variance estimations were always accurate and the arms were all assigned to the desired groups, it would be relatively easy to show that the algorithm makes progress in each iteration (where "progress" is defined to be an multiplicative reduction of the total variances of the remaining arms). However, in our analysis, we need substantial technical effort to deal with the mis-placed arms, which is achieved by making very refined upper bounds for the number of mis-placed arms according to the severity of the mistake.

In Section 5, we use BESTARMEST as a helper procedure to build our main algorithm. The high level idea here is similar to that of the exponential gap algorithm introduced in Karnin et al. [2013]. However, due to the non-uniformity of variances among the arms, we have to design a new stopping condition for our iterative algorithm. In Appendix 8, we prove the variance-dependent lower bound result. Finally we conclude the paper by mentioning a few future directions in Section 6.

2 VARIANCE-DEPENDENT MEAN ESTIMATION

We first build a few subroutines to estimate the variance of the rewards of a given arm (Section 2.1), as well as the arm's mean reward based on the variance estimation (Section 2.2). These procedures will be useful in building blocks to design our main algorithm. All missing proofs in this section are deferred to Appendix 4.

2.1 VARIANCE ESTIMATION

Our goal of this subsection is to design a procedure to estimate order of the variance of the rewards of a given arm. More specifically, our VAREST(i, δ, ℓ) (Algorithm 1) takes arm i, confidence level δ and a positive number $\ell > 0$ (which is used to control the precision of the estimation) as input, and returns an estimate of the variance σ_i^2 up to precision $\Theta(2^{-\ell})$. We also need a helper procedure VARTEST(i, τ, δ, c) (Algorithm 2), which takes arm i, threshold τ, confidence parameters δ and a positive number $c \geqslant 1$ as input, and checks whether σ_i^2 is above the threshold τ.

Algorithm 1 Variance Estimation, VAREST(i, δ, ℓ) $(\ell > 0)$

Input: Arm i, confidence level δ and a positive number ℓ
for $r \leftarrow 1, 2, 3, \ldots$ **do**
  $\tau_r \leftarrow 1/2^r$
  if $\tau_r \leqslant \ell$ ***or*** VARTEST$(i, \tau_r, \delta/e, 80)$ **then**
    Output: $\tau = \tau_r$ as the estimated variance of the rewards of arm i
  end
end

The following lemma shows the guarantee for the procedure VARTEST.

Algorithm 2 Variance Test, $\textsc{VarTest}(i, \tau, \delta, c)$ $(c \geqslant 1)$

Input: Arm i, threshold τ, confidence level δ and a positive number c

$T \leftarrow \frac{c}{\tau} \ln \delta^{-1}$

Sample arm i for $2T$ times and let $x_1, \ldots, x_{2T}$ be the empirical rewards in sequence

$\widehat{\sigma}_i^2 \leftarrow \frac{1}{T} \sum_{r=1}^{T} \frac{(x_r - x_{r+T})^2}{2}$

if $\widehat{\sigma}_i^2 > \tau$ **then** **Output:** true **else** false

Lemma 2.1 *Suppose $\delta \leqslant e^{-1}$. If $\sigma_i^2 \geqslant 2\tau$, with probability at least $1 - \delta \cdot \left(\frac{\tau}{\sigma_i^2}\right)^c$, $\textsc{VarTest}(i, \tau, \delta, c)$ outputs* true. *If $\sigma_i^2 \leqslant \tau/2$, with probability at least $1 - \delta \cdot \left(\frac{\sigma_i^2}{\tau}\right)^c$, $\textsc{VarTest}(i, \tau, \delta, c)$ outputs* false. *Moreover, the sample complexity is $\frac{2c}{\tau} \ln \delta^{-1}$.*

Now we present the lemma on the guarantee of the procedure VarEst. Note that Lemma 2.2 not only shows a lower bound on the success probability of $\textsc{VarEst}(i, \delta, \ell)$, but also provides an upper bound on the error probability that depends on the logarithmic distance between the algorithm's output and the real variance σ_i^2.

Lemma 2.2 *Suppose $\textsc{VarEst}(i, \delta, \ell)$ returns τ. Let $r_m = \lceil |\log_2 \frac{\sigma_i^2}{\tau}| \rceil$ denote the logarithmic mistake ratio. The algorithm has the following three properties.*

(a) It always holds that $\tau > \ell/2$ and the sample complexity is $O(\frac{1}{\ell} \ln \delta^{-1})$;

(b) If $\sigma_i^2 \in (\ell, 1]$, with probability at least $1 - \delta$, we have $\tau \in [\sigma_i^2/4, 2\sigma_i^2)$ and the sample complexity is $O\left(\frac{1}{\sigma_i^2} \ln \delta^{-1}\right)$. We also have $\Pr[\tau \geqslant x] \leqslant \delta \cdot 2^{-20 r_m}$ when $x \geqslant 2\sigma_i^2$ and $\Pr[\tau \leqslant x] \leqslant \delta \cdot 2^{-20 r_m}$ when $x < \sigma_i^2/4$;

(c) If $\sigma_i^2 \leqslant 2\ell$, we have $\Pr[\tau \geqslant x] = O\left(\delta \cdot 2^{-20 r_m}\right)$ for $x \geqslant \max\{2\ell, 2\sigma_i^2\}$.

2.2 VARIANCE-DEPENDENT MEAN ESTIMATION

In this section, we present $\textsc{MeanEst}(i, \epsilon, \delta)$ (Algorithm 3) which estimates the mean reward of a given arm i up to ϵ additive error with probability at least $1 - \delta$ with sample complexity depending on σ_i^2.

At a high level, we first estimate the variance of the rewards of a given arm, then apply Proposition A.2 (Bernstein's Inequality) to control the number of samples needed for an estimate up to the given precision requirement. We show the following lemma.

Lemma 2.3 *With probability at least $1 - \delta$, $\textsc{MeanEst}(i, \epsilon, \delta)$ outputs an estimate (namely $\widehat{\theta}_i$) of the mean reward of arm i such that $|\widehat{\theta}_i - \theta_i| \leqslant \epsilon$ and the sample complexity is $O\left(\left(\frac{\sigma_i^2}{\epsilon^2} + \frac{1}{\epsilon}\right) \ln \delta^{-1}\right)$.*

Algorithm 3 Mean Estimation, $\textsc{MeanEst}(i, \epsilon, \delta)$

Input: Arm i, accuracy ϵ, and confidence level δ

$\widehat{\sigma}_i^2 \leftarrow \textsc{VarEst}(i, \delta/2, \epsilon)$

Sample arm i for $\left(\frac{8\widehat{\sigma}_i^2}{\epsilon^2} + \frac{2}{3\epsilon}\right) \ln \frac{4}{\delta}$ times and let $\widehat{\theta}_i$ denote its empirical mean reward

Output: $\widehat{\theta}_i$ as the estimated mean reward of arm i

Now we prove a few stronger properties of MeanEst which will be useful for building our main algorithm.

Lemma 2.4 *Let Q be the samples used by $\textsc{MeanEst}(i, \epsilon, \delta)$. There exists a constant $c > 0$ such that*

(a) $Q \leqslant \frac{c}{\epsilon^2} \ln \delta^{-1}$;

(b) for integers $j \geqslant 3$, we have $\Pr\left[Q \leqslant c\left(\frac{j\sigma_i^2}{\epsilon^2} + \frac{1}{\epsilon}\right) \ln \delta^{-1}\right] \geqslant 1 - \delta \cdot 2^{-20j}$.

3 WARM-UP: NAÏVE VARIANCE-DEPENDENT BEST-ARM IDENTIFICATION

In this section, we present a straightforward way (Algorithm NaiveBestArm) of using the variance-dependent procedure MeanEst to iteratively reject non-optimal arms and finally identify the best arm. The analysis adopts the union bound on all arms and therefore introduces an extra $\ln |S|$ (where S is the input candidate arms) factor in the sample complexity. In particular, we show the following theorem. The algorithm and missing proofs in this section are deferred to Appendix 5.

Theorem 3.1 *With probability at least $1 - \delta$, the $\textsc{NaiveBestArm}(S, \delta)$ algorithm outputs the best arm in S and the sample complexity is $O\left(\sum_{i \in S} \left(\frac{\sigma_i^2}{\Delta_i^2} + \frac{1}{\Delta_i}\right) (\ln \delta^{-1} + \ln \ln \Delta_i^{-1} + \ln |S|)\right)$.*

It is also straightforward to get the following PAC-style statement where an ϵ-optimal arm denotes an arm whose mean reward is ϵ-close to that of the best arm in S.

Corollary 3.2 *There exists an algorithm that with probability at least $1 - \delta$, finds an ϵ-optimal arm in S using at most $O\left(\sum_{i \in S} \left(\frac{\sigma_i^2}{(\Delta_i^\epsilon)^2} + \frac{1}{\Delta_i^\epsilon}\right) (\ln \delta^{-1} + \ln \ln (\Delta_i^\epsilon)^{-1} + \ln |S|)\right)$ samples, where $\Delta_i^\epsilon = \max\{\Delta_i, \epsilon\}$. We use $\textsc{NaiveBestArmEst}(S, \epsilon, \delta)$ to denote this algorithm.*

4 FIND AN ϵ-OPTIMAL ARM

Now we start to develop our main algorithm. We use $S_{[i]}$ to denote the index of the i-th best arm in S. When there is a tie, we break it arbitrarily. In this section, we design a procedure BESTARMEST(S, ϵ, δ) (described in Algorithm 4) which returns an ϵ-optimal arm. In particular, we prove the following theorem. All missing proofs in this section are deferred to Appendix 6.

Theorem 4.1 *With probability at least* $1 - \delta$, BESTARMEST(S, ϵ, δ) *outputs an arm (denoted by* a*) satisfying* $|\theta_a - \theta_{S_{[1]}}| \leqslant \epsilon$ *and uses* $O\left(\sum_{i \in S}\left(\frac{\sigma_i^2}{\epsilon^2} + \frac{1}{\epsilon}\right)(\ln \delta^{-1} + \ln\ln \epsilon^{-1})\right)$ *samples.*

Algorithm 4 Best Arm Estimation, BESTARMEST(S, ϵ, δ)

Input: A set of arms S, accuracy ϵ, and confidence level δ
$S_1 \leftarrow$ ITERELIM$(S, \epsilon/3, \delta/3)$
if $\epsilon^{-1} \leqslant \ln|S|$ **then** $S_2 \leftarrow S_1$ **else** $S_2 \leftarrow$ ITERELIM$(S_1, \epsilon/3, \delta/3)$
$a \leftarrow$ NAIVEBESTARMEST$(S_2, \epsilon/3, \delta/3)$
Output: Arm a

BESTARMEST can be viewed as an extension of the Median Elimination algorithm. The number of samples used by neither of them depend on the reward gaps. However, our BESTARMEST algorithm explores the variance information and adapts its strategy accordingly. This procedure is the most technical part of our main algorithm. It employs two subroutines ITERELIM and GROUPELIM described in Algorithms 5 and 6.

Algorithm 5 Iterative Elimination, ITERELIM(S, ϵ, δ)

Input: Arm set S, accuracy ϵ, and confidence level δ
Let $\beta \leftarrow \sqrt{255}/16 \cdot e^{.001}$, $\epsilon_r \leftarrow \beta^r(1-\beta)\epsilon$, and $\delta_r \leftarrow e^{-.1r}(1-e^{-.1})\delta$ for $r \geqslant 0$
$T_0 \leftarrow S$, $R_0 \leftarrow \emptyset$, $r \leftarrow 0$
while $|T_r| > 10$ **do**
  $\langle T_{r+1}, R^{r+1}\rangle \leftarrow$ GROUPELIM$(T_r, \epsilon_r, \delta_r)$
  $R_{r+1} \leftarrow R_r \cup R^{r+1}$
  $r \leftarrow r + 1$
end
Output: $T \leftarrow T_r \cup R_r$

Comparing our algorithm with the Median Elimination algorithm in Even-Dar et al. [2002], we note that the major difference is that we use the *grouped median elimination* (GROUPELIM) instead. If, in each iteration, we simply eliminate a constant fraction of the arms according to their empirical means, we cannot guarantee that the samples needed in each iteration reduces at an exponential rate and the total work converges, which is the case in Median Elimination. This is because in our algorithm, the sample complexity relates to the total reward variances of the active arms, rather than the number of active arms. This non-uniformity among the arms may admit the scenario where the eliminated arms have small reward variances and the elimination process does not reduce the total variances by a constant fraction after each iteration.

Algorithm 6 Grouped Median Elimination, GROUPELIM(S, ϵ, δ)

Input: Arm set S, accuracy ϵ, and confidence level δ
Let $N \leftarrow \lceil \log_2(2/\epsilon) \rceil$ be the number of buckets
for $i \in S$ **do** $\widehat{\sigma}_i^2 \leftarrow$ VAREST$(i, \delta/(2N^2), \epsilon)$
Define bucket $\widehat{B}_j \leftarrow \{i \in S | 2^{-j} < \widehat{\sigma}_i^2 \leqslant 2^{-j+1}\}$ for $j = [N]$, and let $T \leftarrow \emptyset$
for $j \leftarrow 1$ ***to*** N **do**
  if $|\widehat{B}_j| \geqslant 2$ **then**
    Let $\widehat{\theta}_i \leftarrow$ MEANEST$(i, \epsilon/2, \delta/(9N))$ for all $i \in \widehat{B}_j$
    Let $\widehat{m}_j$ be the median of the empirical means of the arms in $\widehat{B}_j$
    $T_j \leftarrow \widehat{B}_j \backslash \{i \in \widehat{B}_j | \widehat{\theta}_i < \widehat{m}_j\}$
    $T \leftarrow T \cup T_j$
  else
    Put arm in $\widehat{B}_j$ into the recycle bin R
  end
end
Output: T and R

To solve this problem, our GROUPELIM procedure partitions the arms into buckets according to their empirical reward variances, so that the arms in the same bucket have similar variances of rewards (up to a multiplicative constant factor). If the partition is perfect (i.e., the empirical estimation matches with the true variances and every arm is assigned to the correct bucket), performing median elimination within each group would successfully reduce the total variances by a constant fraction.

To deal with variance estimation noise and imperfect partition, we make considerable effort to upper bound the fraction of arms put in wrong buckets, where the bound is very refined and depends on the distance between the desired and empirical buckets. Another consequence of the noise is that, besides the active arm set T returned by GROUPELIM, we have to introduce a recycle set R of arms. The arms in R do not participate in future rounds of elimination in ITERELIM. However, they appear as the returned arms of ITERELIM. Indeed, the procedure ITERELIM returns a small set of arms instead of the optimal arm. Finally, we use BESTARMEST to examine this small set again to identify the best arm.

We start the sketch of the analysis of our algorithms by presenting the following statement for GROUPELIM.

Theorem 4.2 *With probability at least* $1 - \delta$, GROUPELIM(S, ϵ, δ) *outputs two sets* T *and* R *of arms and has the following four guarantees:*

(a) $|R| = O(\ln \epsilon^{-1})$;
(b) $\sum_{a\in T}(\sigma_a^2+\epsilon) \leqslant \frac{255}{256}\cdot\sum_{a\in S}(\sigma_a^2+\epsilon)$;
(c) $|\theta_{(T\cup R)_{[1]}} - \theta_{S_{[1]}}| \leqslant \epsilon$;
(d) uses $O\left(\sum_{i\in S}\left(\frac{\sigma_i^2}{\epsilon^2}+\frac{1}{\epsilon}\right)(\ln\delta^{-1}+\ln\ln\epsilon^{-1})\right)$ *samples.*

The proof of Theorem 4.2 is split into three subsections. The first claim is easy to verify and shown in the form of the short Lemma 4.6. In Section 4.1, we define an event $\mathcal{E}$ (Equation (5)) concerning about the fraction of the arms put in wrong buckets, and use Lemma 4.7 to show that $\mathcal{E}$ holds with high probability $1-\delta/3$. In Section 4.2, we prove Lemma 4.13, i.e., $\mathcal{E}$ implies the second claim of the theorem. In Appendix 6.10, we prove Lemmas F.2 and F.4, showing that both the probabilities that the third and the fourth claims of the theorem hold are at least $1-\delta/3$. Finally the theorem is proved by a straightforward union bound.

The following theorem shows the guarantee of ITERELIM, and will be proved in Appendix 6.11.

Theorem 4.3 *With probability at least* $1-\delta$, ITERELIM(S,ϵ,δ) *outputs an arm set* T *and has the following three guarantees,*

(a) $|T| = O((\ln|S|)^2\ln\epsilon^{-1})$;
(b) $|\theta_{T_{[1]}} - \theta_{S_{[1]}}| \leqslant \epsilon$;
(c) uses $O\left(\sum_{i\in S}\left(\frac{\sigma_i^2}{\epsilon^2}+\frac{1}{\epsilon}\right)(\ln\delta^{-1}+\ln\ln\epsilon^{-1})\right)$ *samples.*

Finally, with the help of Theorems 4.2 and 4.3, we prove the main theorem on BESTARMEST in Section 4.3.

4.1 UPPER BOUNDS ON FRACTION OF ARMS IN WRONG BUCKETS

For notational convenience, for each $\widehat{B}_j$ $(j=1,2,\ldots,N)$, we set $l(\widehat{B}_j)=2^{-j}$ and $u(\widehat{B}_j)=2^{-j+1}$ as the lower and upper bounds on the estimated reward variances of the arms in $\widehat{B}_j$. We also introduce the "ideal" partition $B_j=\{i\in S \mid 2^{-j}<\sigma_i^2\leqslant 2^{-j+1}\}$ for $j=1,2,\ldots,N-1$ and $B_N=\{i\in S\mid 0\leqslant\sigma_i^2\leqslant 2^{-N+1}\}$. Similarly, we set $l(B_j)=2^{-j}$ for $j=1,2,\ldots N-1$ and $u(B_j)=2^{-j+1}$ for $j=1,2,\ldots,N$, with the exception that $l(B_N)=0$.

Now we list the following simple facts about the procedure GROUPELIM.

Lemma 4.4 $\{\widehat{B}_1,\widehat{B}_2,\ldots,\widehat{B}_N\}$ *is a partition of* S.

Lemma 4.5 *If* $|\widehat{B}_j|\geqslant 2$, *there is* $|T_j|\leqslant\frac{2}{3}|\widehat{B}_j|$.

Lemma 4.6 $|R| = O(\ln\epsilon^{-1})$.

We define $\mathcal{E}$ to be the event

$$\Big\{|B_i\cap\widehat{B}_j| < |B_i|\cdot 2^{-10|i-j|}\cdot N^{-1} \text{ for } \forall|i-j|\geqslant 3\Big\}. \quad (5)$$

In words, it means that the fraction of the arms that are empirically put in a wrong bucket becomes exponentially small as the error distance increases. We now show such an event happens with high probability, which is the main statement of this subsection.

Lemma 4.7 $\Pr[\mathcal{E}]\geqslant 1-\delta/3$.

4.2 PROCEDURE GROUPELIM: MULTIPLICATIVE REDUCTION OF THE TOTAL VARIANCES

We say that B_i *pollutes* $\widehat{B}_j$ (or $\widehat{B}_j$ is *polluted* by B_i) if and only if $|B_i\cap\widehat{B}_j| > |\widehat{B}_j|\cdot 2^{-5|i-j|}$. Intuitively, this means that too many arms (those are supposed to be in B_i) are incorrectly put in $\widehat{B}_j$. Note that the definition of "too many" is in terms of the fraction compared to $|\widehat{B}_j|$ rather than $|B_i|$ as defined in the event $\mathcal{E}$. If $\widehat{B}_j$ is polluted by some B_i where $|i-j|\geqslant 3$, we say that $\widehat{B}_j$ is *bad*. Otherwise, we say that $\widehat{B}_j$ is *good*.

The following lemma shows that for a good bucket $\widehat{B}_j$, as long as it is not the last three buckets, the arms discarded from the bucket aggregate a constant fraction of variances.

Lemma 4.8 *Given that* $j\leqslant N-3$, *if* $|\widehat{B}_j|\geqslant 2$ *and* $\widehat{B}_j$ *is good, there is* $\sum_{a\in T_j}\sigma_a^2\leqslant\frac{127}{128}\cdot\sum_{a\in\widehat{B}_j}\sigma_a^2$.

Corollary 4.9 *Given that* $j\leqslant N-3$, *if* $|\widehat{B}_j|\geqslant 2$ *and* $\widehat{B}_j$ *is good, there is* $\sum_{a\in T_j}(\sigma_a^2+\epsilon)\leqslant\frac{127}{128}\cdot\sum_{a\in\widehat{B}_j}(\sigma_a^2+\epsilon)$.

We now prove a similar statement as Corollary 4.9, but for the last three buckets.

Lemma 4.10 *Given that* $j\geqslant N-2$, *if* $|\widehat{B}_j|\geqslant 2$ *and* $\widehat{B}_j$ *is good, there is* $\sum_{a\in T_j}(\sigma_a^2+\epsilon)\leqslant\frac{127}{128}\cdot\sum_{a\in\widehat{B}_j}(\sigma_a^2+\epsilon)$.

The following two lemmas control the total reward variances of the arms in a polluted bucket.

Lemma 4.11 *Conditioning on* $\mathcal{E}$, *if* $\widehat{B}_j$ *is polluted by some* B_i *where* $i\leqslant N-1$ *and* $|i-j|\geqslant 3$, *we have* $\sum_{a\in\widehat{B}_j}\sigma_a^2\leqslant N^{-1}\cdot\sum_{a\in S}\sigma_a^2\cdot\frac{1}{256}$.

Lemma 4.12 *Conditioning on* $\mathcal{E}$, *if* $\widehat{B}_j$ *is only polluted by* B_N *where* $|N-j|\geqslant 3$, *we have* $\sum_{a\in\widehat{B}_j}\sigma_a^2\leqslant N^{-1}\cdot\sum_{a\in S}(\sigma_a^2+\epsilon)\cdot\frac{1}{1024}$.

Now, we show that with high probability the total reward variances of the active arms reduce by a constant fraction after the procedure GROUPELIM. In particular, we prove the following lemma.

Lemma 4.13 *Conditioning on event $\mathcal{E}$, we have* $\sum_{a\in T}(\sigma_a^2+\epsilon) \leqslant \frac{255}{256}\sum_{a\in S}(\sigma_a^2+\epsilon)$.

Proof According to Lemma 4.11 and Lemma 4.12, if $\widehat{B}_j$ is polluted by some B_i where $|i-j| \geqslant 3$, there is $\sum_{a\in\widehat{B}_j}\sigma_a^2 \leqslant N^{-1}\cdot\sum_{a\in S}(\sigma_a^2+\epsilon)\cdot\frac{1}{256}$ which implies $\sum_{j,\widehat{B}_j \text{ is good}}\sum_{a\in\widehat{B}_j}\sigma_a^2 \geqslant \frac{255}{256}\sum_{a\in S}(\sigma_a^2+\epsilon)$. Hence there is

$$\begin{aligned}
&\frac{\sum_{a\in S}(\sigma_a^2+\epsilon)-\sum_{a\in T}(\sigma_a^2+\epsilon)}{\sum_{a\in S}(\sigma_a^2+\epsilon)}\\
\geqslant\ &\frac{\sum_{j,\widehat{B}_j \text{ is good}}\sum_{a\in\widehat{B}_j\setminus T_j}\sigma_a^2}{\frac{256}{255}\cdot\sum_{j,\widehat{B}_j \text{ is good}}\sum_{a\in\widehat{B}_j}\sigma_a^2}\\
\geqslant\ &\frac{255}{256}\cdot\min_{j,\widehat{B}_j \text{ is good}}\frac{\sum_{a\in\widehat{B}_j\setminus T_j}\sigma_a^2}{\sum_{a\in\widehat{B}_j}\sigma_a^2}. \qquad (6)
\end{aligned}$$

When $|\widehat{B}_j|=1$, $T_j=\emptyset$ which implies $\frac{\sum_{a\in(\widehat{B}_j-T_j)}\sigma_a^2}{\sum_{a\in\widehat{B}_j}\sigma_a^2}=1$. When $\widehat{B}_j$ is good and $|\widehat{B}_j|\geqslant 2$, according to Corollary 4.9 and Lemma 4.10, there is $\frac{\sum_{a\in(\widehat{B}_j-T_j)}\sigma_a^2}{\sum_{a\in\widehat{B}_j}\sigma_a^2}\geqslant\frac{1}{128}$. Therefore, we have (6) $\geqslant\frac{255}{256}\cdot\frac{1}{128}\geqslant\frac{1}{256}$, which concludes the proof of this lemma.

4.3 ANALYSIS OF THE BESTARMEST ALGORITHM

Now we are ready to analyze the BESTARMEST algorithm and prove the main theorem (Theorem 4.1) of this subsection.

First, we define the following three events about the BESTARMEST procedure. Let c be the hidden constant in Corollary 3.2 and Theorem 4.3.

- Let $\mathcal{E}_1$ denote the event $|S_1|\leqslant c(\ln|S|)^2\ln\epsilon^{-1}$, $|\theta_{(S_1)_{[1]}}-\theta_{S_{[1]}}|\leqslant\epsilon/3$, and the sample complexity of Line 2 is at most $c\sum_{i\in S}\left(\frac{\sigma_i^2}{\epsilon^2}+\frac{1}{\epsilon}\right)(\ln\delta^{-1}+\ln\ln\epsilon^{-1})$.
- Let $\mathcal{E}_2$ denote the event $|S_2|=c(\ln|S_1|)^2\ln\epsilon^{-1}$, $|\theta_{(S_2)_{[1]}}-\theta_{(S_1)_{[1]}}|\leqslant\epsilon/3$, and the sample complexity of Line 3 is at most $c\sum_{i\in S_1}\left(\frac{\sigma_i^2}{\epsilon^2}+\frac{1}{\epsilon}\right)(\ln\delta^{-1}+\ln\ln\epsilon^{-1})$.
- Let $\mathcal{E}_3$ denote the event $|\theta_a-\theta_{(S_2)_{[1]}}|\leqslant\epsilon/3$ and the sample complexity of Line 4 is at most $c\sum_{i\in S_2}\left(\frac{\sigma_i^2}{\epsilon^2}+\frac{1}{\epsilon}\right)(\ln\delta^{-1}+\ln\ln\epsilon^{-1}+\ln|S_2|)$.

Proof of Theorem 4.1 By Theorem 4.3, we have $\Pr[\mathcal{E}_1]\geqslant 1-\delta/3$ and $\Pr[\mathcal{E}_2]\geqslant 1-\delta/3$. By Corollary 3.2, we have $\Pr[\mathcal{E}_3]\geqslant 1-\delta/3$. Conditioning on event $\mathcal{E}_1\wedge\mathcal{E}_2\wedge\mathcal{E}_3$ which happens with probability $1-\delta$, we will show both claims of Theorem 4.1 hold.

The first claim is because of $|\theta_a-\theta_{S_{[1]}}|\leqslant|\theta_a-\theta_{(S_2)_{[1]}}|+|\theta_{(S_2)_{[1]}}-\theta_{(S_1)_{[1]}}|+|\theta_{(S_1)_{[1]}}-\theta_{S_{[1]}}|\leqslant\epsilon$.

Now we focus on the second claim (about the sample complexity). It suffices to show that the sample complexity of Line 4 meets the desired asymptotic upper bound. We discuss the following two cases.

Case 1: $\epsilon^{-1}\leqslant\ln|S|$**.** Note that $S_2=S_1$ and $O\left(\sum_{i\in S_2}\left(\frac{\sigma_i^2}{\epsilon^2}+\frac{1}{\epsilon}\right)\ln|S_2|\right)=O\left(\frac{|S_1|\ln|S_1|}{\epsilon^2}\right)=O\left(\frac{\ln|S||S_1|\ln|S_1|}{\epsilon}\right)=O\left(\frac{|S|}{\epsilon}\right)$, where the last equality is due to $|S_1|=O((\ln|S|)^2\ln\epsilon^{-1})$. Hence, the sample complexity of Line 4 is

$$\begin{aligned}
&O\left(\sum_{i\in S_2}\left(\frac{\sigma_i^2}{\epsilon^2}+\frac{1}{\epsilon}\right)(\ln\delta^{-1}+\ln\ln\epsilon^{-1}+\ln|S_2|)\right)\\
=\ &O\left(\sum_{i\in S_2}\left(\frac{\sigma_i^2}{\epsilon^2}+\frac{1}{\epsilon}\right)(\ln\delta^{-1}\right.\\
&\left.+\ln\ln\epsilon^{-1})\right)+O\left(\sum_{i\in S_2}\left(\frac{\sigma_i^2}{\epsilon^2}+\frac{1}{\epsilon}\right)\ln|S_2|\right)\\
=\ &O\Big(\sum_{i\in S}\left(\frac{\sigma_i^2}{\epsilon^2}+\frac{1}{\epsilon}\right)(\ln\delta^{-1}+\ln\ln\epsilon^{-1})\Big)+O\left(\frac{|S|}{\epsilon}\right)\\
=\ &O\left(\sum_{i\in S}\left(\frac{\sigma_i^2}{\epsilon^2}+\frac{1}{\epsilon}\right)(\ln\delta^{-1}+\ln\ln\epsilon^{-1})\right).
\end{aligned}$$

Case 2: $\epsilon^{-1}>\ln|S|$**.** Note that $\ln|S_2|=O(\ln\ln|S_1|+\ln\ln\epsilon^{-1})=O(\ln\ln\ln|S|+\ln\ln\epsilon^{-1})=O(\ln\ln\epsilon^{-1})$, where the first and second equalities are due to $|S_2|=O((\ln|S_1|)^2\ln\epsilon^{-1})$ and $|S_1|=O((\ln|S|)^2\ln\epsilon^{-1})$ respectively. Hence, the sample complexity of Line 4 is $O\Big(\sum_{i\in S_2}\left(\frac{\sigma_i^2}{\epsilon^2}+\frac{1}{\epsilon}\right)(\ln\delta^{-1}+\ln\ln\epsilon^{-1}+\ln|S_2|)\Big)=O\left(\sum_{i\in S}\left(\frac{\sigma_i^2}{\epsilon^2}+\frac{1}{\epsilon}\right)(\ln\delta^{-1}+\ln\ln\epsilon^{-1})\right)$.

In both cases, the sample complexity of Line 4 is $O\left(\sum_{i\in S}\left(\frac{\sigma_i^2}{\epsilon^2}+\frac{1}{\epsilon}\right)(\ln\delta^{-1}+\ln\ln\epsilon^{-1})\right)$. Therefore, the sample complexity of the whole procedure also meets the desired upper bound.

5 THE MAIN VARIANCE-DEPENDENT ALGORITHM

Now we are ready to present the main variance-dependent best arm identification algorithm VD-BESTARMID(n,δ) with the help of MEANEST and BESTARMEST developed

in previous sections. All missing proofs in this section are deferred to Appendix 7.

Theorem 5.1 *With probability at least* $1-\delta$, VD-BESTARMID(n,δ) *outputs the best arm and the number of samples used is* $O\left(\sum_{i=1}^{n}\left(\frac{\sigma_i^2}{\Delta_i^2}+\frac{1}{\Delta_i}\right)(\ln\delta^{-1}+\ln\ln\Delta_i^{-1})\right)$.

Algorithm 7 Variance-Dependent Best Arm Identification, VD-BESTARMID(n,δ)

Input: Arm set $S=[n]$ and confidence level δ
$S_1 \leftarrow S$, $r \leftarrow 1$
while $|S_r|>1$ **do**
  Set $\epsilon_r \leftarrow 1/2^{r+2}$ and $\delta_r \leftarrow 1/(2r^2)\cdot\delta$
  for $i \in S_r$ **do** $\widehat{\theta}_i^r \leftarrow$ MEANEST$(i, \frac{\epsilon_r}{2}, \frac{\delta_r}{18})$
  $a_r \leftarrow$ BESTARMEST$(S_r, \frac{\epsilon_r}{2}, \frac{\delta_r}{18})$
  $a_r^* \leftarrow$ BESTARMEST$(S_r\backslash\{a_r\}, \frac{\epsilon_r}{2}, \frac{\delta_r}{18})$
  if $|\widehat{\theta}_{a_r}^r - \widehat{\theta}_{a_r^*}^r| > 2\epsilon_r$ **then Output:** a_r
  $S_{r+1} \leftarrow S_r\backslash\{i\in S_r | \widehat{\theta}_i^r < \widehat{\theta}_{a_r}^r - \epsilon_r\}$
  $r \leftarrow r+1$
end
Output: The remaining arm in S_r

We present the details of VD-BESTARMID(n,δ) in Algorithm 7. It has a similar structure to that of the Exponential Gap Elimination algorithm in Karnin et al. [2013] as our algorithm also keeps a confidence interval ϵ_r which halves after each round. Within a round, we estimate the mean reward of each arm up to confidence interval ϵ_r and an arm will be discarded if its estimation is ϵ_r below that of the best arm. However, due to non-uniformity of the reward variances of the arms, we cannot repeat this process until there is only one arm left (as is done in the Exponential Gap Elimination algorithm), otherwise the sample complexity would not satisfy the desired upper bound. Instead, we design a new stopping condition (Line 8) which may be triggered earlier.

The proof of Theorem 5.1 is split into two parts: correctness (the best arm is identified with high probability proved by Lemma 5.4 in Section 5.1) and sample complexity (proved by Lemma 5.10 in Section 5.2). We finally obtain Theorem 5.1 by combining these two lemmas with a union bound.

The rest of this section is devoted to the proof of Theorem 5.1.

5.1 CORRECTNESS

We use $\mathcal{M}_1$ to denote the event $\widehat{\theta}_{S_{[1]}}^r \geqslant \widehat{\theta}_{a_r}^r - \epsilon_r$ for every round r, and use $\mathcal{M}_2$ to denote the event that VD-BESTARMID(n,δ) terminates with $r = O(\ln\Delta_2^{-1})$ and returns the best arm. We have the following two lemmas.

Lemma 5.2 $\Pr[\mathcal{M}_1] \geqslant 1-\delta/9$.

Lemma 5.3 $\Pr[\mathcal{M}_2|\mathcal{M}_1] \geqslant 1-2\delta/9$.

We now show the correctness lemma as follows.

Lemma 5.4 *With probability at least* $1-\delta/3$, VD-BESTARMID(n,δ) *terminates with* $r = O(\ln\Delta_2^{-1})$ *and returns the best arm.*

Proof It suffices to prove $\Pr[\mathcal{M}_2] \geqslant 1-\delta/3$. By Lemma 5.2 and 5.3, we have $\Pr[\mathcal{M}_2] \geqslant \Pr[\mathcal{M}_2|\mathcal{M}_1]\Pr[\mathcal{M}_1] \geqslant 1-\delta/3$.

5.2 SAMPLE COMPLEXITY

For each $1 \leqslant s \leqslant \lceil\log_2(1/\Delta)+1\rceil$, we define the set $A_s = \{i \in S \mid 2^{-s} < \Delta_i \leqslant 2^{-s+1}\}$, and let $n_s = |A_s|$. Also, we denote the set of arms from A_s surviving after round r by $S_{r,s} = S_r \cap A_s$.

We will show that from round s onwards, every sub-optimal arm in A_s is eliminated with high probability. Specifically, we show the following lemma.

Lemma 5.5 *Conditioning on* $\mathcal{M}_1$, *with probability at least* $1-\delta_r/4$, *we have* $\widehat{\theta}_i^r < \widehat{\theta}_{a_r}^r - \epsilon_r$ *for any arm* $i \in S_{r-1,s}$ *and round* $r \geqslant s$.

Let I_i^r denote the random variable $\mathbb{1}\{i \in S_r\}$. We also define $T_i^r = \left(\frac{\sigma_i^2}{\epsilon_r^2}+\frac{1}{\epsilon_r}\right)(\ln\delta_r^{-1}+\ln\ln\epsilon_r^{-1})$.

In the desired event (which is explicitly defined by event $\mathcal{M}_3$ and analyzed in Lemma 5.9 soon afterwards), we may bound the number of pulls to arm i in round r by $I_i^r T_i^r$. In light of this, the following two lemmas help to upper-bound the number of pulls to the sub-optimal arms where c is a constant.

Lemma 5.6 *Conditioning on* $\mathcal{M}_1$, *we have that with probability at least* $1-\left(\frac{\delta}{8}\right)^j$, $\sum_{r=1}^{+\infty} I_i^r T_i^r \leqslant c4^j\left(\frac{\sigma_i^2}{\Delta_i^2}+\frac{1}{\Delta_i}\right)(\ln\delta^{-1}+\ln\ln\Delta_i^{-1})$ *for* $i \neq S_{[1]}$.

Lemma 5.7 *Conditioning on* $\mathcal{M}_1$, *we have that with probability at least* $1-\frac{\delta}{18}$, $\sum_{i\neq S_{[1]}}\sum_{r=1}^{+\infty} I_i^r T_i^r \leqslant O\left(\sum_{i\neq S_{[1]}}\left(\frac{\sigma_i^2}{\Delta_i^2}+\frac{1}{\Delta_i}\right)(\ln\delta^{-1}+\ln\ln\Delta_i^{-1})\right)$.

The following lemma helps to upper-bound the number of the pulls to the best arm.

Lemma 5.8 *When $\mathcal{M}_2$ happens, we have* $\sum_{r=1}^{+\infty} I^r_{S_{[1]}} T^r_{S_{[1]}} = O\left(\left(\frac{\sigma_1^2}{\Delta_1^2} + \frac{1}{\Delta_1}\right)(\ln \delta^{-1} + \ln\ln \Delta_1^{-1})\right)$.

We use $\mathcal{M}_3$ to denote the event that, for each r, the number of samples used in round r is $\sum_{i=1}^n O(I_i^r T_i^r)$. The following lemma shows that $\mathcal{M}_3$ happens with high probability.

Lemma 5.9 $\Pr[\mathcal{M}_3] \geqslant 1 - \delta/6$.

We are now ready to prove the following lemma on the sample complexity of VD-BESTARMID.

Lemma 5.10 *With probability at least $1-2\delta/3$, the sample complexity of* VD-BESTARMID(n, δ) *is*

$$O\left(\sum_{i=1}^n \left(\frac{\sigma_i^2}{\Delta_i^2} + \frac{1}{\Delta_i}\right)(\ln \delta^{-1} + \ln\ln \Delta_i^{-1})\right).$$

Proof Note that $\Pr[\mathcal{M}_1] \geqslant 1 - \delta/9$ by Lemma 5.2. Further by Lemma 5.7, with probability at least $(1 - \delta/9)(1 - \delta/18) \geqslant 1 - \delta/6$, we have $\sum_{i \neq S_{[1]}} \sum_{r=1}^{+\infty} I_i^r T_i^r = O\left(\sum_{i \neq S_{[1]}} \left(\frac{\sigma_i^2}{\Delta_i^2} + \frac{1}{\Delta_i}\right)(\ln \delta^{-1} + \ln\ln \Delta_i^{-1})\right)$. Note that $\Pr[\mathcal{M}_2] \geqslant 1 - \delta/3$ by Lemma 5.4. Further by Lemma 5.8, with probability at least $1 - \delta/3$, it holds that $\sum_{r=1}^{+\infty} I^r_{S_{[1]}} T^r_{S_{[1]}} = O\left(\left(\frac{\sigma_1^2}{\Delta_1^2} + \frac{1}{\Delta_1}\right)(\ln \delta^{-1} + \ln\ln \Delta_1^{-1})\right)$. Via a union bound, with probability at least $1 - \delta/2$,

$$\sum_{i=1}^n \sum_{r=1}^{+\infty} I_i^r T_i^r = O\left(\sum_{i=1}^n \left(\frac{\sigma_i^2}{\Delta_i^2} + \frac{1}{\Delta_i}\right)(\ln \delta^{-1} + \ln\ln \Delta_i^{-1})\right). \quad (7)$$

Note that $\Pr[\mathcal{M}_3] \geqslant \delta/6$ by Lemma 5.9. Conditioning on (7) and event $\mathcal{M}_3$ which happens with probability at least $1 - 2\delta/3$ (via a union bound), the sample complexity of algorithm VD-BESTARMID(n, δ) is $\sum_{r=1}^{+\infty} \sum_{i=1}^n O(I_i^r T_i^r) = O\left(\sum_{i=1}^n \sum_{r=1}^{+\infty} I_i^r T_i^r\right) = O\left(\sum_{i=1}^n \left(\frac{\sigma_i^2}{\Delta_i^2} + \frac{1}{\Delta_i}\right)(\ln \delta^{-1} + \ln\ln \Delta_i^{-1})\right)$.

6 CONCLUSION AND FUTURE WORKS

In this paper, we present a variance-dependent best arm identification algorithm and the nearly matching sample complexity lower bound.

While our algorithm almost achieves theoretical optimality, its empirical performance suffers from the large constant factors introduced by multiple subroutines. It is worthwhile to design algorithms with better empirical performance and the same sample complexity bound. The UCB-style algorithms (e.g. lil'UCB in Jamieson et al. [2014]) are a very promising direction towards this end.

On the theoretical side, we believe that it is promising to combine our approach with the ideas in Chen et al. [2017] and improve the doubly-logarithmic terms in our sample complexity bound. It is very interesting to investigate the ultimate sample complexity of the problem.

Acknowledgements

We want to thank Yuan Zhou for providing valuable ideas and many helpful discussions. Pinyan Lu is supported by Science and Technology Innovation 2030 –"New Generation of Artificial Intelligence" Major Project No.(2018AAA0100903), NSFC grant 61922052 and 61932002, Innovation Program of Shanghai Municipal Education Commission, Program for Innovative Research Team of Shanghai University of Finance and Economics, and the Fundamental Research Funds for the Central Universities. Chao Tao is supported in part by NSF IIS-1633215, NSF CCF-1844234, and NSF CCF-2006591.

References

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

Chun-hung Chen and Loo Hay Lee. *Stochastic simulation optimization: an optimal computing budget allocation*, volume 1. 2011.

Lijie Chen and Jian Li. On the optimal sample complexity for best arm identification. *arXiv preprint arXiv:1511.03774*, 2015.

Lijie Chen, Jian Li, and Mingda Qiao. Towards instance optimal bounds for best arm identification. In *COLT*, 2017.

Eyal Even-Dar, Shie Mannor, and Yishay Mansour. Pac bounds for multi-armed bandit and markov decision processes. In *COLT*, 2002.

Roger H Farrell. Asymptotic behavior of expected sample size in certain one sided tests. *The Annals of Mathematical Statistics*, pages 36–72, 1964.

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

Aurélien Garivier and Olivier Cappé. The KL-UCB algorithm for bounded stochastic bandits and beyond. In *COLT*, 2011.

Kevin Jamieson, Matthew Malloy, Robert Nowak, and Sébastien Bubeck. lil'ucb: An optimal exploration algorithm for multi-armed bandits. In *COLT*, 2014.

Zohar Shay Karnin, Tomer Koren, and Oren Somekh. Almost optimal exploration in multi-armed bandits. In *ICML*, 2013.

Emilie Kaufmann and Shivaram Kalyanakrishnan. Information complexity in bandit subset selection. In Shai Shalev-Shwartz and Ingo Steinwart, editors, *COLT*, 2013.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 17(1):1–42, 2016.

Pushmeet Kohli, Mahyar Salek, and Greg Stoddard. A fast bandit algorithm for recommendation to users with heterogenous tastes. In *AAAI*, 2013.

Odalric-Ambrym Maillard, Rémi Munos, and Gilles Stoltz. A finite-time analysis of multi-armed bandits problems with kullback-leibler divergences. In *COLT*, 2011.

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

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

Ervin Tanczos, Robert Nowak, and Bob Mankoff. A KL-LUCB algorithm for large-scale crowdsourcing. In *NIPS*, 2017.

Yuan Zhou, Xi Chen, and Jian Li. Optimal pac multiple arm identification with applications to crowdsourcing. In *ICML*, 2014.