Approximate Top-m Arm Identification with Heterogeneous Reward Variances

Ruida Zhou
Texas A&M University

Chao Tian
Texas A&M University

Abstract

We study the effect of reward variance heterogeneity in the approximate top-m arm identification setting. In this setting, the reward for the i-th arm follows a σ_i^2-sub-Gaussian distribution, and the agent needs to incorporate this knowledge to minimize the expected number of arm pulls to identify m arms with the largest means within error ϵ out of the n arms, with probability at least $1 - \delta$. We show that the worst-case sample complexity of this problem is

$$\Theta\left(\sum_{i=1}^{n} \frac{\sigma_i^2}{\epsilon^2} \ln \frac{1}{\delta} + \sum_{i \in G^m} \frac{\sigma_i^2}{\epsilon^2} \ln(m) + \sum_{j \in G^l} \frac{\sigma_j^2}{\epsilon^2} \mathrm{Ent}(\sigma_{G^r}^2)\right), \tag{1}$$

where G^m, G^l, G^r are certain specific subsets of the overall arm set $\{1, 2, \ldots, n\}$, and $\mathrm{Ent}(\cdot)$ is an entropy-like function which measures the heterogeneity of the variance proxies. The upper bound of the complexity is obtained using a divide-and-conquer style algorithm, while the matching lower bound relies on the study of a dual formulation.

1 Introduction

In the multi-armed bandit (MAB) model, an agent interacts with a slot machine by pulling one of the many arms and observing the corresponding reward at each time step (Lattimore and Szepesvári, 2020). The goal in the canonical MAB setting is to maximize the cumulative reward. In order to accomplish this goal, algorithms must be designed to balance exploration and exploitation during this online learning process; the objective in this setting is usually referred to as regret minimization. In many applications, the true goal is in fact not to maximize the cumulative reward, but to identify the best arm among all the arms, and regret minimization does not match the true goal in such cases. Instead, the best arm identification problem is the more suitable formulation, and it is a pure exploration problem (Bubeck et al., 2009) that aims to identify the best arm as *fast* and *accurately* as possible.

Approximate best arm identification with fixed confidence is a formal PAC-learning formulation for the best arm identification setting, where the agent is required to identify an arm whose expected reward is not less than that of the best arm by ϵ, with a confidence at least $1 - \delta$. A more general version of this problem is to identify the top-m arms, where the expected rewards of the m arms identified are not less than that of the m-th best arm by ϵ, with a confidence at least $1-\delta$. In this setting, the algorithms will have a performance guarantee in terms of the confidence of success. We refer to these settings as (ϵ, δ) best arm identification, and (ϵ, δ) top-m arm identification, respectively.

In most previous works on multi-armed bandit, an inherent assumption is that the reward distribution of each arm is sub-Gaussian, and moreover, the variance proxies are homogeneous among all the arms. Such an assumption may be natural when the rewards are bounded in a range, or it is reasonable to view the arms as of the same randomness nature (except the mean rewards of the arms). In other applications, this assumption is less suitable, since the reward distributions are naturally heterogeneous. In this work, we consider (ϵ, δ) best m-arm identification with sub-Gaussian distributed rewards when the variance proxies are heterogeneous and known. Our goal is to understand the worst-case sample complexity of the problem as a function of the number of arms n, the number of best arms to be identified m, the variance proxy vector $(\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2)$, i.e., the worst in the class of possible reward distributions with the given variance proxy vector; see Section 3 for a more precise definition.

For a more concrete example application of the problem setting, consider a remote sensing setting, where multiple underground sensors will need to communi-

Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151. Copyright 2022 by the author(s).

cate with the central controller on a wireless link to find the best location to drill for natural gas. The channel noises in the wireless link can be viewed as additive noises on the sensing values themselves, and such channel statistics are usually obtained independent of the sensing but by sending and receiving pilot signals. Therefore, the variances of the arms are indeed known, and the goal is to identify several "best" arms. Although the problem is well motivated by practical applications, our approach to study it is largely theoretical, and the obtained result is theoretical in nature. Particularly, it is well-known that the worst-case sample complexity in the homogeneous (ϵ, δ) top-m arm identification setting is $\Theta\left(\sum_{i\in[n]} \frac{\sigma_i^2}{\epsilon^2}(\ln\frac{1}{\delta} + \ln m)\right)$ (c.f., (Kalyanakrishnan et al., 2012)). However, we observe that the structure of the sample complexity may evolve (specifically, the factor $\ln(m)$ will diminish) as the setting transitions from the homogeneous setting to the heterogeneous setting. Therefore, we focus on this transition behavior and aim to provide a precise characterization *theoretically* .

Several well-known algorithms can be straightforwardly adapted to the problem under consideration. We first consider adapting the naive elimination approach and the median elimination algorithm (Even-Dar et al., 2002; Kalyanakrishnan and Stone, 2010), as well the LUCB (Kalyanakrishnan et al., 2012) and UGapE (Gabillon et al., 2012) algorithms. We observe that the adapted algorithms only perform well in some respective cases. More precisely, the adapted naive elimination algorithm performs well when the heterogeneity is more significant, and the adapted median elimination algorithm performs well when the heterogeneity is less significant. Given this observation, we seek for a new algorithm that can naturally account for the heterogeneity, and propose the variance-grouped median elimination algorithm. There is no need to artificially ascribe an instance as having either high or low heterogeneity in this algorithm, and its performance adapts naturally.

We further establish a matching lower bound by reformulating it into an optimization problem, and considering its dual. Combined with this lower bound, we show the proposed algorithm is in fact optimal. The worst-case sample complexity, as given in (1), is in general proportional to the sum of the reward variances, and has three components. The first component (with $\ln\frac{1}{\delta}$) reflects the effect of the confidence parameter, the second component reflects the impact of the more homogeneous subset of the arms, and the last term (with the $\mathrm{Ent}(\cdot)$ function) reflects the impact of the more heterogeneous subset of the arms. The result naturally degrades if the reward variances are indeed homogeneous, which essentially has only the first two components. The third component captures the impact of the heterogeneity, which is not critically related to m, but on the variances $\sigma_{1:n}^2$ through an entropy-like function. For highly heterogeneous variances, the second term will in fact disappear, and $\mathrm{Ent}(\sigma_{G^r}^2)$ can be of order $O(1)$, thus becoming independent of m completely.

2 Related Works

Multi-armed bandit problems have been extensively studied in the machine learning community in the past decades. A canonical setting is to maximize the cumulative reward, whose asymptotically optimal behavior was first characterized in the seminal work by Lai and Robbins (1985). Good tutorials and books (Bubeck and Cesa-Bianchi, 2012; Slivkins, 2019; Lattimore and Szepesvári, 2020) are readily available.

An alternative setting is to instead identify the best arm. There are in general two lines of research: minimizing the mis-identification probability within a fixed budget of samples (Audibert et al., 2010; Bubeck et al., 2013; Carpentier and Locatelli, 2016), and fast identification with a fixed confidence guarantee (Jamieson and Nowak, 2014). The (ϵ, δ) best arm identification problem belongs to the latter and was introduced in (Even-Dar et al., 2002, 2006), where several elimination based algorithms, such as naive elimination, successive elimination and median elimination algorithms, were proposed. The median elimination algorithm was shown to be worst-case optimal for which a matching lower bound was derived by Mannor and Tsitsiklis (2004). The asymptotic (large number of arms) optimal elimination algorithm was recently discovered (Hassidim et al., 2020), which was inspired by the idea of identifying the "good arms" (Katz-Samuels and Jamieson, 2020). The case of exact best arm identification, i.e., $\epsilon = 0$, motivated algorithms that adapt to the underlining model and usually performs well in an instance-dependent manner (Karnin et al., 2013; Jamieson et al., 2014; Chen and Li, 2015; Garivier and Kaufmann, 2016; Kaufmann et al., 2016).

There are multiple variants of the problem (Zhou et al., 2014; Shen, 2019; Jin et al., 2019; Assadi and Wang, 2020; Chaudhuri and Kalyanakrishnan, 2019). One of the most natural generalization of the best arm identification problem is to identify multiple best arms. The (ϵ, δ) top-m arm identification was studied in (Kalyanakrishnan and Stone, 2010), in which an algorithm named "halving" was proposed, and it bears similarity to the median elimination algorithm. It was later shown that the halving algorithm is indeed worst-case optimal (Kalyanakrishnan et al., 2012). Though more adaptive algorithms were proposed later, such as LUCB (Kalyanakrishnan and Stone, 2010) and UGapE (Gabillon et al., 2012; Kalyanakrishnan et al., 2012), they are not worst-case optimal. For the case of exact top-m arm identification, efforts toward understanding the instance-dependent sample complexity were also made (Kaufmann and Kalyanakrishnan, 2013; Chen et al., 2017; Simchowitz et al., 2017).

Gaussian rewards with heterogeneous variances was considered in the earliest work on best arm identification (Bechhofer, 1954) in the fixed confidence setting, though without a theoretical analysis on the stopping time. The possible variance heterogeneity among arms gained attentions recently in the fixed budget setting (Faella et al., 2020), where the confidence bounds are designed based on central limit theorem. Identifying the best arms in multiple bandits with possible heterogeneous variances was studied in the fixed budget setting (Gabillon et al., 2011), where an elimination based algorithm was proposed to take variances into designing confidence bound. In the addition to the fixed budget setting, most recently Lu et al. (2021) studied the best arm identification with unknown heterogeneous variances in the fixed confidence setting. They assumed the support of reward distribution is bounded, and proposed an elimination-based algorithm by first estimating the variances (with known upper bound on the variances) then utilizing the estimated variances in identifying the unique best arm based on Bernstein-style confidence bounds. The algorithm achieves near-optimal instance dependent performance. In comparison, we aim to study the *worst-case sample complexity* with known variance proxies as inputs (the support of reward distribution may be unbounded), in the *top-m identification* problem. We propose an optimal algorithm with an *exact matching* lower bound, and studied the impact of variances transition from the homogeneous setting to the heterogeneous setting in terms of the parameter m.

3 Preliminary

System model: We largely follow the canonical sub-Gaussian bandit model, except the additional component related to the reward variances. A bandit instance I is represented by a set of arm indices $[n] := \{1, 2, \dots, n\}$ and the tuple of reward distributions $(\nu_1, \nu_2, \dots, \nu_n)$. For any $i \in [n]$, pulling the i-th arm returns a reward observation, which is independently sampled from distribution ν_i, where ν_i is a sub-Gaussian distribution with mean μ_i and variance proxy σ_i^2[1]. An arm is ϵ-approximate top-m if the mean reward of that arm is at least $\max^m_{i\in[n]} \mu_i - \epsilon$, where $\max^m_{i\in[n]}$ indicates the m-th largest (mean reward) value among the arms in $[n]$. With the knowledge of variance proxy values $\sigma^2_{1:n}$, but without the knowledge of mean values $\mu_{1:n}$, the agent actively learns the parameters of the sub-Gaussian bandit instance I by observing independent reward samples. When there is no ambiguity from the context, we omit "proxy" and simply refer to $\sigma^2_{1:n}$ as the reward variances.

(ϵ, δ) top-m arm identification: In the (ϵ, δ) top-m arm identification problem, the agent is required to identify some subset $R \subset [n]$ with $|R| = m$, such that, with probability at least $1 - \delta$, any arm in R is ϵ-approximate top-m.

Algorithm class: Taking the parameters $(\epsilon, \delta, m, [n], \sigma^2_{1:n})$ as input, an algorithm A deployed by the agent is represented by a tuple $(\pi_t, \rho_t)_{t\geq 1}$. During the learning process, the function π_t selects an arm in $[n]$ based on the inputs of the algorithm as well as the previous observations before time step t (i.e., the arms that were pulled). The function ρ_t decides whether to stop based on the inputs of the algorithm as well as the available observations (the current observation and the previous observations before time step t). If ρ_t decides to stop, it returns a set of arms $R^{\mathsf{A}} \subset [n]$; otherwise, the process continues. Let T^{A} be the time that the process stops, which is the number of samples observed by algorithm A. We only study the *valid* algorithms that solve the (ϵ, δ) top-m arm identification when dealing with any bandit instance.

Worst-case sample complexity: The number of samples observed by the algorithm T^{A} is a stopping time, whose expectation the agent aims to minimize. We study the *worst-case sample complexity* for (ϵ, δ) top-m arm identification, which is an intrinsic quantity that measures the difficulty of the problem, and thus independent of the algorithm and $\mu_{1:n}$. Formally, the worst-case sample complexity of the (ϵ, δ) top-m arm identification problem under algorithm inputs $(\epsilon, \delta, m, [n], \sigma^2_{1:n})$ is

$$\mathrm{SC}(\epsilon, \delta, m, [n], \sigma^2_{1:n}) := \inf_{\mathsf{A}} \sup_{I \in \mathcal{I}(\sigma^2_{1:n})} \mathbb{E}_I[T^{\mathsf{A}}], \tag{2}$$

where the infimum is taken over all valid algorithms, the supremum is taken over the instance class $\mathcal{I}(\sigma^2_{1:n})$ containing all the distribution tuples $\nu_{1:n}$ with variances $\sigma^2_{1:n}$, and the subscript I in the expectation $\mathbb{E}_I[\cdot]$ indicates that it is with respect to the bandit model I.

Measure of heterogeneity: For any positive vector $a_{1:n}$, define the entropy function as $\mathrm{Ent}(a_{1:n}) := -\sum_{j=1}^n \hat{a}_i \ln \hat{a}_i$ with $\hat{a}_i = \frac{a_j}{\sum_{i=1}^n a_i}$. It measures the heterogeneity of the vector $a_{1:n}$, and takes value within

[1] A random variable X follows some σ^2-sub-Gaussian distribution, if $\ln \mathbb{E}[e^{\lambda(X - \mathbb{E}[X])}] \leq \frac{\sigma^2\lambda^2}{2}$, $\forall \lambda \in \mathbb{R}$, and σ^2 is called the variance proxy.

$(0, \ln(n)]$. Note that the entropy function is usually defined on the probability simplex, and we had slightly abused the notation by defining it for a positive vector. In this paper, we study the worst-case sample complexity, which is gap-independent.

4 Main Result: Worst-case Sample Complexity

The main result of this work is the characterization of the worst-case sample complexity $\mathrm{SC}(\epsilon, \delta, m, [n], \sigma^2_{1:n})$. To present this result, we first introduce some additional notation. Let $\underline{\sigma} := \min_{i\in[n]} \sigma_i$, and partition $[n]$ into k disjoint subsets $G_1, \dots, G_k$, such that for any $j \in [k]$,

$$G_j := \{i \in [n] : 2^{j-1} \leq \sigma_i^2/\underline{\sigma}^2 < 2^j\}. \tag{3}$$

Define two disjoint sets

$$G^m := \cup_{j:|G_j|>2m} G_j, \quad G^l := \cup_{j:|G_j|\leq 2m} G_j, \tag{4}$$

where $|\cdot|$ denotes the cardinality of the set. For each j with $G_j \subset G^l$, let $G'_j = G_j$; for each j with $G_j \subset G^m$, select $G'_j \subset G_j$ with $|G'_j| = 2m$, and denote $G^r := \cup_{j\geq 1} G'_j$ as a subset of the arms, such that $\mathrm{Ent}\left(\sigma^2_{G^r}\right)$ is maximized. (The superscripts of G^m, G^l, G^r indicate "more", "less" and "reduced", respectively).

The worst-case sample complexity $\mathrm{SC}(\epsilon, \delta, m, [n], \sigma^2_{1:n})$ is summarized in the following theorem.

Theorem 4.1. *Suppose $n > 2m$, $\epsilon > 0$ and $0 < \delta < 0.1$, then the worst-case sample complexity is*

$$\mathrm{SC}(\epsilon, \delta, m, [n], \sigma^2_{1:n}) =$$

$$\Theta\left(\sum_{i\in[n]} \frac{\sigma_i^2}{\epsilon^2} \ln\frac{1}{\delta} + \sum_{i\in G^m} \frac{\sigma_i^2}{\epsilon^2} \ln(m) + \sum_{j\in G^l} \frac{\sigma_j^2}{\epsilon^2} \mathrm{Ent}(\sigma^2_{G^r})\right). \tag{5}$$

The following lemma upper bounds the entropy $\mathrm{Ent}(\sigma^2_{G^r})$ in the third component.

Lemma 4.2. *For any $m \geq 2$, $\mathrm{Ent}(\sigma^2_{G^r}) \leq 8\ln(m)$.*

This lemma indicates that the worst-case sample complexity in the heterogeneous variance setting is upper bounded by $O\left(\sum_{i\in[n]} \frac{\sigma_i^2}{\epsilon^2} \ln\frac{m}{\delta}\right)$ in general. In certain sense, the heterogeneity in fact makes the problem "easier" to solve. To further illustrate this point, let us consider two special cases:

- When the variances are more homogeneous, e.g., in the extreme case $\sigma_i^2 = \sigma^2$, $\forall i \in [n]$, we have $G^m = [n]$ and $G^l = \emptyset$. Theorem 4.1 naturally degrades to the worst-case sample complexity in the homogeneous setting characterized in (Kalyanakrishnan et al., 2012), which is $\Theta\left(\frac{n\sigma^2}{\epsilon^2} \ln\frac{m}{\delta}\right)$.
- When the variances are highly heterogeneous, e.g., in the extreme case $|G_j| = 1, \forall j = 1, 2, \dots, k$, we have $G^m = \emptyset$ and $G^l = [n]$. Theorem 4.1 shows that the worst-case sample complexity is $\Theta\left(\sum_{i\in[n]} \frac{\sigma_i^2}{\epsilon^2} \ln\frac{1}{\delta}\right)$, which is independent of m.

Comparing the two cases and assuming the sum of the variances remain the same, the latter clearly has a more desirable sample complexity. The sets G^m and G^l describe the transition between the homogeneous and the heterogeneous. In the rest of this article, we present the optimal algorithm and the matching lower bound to establish Theorem 4.1.

5 Algorithms

We first revisit several existing algorithms designed mostly under the assumption of homogeneous variances. By adapting them to the heterogeneous variance case, we analyze their advantages and disadvantages. As will become clear shortly, these adapted algorithms still perform well in certain respective cases. Based on this observation, we will propose an optimal divide-and-conquer style algorithm.

5.1 Adapting Existing Algorithms

Weighted naive elimination: In this adapted algorithm, the agent simply pulls each arm-i a total of $\frac{2\sigma_i^2}{(\epsilon/2)^2} \ln\frac{1}{\omega_i}$ times, calculates the sample mean $\hat{\mu}_i$, and returns the m arms with the largest sample means. We call it "weighted" because the numbers of pulls for the arms are determined by the reward variances $\sigma^2_{1:n}$ and the confidence parameters $\omega_{1:n}$. The parameters $\omega_{1:n}$ need to be optimized in order to provide the performance guarantee, and the following lemma provides one such assignment of the optimized $\omega_{1:n}$.

Lemma 5.1. *Let $\omega_i = \delta \frac{\sigma_i^2}{\sum_{j=1}^n \sigma_j^2}$, the weighted naive elimination algorithm takes*

$$8 \sum_{i\in[n]} \frac{\sigma_i^2}{\epsilon^2} \left(\ln\frac{1}{\delta} + \mathrm{Ent}(\sigma^2_{1:n})\right) \tag{6}$$

samples, and solves the (ϵ, δ) top-m arm identification problem for any $\epsilon > 0$ and $0 < \delta < 1$.

We will use $\mathsf{WNElim}(\epsilon, \delta, m, [n], \sigma^2_{1:n})$ to denote the weighted naive elimination algorithm with the choices of $\omega_{1:n}$ in Lemma 5.1. The entropy function $\mathrm{Ent}(\sigma^2_{1:n})$ appears naturally as a multiplicative factor in the second item of Equation (37), which measures the heterogeneity of the variances. If the variance heterogeneity is high, the entropy term $\mathrm{Ent}(\sigma^2_{1:n})$ can be significantly less than $\log n$. As mentioned earlier, when $\sigma_i^2 = 2^i$, the entropy term is in fact $O(1)$, i.e., no longer a function of n and m. On the other hand, by the principal of maximum entropy (Cover, 1999), it has the maximum value $\ln(n)$ when the variances are homogeneous. Thus the weighted naive elimination algorithm will provide good performance when the arm variances are highly heterogeneous, but will lose efficiency when they are more homogeneous.

Adapted median elimination: Median Elimination ("Halving" algorithm in (Kalyanakrishnan and Stone, 2010)) is known to achieve the worst-case optimal performance in the homogeneous variance setting. One simple method to adapt it to the heterogeneous setting is to ignore the knowledge of the heterogeneity, and simply assume that all the arms have the largest variance $\max_{i\in[n]}\sigma_i^2$. The original median elimination algorithm can be applied without any change, and the expected number of samples taken is thus $O\left(\frac{n\max_{i\in[n]}\sigma_i^2}{\epsilon^2}\left(\ln\frac{1}{\delta}+\ln m\right)\right)$, as shown in (Kalyanakrishnan and Stone, 2010). In the appendix, we provide another method to adapt the median elimination algorithm, which improves the $n\max_{i\in[n]}\sigma_i^2$ term by roughly halving the sum of variances in each round.

If the variances are more homogeneous, e.g., $\sigma_i^2/\sigma_j^2 \le 2, \forall i,j \in [n]$, then $\sum_{i\in[n]}\sigma_i^2 \le n\max_{i\in[n]}\sigma_i^2 \le 2\sum_{i\in[n]}\sigma_i^2$ and the expected number of samples is $O\left(\frac{\sum_{i\in[n]}\sigma_i^2}{\epsilon^2}\left(\ln\frac{1}{\delta}+\ln m\right)\right)$. For the same example, the weighted naive elimination uses $O\left(\frac{\sum_{i\in[n]}\sigma_i^2}{\epsilon^2}\left(\ln\frac{1}{\delta}+\ln n\right)\right)$ samples. Thus this simple adaptation of the median elimination algorithm is able to perform well for the highly homogeneous case, but will induce a loss of performance for the more heterogeneous cases.

Adapting other algorithms: The adaptation of several instance dependent algorithms, such as LUCB and UGapE, is straightforward. For the problem in consideration, both algorithms require $O\left(\frac{\sum_{i\in[n]}\sigma_i^2}{\epsilon^2}\left(\ln\frac{1}{\delta}+\ln\frac{\sum_{i\in[n]}\sigma_i^2}{\epsilon^2}\right)\right)$ number of samples in expectation in the worst case. They are not worst-case optimal in the homogeneous variance setting, and certainly not in the heterogeneous variance setting since the latter is a more general setting.

5.2 The Optimal Variance-Grouped MedElim Algorithm

It was shown in the previous subsection that the weighted naive elimination algorithm and the median elimination algorithm have advantages in the respective cases. In order to retain the advantages in both algorithms, we take a "divide and conquer" approach. Recall the minimum variance is $\underline{\sigma} = \min_{i\in[n]}\sigma_i$, and the disjoint subsets $G_1,\dots,G_k$ form a partition of $[n]$, and for any $j\in[k]$,

$$G_j = \left\{i\in[n] : 2^{j-1} \le \sigma_i^2/\underline{\sigma}^2 < 2^j\right\}. \qquad (7)$$

The largest variance ratio within each subset is at most 2, while the variances among subsets are well separated. We wish to apply median elimination to each subset and select "good" arms within that subset, and then apply weighted naive elimination over all the selected "good" arms. However, the "good" arms within a subset can in fact be "bad" in terms of the overall arm set $[n]$. To see this, consider the following example instance: m arms have a mean reward ϵ, and the rest of $n-m$ arms have a mean reward $-\epsilon$. Then any ϵ-approximate top-m arms need to have mean ϵ. Suppose the subset G_1 contains $m' < m$ arms with mean ϵ and some other arms with mean $-\epsilon$. Ideally we would like to apply median elimination to find those top-m' arms with mean ϵ within G_1. However, parameter m' is not known, and we will apply median elimination on G_1 by selecting some l arms. If $l < m'$, then the returned l arms will not include all the top-m' arms in G_1, and therefore fail to identify the final top-m arms. On the other hand, if $l > m'$, then $\max^l_{i\in G_1}\mu_i = -\epsilon$. Any arm in G_1 is ranked in the top-l within G_1, and the problem is trivial to solve. The returned l arms, even though are top-l within G_1, are not guaranteed to contain those top-m' arms with mean reward ϵ.

To successfully apply the divide-and-conquer approach, we need a "blind" algorithm that returns a subset containing the approximate top-m' arms, ideally with certain graceful transition of the confidence values.

Definition 5.1. *The algorithm $A(\epsilon,\delta,m,[n],\sigma^2_{1:n})$ is said to satisfy the (ϵ,δ') top-m' condition, where $m'\le m$, if with probability at least $1-\delta'$, $\max^{m'}_{j\in R^A}\mu_j \ge \max^{m'}_{j\in[n]}\mu_i - \epsilon$.*

The condition is equivalent to the standard (ϵ,δ) top-m arm identification requirement, if $m' = m$ and $\delta' = \delta$. We first restate the median elimination algorithm presented in Algorithm 1 (the halving algorithm (Kalyanakrishnan and Stone, 2010)), with the necessary changes on the constants and the variance values taken into account (note the input $2m$).

The following lemma summarizes the sample complexity of the MedElim algorithm with the aforementioned transition in the confidence values for $m' = 1,2,\dots,m$ for the $2m$ return arms. This algorithm will be used as a building block for the variance-grouped median

Algorithm 1 $\mathsf{MedElim}(\epsilon, \delta, 2m, [n], \sigma^2_{1:n})$

```
  Initialize $S_1 = [n]$, $\ell = 1$ and $\epsilon_\ell = (\epsilon/3)\frac{3^\ell}{4^\ell}$, $\delta_\ell = \frac{\delta/4}{2^\ell}$

  while $|S_\ell| > 2m$ do
    Pull arm-$i$ $t_{i,\ell} = \frac{2\sigma_i^2}{(\epsilon_\ell/2)^2} \ln \frac{m}{\delta_\ell}$ times and calculate their sample mean $\hat{\mu}_{i,\ell}$ for each $i \in S_\ell$
    Update the candidate set as $S_{\ell+1} = \arg\max_{i \in S_\ell}^{1:\max(\lfloor |S_\ell|/2 \rfloor, 2m)} \hat{\mu}_{i,\ell}$
    Let $\ell = \ell + 1$
  Return: $S_\ell$
```

elimination algorithm given next. The proof of this lemma can be found in the appendix.

Lemma 5.2. *For any $\sigma^2_{1:n}$, if $\max_{i\in[n]} \sigma_i^2 / \min_{j \in [n]} \sigma_j^2 \leq 2$, the* MedElim *algorithm has an expected stopping time*

$$O\left(\frac{\sum_{i\in[n]} \sigma_i^2}{\epsilon^2}\left(\ln\frac{1}{\delta} + \ln(m)\right)\right). \tag{8}$$

Moreover, for any $m' \leq m$, the MedElim algorithm satisfies the $(\epsilon, \frac{m'}{m}\delta)$ top-m' condition.

Now we are in a position to provide the proposed algorithm below, which we refer to as the variance-grouped median elimination algorithm.

Algorithm 2 $\mathsf{V\text{-}MedElim}(\epsilon, \delta, m, [n], \sigma^2_{1:n})$

```
  Partition $[n]$ into groups $G_1, \ldots, G_k$ by (7)
  for $j \in 1 : k$ do
    $R_j = \mathsf{MedElim}(\epsilon/2, \delta/2, 2m, G_j, \sigma^2_{G_j})$;
  Let $G = \cup_{j=1}^k R_j$;
  $R = \mathsf{WNElim}(\epsilon/2, \delta/2, m, G, \sigma^2_G)$;
  Return: $R$
```

The performance of proposed algorithm is summarized in the following theorem.

Theorem 5.3. *The variance-grouped median elimination algorithm solves the (ϵ, δ) top-m arm identification problem for any $\epsilon > 0$ and $0 < \delta < 1$, and the expected number of samples is*

$$O\left(\sum_{i\in[n]} \frac{\sigma_i^2}{\epsilon^2} \ln\frac{1}{\delta} + \sum_{i \in G^m} \frac{\sigma_i^2}{\epsilon^2} \ln(m) + \sum_{j \in G^l} \frac{\sigma_j^2}{\epsilon^2} \mathrm{Ent}(\sigma^2_{G^r})\right). \tag{9}$$

Proof of Theorem 5.3. Without loss of generality, assume $[m]$ is the set of top-m arms. For any j with $G_j \cap [m] \neq \emptyset$ and $i \in G_j \cap [m]$, arm-i must be one of top-$|G_j \cap [m]|$ arms in G_j. Let $m'_j = |G_j \cap [m]|$ be the number of top-m arms contained in G_j. By Lemma 5.2, with probability at least $1 - \frac{m'_j}{m}\frac{\delta}{2}$,

$$\begin{aligned}\max{}^{m'_j}_{l \in R_j} \mu_l &\geq \max{}^{m'_j}_{l \in G_j \cap [m]} \mu_l - \epsilon/2 \\ &\geq \max{}^{m}_{l \in [n]} \mu_l - \epsilon/2. \end{aligned} \tag{10}$$

It implies that with probability at least $1 - \sum_{j=1}^k \frac{m'_j}{m}\frac{\delta}{2} = 1 - \frac{\delta}{2}$, there are at least $\sum_{j=1}^k m'_j = m$ arms in $G = \cup_{j=1}^k R_j$ that are $\epsilon/2$-approximate top-m. In other words, event $\max^m_{l \in G} \mu_l \geq \max^m_{l\in[n]} \mu_l - \epsilon/2$ occurs with probability at least $1 - \frac{\delta}{2}$.

Conditioned on this event occurring, Lemma 5.1 implies that with probability at least $1 - \frac{\delta}{2}$, the returned set R of the weighted naive elimination over $G = \cup_{j=1}^k R_j$ satisfies

$$\min_{l \in R} \mu_l \geq \max{}^m_{l \in G} \mu_l - \epsilon/2 \geq \max{}^m_{l\in[n]} \mu_l - \epsilon. \tag{11}$$

Thus with probability at least $1 - \delta$, all arms in R are ϵ-approximate top-m.

Recall the definition of G^l, G^m, G^r in Section 4. The total number of samples used in the median elimination subroutine is $O\left(\sum_{i \in G^m} \frac{\sigma_i^2}{\epsilon^2}\left(\ln\frac{1}{\delta} + \ln(m)\right)\right)$. The number of samples used in the weighted naive elimination subroutine is $O\left(\sum_{i\in[n]} \frac{\sigma_i^2}{\epsilon^2}\left(\ln\frac{1}{\delta} + \mathrm{Ent}(\sigma^2_{G^r})\right)\right)$. By Lemma 4.2, the expected total number of samples is $O\Big(\sum_{i\in[n]} \frac{\sigma_i^2}{\epsilon^2} \ln\frac{1}{\delta} + \sum_{i\in G^m} \frac{\sigma_i^2}{\epsilon^2} \ln(m)$ $+ \sum_{j \in G^l} \frac{\sigma_j^2}{\epsilon^2} \mathrm{Ent}(\sigma^2_{G^r})\Big)$. □

An illustrative example: In the following example, we show the number of required samples by the variance-grouped median elimination algorithm given in Theorem 5.3 achieves an order-wise improvement over $\frac{\sum_{i\in[n]}\sigma_i^2}{\epsilon^2}(\ln(1/\delta) + \mathrm{Ent}(\sigma^2_{1:n}))$ and $\frac{\sum_{i\in[n]}\sigma_i^2}{\epsilon^2}(\ln(1/\delta) + \ln(m))$. Take some integer $k \geq 2$ as an auxiliary parameter in this problem setting, and denote $\ell = \lceil \log(k) \rceil$. Let $\log(m) = k$ and $\log(n) = k^2$. We aim to approximately identify the top-m arms out of n arms. Among these n arms, there are 2^i arms with the same variance 2^{-i} for each $i = 0, 1, \ldots, \ell - 1$, and the rest $n - \sum_{i=0}^{\ell-1} 2^i = 2^{k^2} - 2^\ell + 1$ arms have the same variance $2^{-k^2}\ell/k$. Then G^m is the set of arms with variances $2^{-k^2}\ell/k$, and G^l is the set of arms with variances 2^{-i} for $i = 0, 1, \ldots, \ell - 1$. It is seen that

$$\sum_{j \in G^m} \sigma_j^2 = (2^{k^2} - 2^\ell + 1)2^{-k^2}\ell/k = \Theta(\ell/k), \tag{12}$$

$$\sum_{j \in G^l} \sigma_j^2 = \sum_{i=0}^{\ell-1} 2^i 2^{-i} = \ell = \Theta(\log(k)), \tag{13}$$

which implies $\sum_{j\in[n]}\sigma_j^2 = \Theta(\log(k))$. Furthermore, we can calculate that

$$\mathrm{Ent}(\sigma^2_{G^r}) = \Theta(\mathrm{Ent}(\sigma^2_{G^l})) = \Theta(\log(k)). \tag{14}$$

Thus the number of required samples by the variance-grouped median elimination algorithm is of order

$$\Theta(\ln(k)\ln(1/\delta) + \ln(k)^2/\epsilon^2). \tag{15}$$

Since $\mathrm{Ent}(\sigma^2_{1:n}) = \Theta(k)$ and $\ln(m) = \Theta(k)$, it is seen that $\frac{\sum_{i\in[n]}\sigma_i^2}{\epsilon^2}(\ln(1/\delta) + \mathrm{Ent}(\sigma^2_{1:n}))$ and $\frac{\sum_{i\in[n]}\sigma_i^2}{\epsilon^2}(\ln(1/\delta) + \ln(m))$ are of the same order

$$\Theta(\ln(k)\ln(1/\delta) + k\ln(k)/\epsilon^2). \tag{16}$$

The detailed calculation of the entropy values used above is given in the supplementary material. Fix $\delta > 0$ as constant, comparing the numbers of required samples in (15) and (16), which are of order $\Theta(\ln(k)^2/\epsilon^2)$ and $\Theta(k\ln(k)/\epsilon^2)$, respectively, it is seen that the variance-grouped median elimination algorithm provides an order-wise improvement in this example setting by reducing a factor k to $\ln(k)$.

Remark. Our result establishes the theoretical optimality of the proposed algorithm through a matching lower bound provided in the following section. However, the empirical performance of the proposed algorithm suffers from large multiplicative factors introduced by the Median Elimination subroutine. More aggressive elimination based algorithm, such as the algorithms proposed in (Hassidim et al., 2020), can be used as a subroutine to improve the multiplicative factor while maintaining the same order.

6 The Lower Bound

In the homogeneous variance setting, the previous lower bound (Kalyanakrishnan et al., 2012) on worst-case (ϵ, δ)-PAC top-m identification leveraged the change-of-measure technique and was proved by contradiction. The approach leads to a large multiplicative factor and is also difficult to utilize in the heterogeneous variance case. The lower bound was later tightened and generalized to the instance-dependent case in (Chen et al., 2017) and (Simchowitz et al., 2017). Their approach assumed that the algorithms have a uniform preference over the arms at the beginning, which is reasonable in the homogeneous setting but not in the heterogeneous setting.

We derive a flexible simple inequality to better take into account the heterogeneous variances, given in Lemma 6.2. Applying this lemma, we formulate the lower bound as an optimization problem, whose dual formulation (Lemma 6.3) is then studied. The eventual lower bound is given in the following theorem, obtained by considering several feasible solutions to the dual problem.

Theorem 6.1. *There exists some universal constant $c > 0$, that for any $0 < \epsilon$, $0 < \delta < 0.1$, $m < n/2$, $\sigma^2_{1:n}$ and any valid algorithm, there exists an instance with the given variances such that the expected number of samples of the algorithm is at least*

$$c\left(\sum_{i\in[n]}\frac{\sigma_i^2}{\epsilon^2}\ln\frac{1}{\delta} + \sum_{i\in G^m}\frac{\sigma_i^2}{\epsilon^2}\ln(m) + \sum_{j\in G^l}\frac{\sigma_j^2}{\epsilon^2}\mathrm{Ent}(\sigma^2_{G^r})\right). \tag{17}$$

6.1 Dual Formulation of the Lower Bound

We first introduce an inequality in the lemma below, which helps us connect the sample complexity with a multi-hypothesis testing problem.

Lemma 6.2. *For any two probability measure P, Q on the same measurable space $(\Omega, \mathcal{F})$, if $\mathcal{E} \in \mathcal{F}$ with $P(\mathcal{E}) \geq 1-\delta > Q(\mathcal{E})$, we have*

$$Q(\mathcal{E}) \geq B(\delta)e^{-\frac{D(P||Q)}{1-\delta}}, \tag{18}$$

where $D(\cdot||\cdot)$ is the Kullback-Leibler divergence and $B(\delta) = e^{-\frac{\mathrm{Ent}(\delta,1-\delta)}{1-\delta}}$ is a strictly decreasing function with $B(0.1) > 0.69$.

Fix any algorithm A with inputs $(\epsilon, \delta, m, [n], \sigma^2_{1:n})$ that solves the (ϵ, δ) top-m arm identification problem. Consider the Gaussian instances where the i-th arm has a Gaussian distribution with variance σ_i^2. Denote P_I as the probability measure induced by the learning process of applying algorithm A on Gaussian bandit instance $I \in \mathcal{I}(\sigma^2_{1:n})$.

Let $\epsilon' > \epsilon$ be some parameter that can be arbitrarily close to ϵ. For any subset $M \subset [n]$ with $|M| = m$ and any index $l \in [n] \setminus M$, we first construct an instance $I_{l,M} \in \mathcal{I}(\sigma^2_{1:n})$ by specifying the reward means of each arm as follows: the l-th arm has mean 0, the arms in M have mean ϵ', and the rest have mean $-\epsilon'$. The only ϵ-approximate top-m arms of instance $I_{l,M}$ are clearly M. Similarly, for each subset $F \subset [n]$ with $|F| = m-1$ and any index $l \in [n]\setminus F$, we then construct an instance $I_{l,F} \in \mathcal{I}(\sigma^2_{1:n})$. In instance $I_{l,F}$, the l-th arm has mean 0, the arms in F have mean ϵ', and the rest arms have mean $-\epsilon'$. The only ϵ-approximate top-m arm set of instance $I_{l,F}$ is clearly $F \cup \{l\}$. These are the possible hypotheses we will consider.

Given an instance $I_{l,M}$, if $F = M\setminus\{i\}$ for some $i \in M$, it is clear that instances $I_{l,M}$ and $I_{l,F}$ differ only at the i-th arm. Denote $t_{l,F,i}$ as the expected number of pulls of the i-th arm by algorithm A on instance $I_{l,F}$. The KL-divergence can be calculated as $D(P_{I_{l,F}}||P_{I_{l,M}}) = \frac{2\epsilon'^2}{\sigma_i^2}t_{l,F,i}$; see Lemma 5.1 in (Lattimore and Szepesvári, 2020) for more details. Since A solves the (ϵ,δ) top-m arm identification problem, we have $P_{I_{l,F}}(R^{\mathsf{A}} = F\cup\{l\}) \geq 1-\delta > \delta \geq P_{I_{l,M}}(R^{\mathsf{A}} = F\cup\{l\})$. Applying Lemma 6.2 on $P_{I_{I,F}}$, $P_{I_{l,M}}$ and event $\{R^{\mathsf{A}} = F\cup\{l\}\}$ gives

$$\begin{aligned} P_{I_{l,M}}(R^{\mathsf{A}} = F\cup\{l\}) &\geq B(\delta)e^{-\frac{D(P_{I_{l,F}}||P_{I_{l,M}})}{1-\delta}} \\ &= B(\delta)e^{-\frac{2\epsilon'^2}{\sigma_i^2}\frac{t_{l,F,i}}{1-\delta}}. \end{aligned} \tag{19}$$

This inequality holds for any $F = M\setminus\{i\}$ with $i\in M$. In addition, events $\{R^{\mathsf{A}} = M\cup\{l\}\setminus\{i\}\}$'s are disjoint for any $i\in M\cup\{l\}$, and they are also disjoint with the event $\{R^{\mathsf{A}} = M\}$. It follows that $\sum_{i\in M} P_{I_{l,M}}\left(R^{\mathsf{A}} = M\cup\{l\}\setminus\{i\}\right) \leq 1 - P_{I_{l,M}}(R^{\mathsf{A}} = M) \leq \delta$. Summing inequality (19) for all $i\in M$ gives

$$\begin{aligned} \delta &\geq \sum_{i\in M} P_{I_{l,M}}(R^{\mathsf{A}} = M\cup\{l\}\setminus\{i\}) \\ &\geq \sum_{i\in M} B(\delta)\exp\left(-\frac{2\epsilon'^2}{\sigma_i^2}\frac{t_{l,M\setminus\{i\},i}}{1-\delta}\right). \end{aligned} \tag{20}$$

In the worst-case, algorithm A takes at least $\max_{F,l\notin F}\sum_{j\notin F\cup\{l\}} t_{l,F,j}$ samples in expectation. Any valid algorithm has to satisfy (20), and thus the sample complexity $\mathrm{SC}(\epsilon,\delta,m,[n],\sigma_{1:n}^2)$ is lower bounded by the optimal value of the following optimization problem:

$$\text{minimize:} \quad \max_{F\subset[n]:|F|=m-1,\ l\notin F}\ \sum_{j\notin F\cup\{l\}} t_{l,F,j} \tag{21}$$

$$\begin{aligned} \text{subject to:} \quad & \sum_{i\in M}\exp\left(-t_{l,M\setminus\{i\},i}/\theta_i\right) \leq \delta', \\ & \forall M\subset[n], |M| = m,\ \forall l\notin M, \end{aligned} \tag{22}$$

where $\theta_i = \frac{(1-\delta)\sigma_i^2}{2\epsilon^2}, \forall i\in[n]$ and $\delta' = \frac{\delta}{B(\delta)}$. Though this problem is convex, it is difficult to solve explicitly. Therefore, we consider its (restricted) dual formulation in the following lemma.

Lemma 6.3. *For $\epsilon > 0$, $\delta < 0.25$, $m < n/2$, $(\sigma_i^2)_{i\in[n]}$, $\mathrm{SC}(\epsilon,\delta,m,[n],\sigma_{1:n}^2) \geq \frac{1-\delta}{2\epsilon^2}v^*$, where v^* is the optimal value of the following optimization problem:*

$$\begin{aligned} \textit{maximize:} \quad & \sum_{M\subset[n]:|M|=m}\left(\sum_{l\in M}\eta_{M\setminus\{l\}}\sigma_l^2\right)\times \\ & \left(\ln\frac{B(\delta)}{\delta} + \mathrm{Ent}(\{\eta_{M\setminus\{l\}}\sigma_l^2\}_{l\in M})\right) \end{aligned} \tag{23}$$

$$\begin{aligned} \textit{subject to:} \quad & \sum_{F\subset[n]:|F|=m-1}\eta_F = 1, \\ & \eta_F \geq 0,\ \forall F\subset[n], |F| = m-1. \end{aligned} \tag{24}$$

Though the dual formulation is still difficult to solve, by the weak duality, we can derive lower bounds for the primal problem by assigning specific feasible values to the dual variables η_F's. In addition, each η_F is a probability mass function and has a clear operational meaning, which is the worst-case prior distribution of the underlining instance being one of $\{I_{l,F}\}_{l\notin F}$.

6.2 Dichotomy of the lower bound

As shown in Theorem 6.1, the lower bound of the sample complexity consists of three terms

$$\underbrace{\sum_{i\in[n]}\frac{\sigma_i^2}{\epsilon^2}\ln\frac{1}{\delta}}_{\text{I}} + \underbrace{\sum_{i\in G^m}\frac{\sigma_i^2}{\epsilon^2}\ln(m)}_{\text{II}} + \underbrace{\sum_{j\in G^l}\frac{\sigma_j^2}{\epsilon^2}\mathrm{Ent}(\sigma_{G^r}^2)}_{\text{III}}. \tag{25}$$

We will discuss each term from the viewpoint of the dual formulation in Lemma 6.3. The optimal value v^* of the optimization in Lemma 6.3 can be lower bounded by the average of the objective function values v_1, v_2, v_3 when assigning the variables certain feasible values in the dual optimization problem, i.e., $v^* = \Omega(v_1+v_2+v_3)$. We construct three sets of feasible dual variables η_F's, the resultant values $v_{1:3}$ will induce Term I-III, respectively.

It is straightforward to see that Term I can be obtained by assigning η_F's uniformly, and thus we can focus on Term II and Term III. More precisely, we aim to lower bound the optimal value of the following optimization problem:

$$\begin{aligned} \text{maximize:} \quad & \sum_{M\subset[n]:|M|=m}\left(\sum_{l\in M}\eta_{M\setminus\{l\}}\sigma_l^2\right)\times \\ & \mathrm{Ent}(\{\eta_{M\setminus\{l\}}\sigma_l^2\}_{l\in M}) \end{aligned} \tag{26}$$

$$\begin{aligned} \text{subject to:} \quad & \sum_{F\subset[n]:|F|=m-1}\eta_F = 1, \\ & \eta_F \geq 0,\ \forall F\subset[n], |F| = m-1. \end{aligned} \tag{27}$$

Firstly, to study the sample complexity induced by $\sigma_{G^m}^2$, we specify a feasible assignment of dual variables η_F's as follows. For any $F\subset G^m$ with $|F| = m-1$, let $\eta_F = \frac{\prod_{i\in F}\sigma_i^2}{\sum_{F'\subset G^m:|F'|=m-1}\prod_{j\in F}\sigma_i^2}$; and for any $F\not\subset G^m$ with $|F| = m-1$, set $\eta_F = 0$. Then $\mathrm{Ent}(\{\eta_{M\setminus\{l\}}\sigma_l^2\}_{l\in M}) = \ln(m)$ for any $M\subset G^m$ with $|M| = m$. Formally, Term II is the introduced by the following lemma.

Lemma 6.4. *The optimal value of the optimization (26) is lower-bounded by $\frac{1}{3}\sum_{j\in G^m}\sigma_j^2\ln(m)$.*

Secondly, to study the complexity induced by $\sigma_{G^l}^2$, we consider the reduced arm set $G^r\supset G^l$. Define $L\subset G^r$

with $|L| = 2m$ as the arms with $2m$ largest variances in G^r. We can verify that $\sum_{i\in L}\sigma_i^2$ dominates $\sum_{j\in G^r}\sigma_j^2$. Moreover, $\mathrm{Ent}(\sigma^2_{G^r})$ and $\mathrm{Ent}(\sigma^2_L)$ behave similarly, and thus we can focus on the arms in L. Rigorously, the following lemma justifies this choice.

Lemma 6.5. *Let* $\eta_F = \binom{2m}{m-1}^{-1}$ *for any* $F \subset L$ *with* $|F| = m-1$ *and* $\eta_F = 0$ *otherwise. The objective function of the optimization problem (26) is at least* $c'\sum_{i\in G^l}\sigma_i^2\mathrm{Ent}(\sigma^2_{G^r}) - \ln(2)\sum_{i\in L}\sigma_i^2$*, for some constant* $c' > 0$.

The first item in Lemma 6.5 is exactly Term III, and the second item $-\ln(2)\sum_{i\in G^l}\sigma_i^2$ can be absorbed into Term I.

7 Conclusion

We study the worst-case sample complexity of (ϵ,δ) top-m arm identification problem with heterogeneous reward variances. The heterogeneity of reward variances is measured by certain entropy-like function. We propose the variance-grouped median elimination algorithm, which combines the advantages of the median elimination algorithm and the weighted naive elimination algorithm in a divide-and-conquer manner. Matching lower bound of the worst-case sample complexity was devised using a dual formulation and finding suitable feasible solutions.

References

Assadi, S. and Wang, C. (2020). Exploration with limited memory: streaming algorithms for coin tossing, noisy comparisons, and multi-armed bandits. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 1237–1250.

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

Bechhofer, R. E. (1954). A single-sample multiple decision procedure for ranking means of normal populations with known variances. *The Annals of Mathematical Statistics*, pages 16–39.

Bubeck, S. and Cesa-Bianchi, N. (2012). Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *arXiv preprint arXiv:1204.5721*.

Bubeck, S., Munos, R., and Stoltz, G. (2009). Pure exploration in multi-armed bandits problems. In *International conference on Algorithmic learning theory*, pages 23–37. Springer.

Bubeck, S., Wang, T., and Viswanathan, N. (2013). Multiple identifications in multi-armed bandits. In *International Conference on Machine Learning*, pages 258–265. PMLR.

Carpentier, A. and Locatelli, A. (2016). Tight (lower) bounds for the fixed budget best arm identification bandit problem. In *Conference on Learning Theory*, pages 590–604. PMLR.

Chaudhuri, A. R. and Kalyanakrishnan, S. (2019). Pac identification of many good arms in stochastic multi-armed bandits. In *International Conference on Machine Learning*, pages 991–1000. PMLR.

Chen, L. and Li, J. (2015). On the optimal sample complexity for best arm identification. *arXiv preprint arXiv:1511.03774*.

Chen, L., Li, J., and Qiao, M. (2017). Nearly instance optimal sample complexity bounds for top-k arm selection. In *Artificial Intelligence and Statistics*, pages 101–110. PMLR.

Cover, T. M. (1999). *Elements of information theory*. John Wiley & Sons.

Even-Dar, E., Mannor, S., and Mansour, Y. (2002). PAC bounds for multi-armed bandit and markov decision processes. In *International Conference on Computational Learning Theory*, pages 255–270. Springer.

Even-Dar, E., Mannor, S., and Mansour, Y. (2006). Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of machine learning research*, 7(Jun):1079–1105.

Faella, M., Finzi, A., and Sauro, L. (2020). Rapidly finding the best arm using variance. In *Proc. of ECAI, 24th European Conference of Artificial Intelligence*.

Gabillon, V., Ghavamzadeh, M., and Lazaric, A. (2012). Best arm identification: A unified approach to fixed budget and fixed confidence. *Advances in Neural Information Processing Systems*, 25:3212–3220.

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

Garivier, A. and Kaufmann, E. (2016). Optimal best arm identification with fixed confidence. In *Conference on Learning Theory*, pages 998–1027.

Hassidim, A., Kupfer, R., and Singer, Y. (2020). An optimal elimination algorithm for learning a best arm. *arXiv preprint arXiv:2006.11647*.

Jamieson, K., Malloy, M., Nowak, R., and Bubeck, S. (2014). lil'ucb: An optimal exploration algorithm for multi-armed bandits. In *Conference on Learning Theory*, pages 423–439.

Jamieson, K. and Nowak, R. (2014). Best-arm identification algorithms for multi-armed bandits in the fixed confidence setting. In *2014 48th Annual Conference on Information Sciences and Systems (CISS)*, pages 1–6. IEEE.

Jin, T., Shi, J., Xiao, X., and Chen, E. (2019). Efficient pure exploration in adaptive round model. *Advances in Neural Information Processing Systems*, 32:6609–6618.

Kalyanakrishnan, S. and Stone, P. (2010). Efficient selection of multiple bandit arms: Theory and practice. In *ICML*.

Kalyanakrishnan, S., Tewari, A., Auer, P., and Stone, P. (2012). PAC subset selection in stochastic multi-armed bandits. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 227–234.

Karnin, Z., Koren, T., and Somekh, O. (2013). Almost optimal exploration in multi-armed bandits. In *International Conference on Machine Learning*, pages 1238–1246.

Katz-Samuels, J. and Jamieson, K. (2020). The true sample complexity of identifying good arms. In *International Conference on Artificial Intelligence and Statistics*, pages 1781–1791. PMLR.

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

Kaufmann, E. and Kalyanakrishnan, S. (2013). Information complexity in bandit subset selection. In *Conference on Learning Theory*, pages 228–251. PMLR.

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

Lattimore, T. and Szepesvári, C. (2020). *Bandit algorithms*. Cambridge University Press.

Lu, P., Tao, C., and Zhang, X. (2021). Variance-dependent best arm identification. *arXiv preprint arXiv:2106.10417*.

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

Shen, C. (2019). Universal best arm identification. *IEEE Transactions on Signal Processing*, 67(17):4464–4478.

Simchowitz, M., Jamieson, K., and Recht, B. (2017). The simulator: Understanding adaptive sampling in the moderate-confidence regime. In *Conference on Learning Theory*, pages 1794–1834. PMLR.

Slivkins, A. (2019). Introduction to multi-armed bandits. *arXiv preprint arXiv:1904.07272*.

Zhou, Y., Chen, X., and Li, J. (2014). Optimal pac multiple arm identification with applications to crowdsourcing. In *International Conference on Machine Learning*, pages 217–225. PMLR.

Supplementary Material: Approximate Top-m Arm Identification with Heterogeneous Reward Variances

A Proofs for Section 4

We will need the following well known inequality frequently.

Lemma A.1 (Hoeffding's inequality)**.** *Let $X_{1:n}$ be n independent random variables follow some σ^2-sub-Gaussian distribution with mean μ. Let $\hat{\mu}$ be their sample mean. Then the following inequalities hold*

$$\mathbb{P}\left(\hat{\mu} - \mu \geq \epsilon\right) \leq e^{-\frac{\epsilon^2 n}{2\sigma^2}}, \quad \mathbb{P}\left(\hat{\mu} - \mu \leq -\epsilon\right) \leq e^{-\frac{\epsilon^2 n}{2\sigma^2}}. \tag{28}$$

Lemma A.2 (Restate Lemma 4.2)**.** *For any $m \geq 2$, $\mathrm{Ent}(\sigma^2_{G^r}) \leq 8\ln(m)$.*

Proof of Lemma 4.2. For any choice of $\sigma^2_{1:n}$. Let $s_j = \sum_{i \in G'_j} \sigma_i^2$ for each $i = 1, \ldots, k$. By the grouping property of entropy, we have

$$\begin{aligned} \mathrm{Ent}(\sigma^2_{G^r}) &= \mathrm{Ent}(s_{1:k}) + \sum_{j=1}^{k} \frac{s_j}{\sum_{i=1}^{k} s_i} \mathrm{Ent}(\sigma^2_{G'_j}) && (29)\\ &\leq \mathrm{Ent}(s_{1:k}) + \ln(2m), && (30) \end{aligned}$$

where the inequality is due to the principal of maximum entropy.

For $j = 1, \ldots, k$, if $|G'_j| > 0$, we have $2^{j-1} \leq s_j/\underline{\sigma}^2 < 2m2^j$, otherwise $s_j = 0$. Without loss of generality, assume $\underline{\sigma}^2 = 1$ and $s_k > 0$. Let $s_{1:k}$ be the assignment with the largest entropy $\mathrm{Ent}(s_{1:k})$. If there are only $2m$ non-zero $s_{1:k}$, we have $\mathrm{Ent}(s_{1:k}) \leq \ln(2m)$ and the lemma is already proved. When there are more than $2m$ non-zero $s_{1:k}$, we have

$$\sum_{j=1}^{k-2m+1} s_j \leq 2m \sum_{j=1}^{k-2m+1} 2^j = 4m(2^{k-2m+1} - 1) < 4m2^{k-2m+1}, \tag{31}$$

and $s_k \geq 2^{k-1}$. It follows that

$$\begin{aligned} \sum_{j=1}^{k-2m+1} s_j &= \frac{\sum_{j=1}^{k-2m+1} s_j}{\sum_{i=k-2m+2}^{k} s_i + \sum_{j=1}^{k-2m+1} s_j} \sum_{j=1}^{k} s_j && (32)\\ &\leq \frac{\sum_{j=1}^{k-2m+1} s_j}{s_k + \sum_{j=1}^{k-2m+1} s_j} \sum_{j=1}^{k} s_j < \frac{4m2^{k-2m+1}}{2^{k-1} + 4m2^{k-2m+1}} \sum_{j=1}^{k} s_j && (33)\\ &= \frac{4m2^{-2m+2}}{1 + 4m2^{-2m+2}} \sum_{j=1}^{k} s_j. && (34) \end{aligned}$$

We can then write

$$\begin{aligned} \mathrm{Ent}(s_{1:k}) &= \mathrm{Ent}(\sum_{j=1}^{k-2m+1} s_j, s_{k-2m+2:k}) + \frac{\sum_{j=1}^{k-2m+1} s_j}{\sum_{j=1}^{k} s_j} \mathrm{Ent}(s_{1:k-2m+1}) && (35)\\ &\leq \ln(2m) + \frac{4m2^{-2m+2}}{1 + 4m2^{-2m+2}} \mathrm{Ent}(s_{1:k}), && (36) \end{aligned}$$

where the equality is by the grouping property of entropy function, and the inequality is by $\mathrm{Ent}(s_{1:k-2m+1}) \le \mathrm{Ent}(s_{1:k})$ since $s_{1:k}$ is the optimal assignment in terms of the largest entropy with k subsets, thus assignment $s_{1:k-2m+1}$ has smaller entropy. It implies $\mathrm{Ent}(s_{1:k}) \le (1 + 4m2^{-2m+2})\ln(2m) \le 3\ln(2m)$. We thus have $\mathrm{Ent}(\sigma^2_{G^r}) \le 4\ln(2m) \le 8\ln(m)$. □

B Proofs for Section 5

Lemma B.1 (Restate Lemma 5.1). *Let* $\omega_i = \delta\frac{\sigma_i^2}{\sum_{j=1}^n \sigma_j^2}$, *the weighted naive elimination algorithm takes*

$$8\sum_{i\in[n]}\frac{\sigma_i^2}{\epsilon^2}\left(\ln\frac{1}{\delta} + \mathrm{Ent}(\sigma^2_{1:n})\right) \tag{37}$$

samples, and solves the (ϵ,δ) *top-*m *arm identification problem for any* $\epsilon > 0$ *and* $0 < \delta < 1$.

Proof of Lemma 5.1. The stopping time is clearly

$$\sum_{i=1}^n \frac{2\sigma_i^2}{(\epsilon/2)^2}\ln\frac{1}{\omega_i} = 8\frac{\sum_{i=1}^n\sigma_i^2}{\epsilon^2}\left(\ln\frac{1}{\delta} + \mathrm{Ent}(\sigma^2_{1:n})\right). \tag{38}$$

After the arms have been pulled and the reward observations collected, by Hoeffding's inequality (Lemma A.1), we have $\mathbb{P}(\hat\mu_i \le \mu_i - \epsilon/2) \le \omega_i$ for any $i \in [m]$ and $\mathbb{P}(\hat\mu_j \ge \mu_j + \epsilon/2) \le \omega_j$ for any $j \in [n]\setminus[m]$. Since $\sum_{i\in[n]}\omega_j = \delta$, the union bound implies that the event $\mathcal{E} = \{\hat\mu_i > \mu_i - \epsilon/2, \forall i \in [m]\} \cap \{\hat\mu_j < \mu_j + \epsilon/2, \forall j \in [n]\setminus[m]\}$ occurs with probability at least $1-\delta$.

Suppose event $\mathcal{E}$ occurs. Consider a threshold $\mu_m - \epsilon/2$. Firstly, for any $i \in [m]$, $\hat\mu_i > \mu_i - \epsilon/2 \ge \mu_m - \epsilon/2$. In addition, any $j \in [n]/[m]$ with $\hat\mu_j > \mu_m - \epsilon/2$ must satisfy $\mu_j + \epsilon/2 > \hat\mu_j > \mu_m - \epsilon/2$, which implies $\mu_j > \mu_m - \epsilon$, i.e., the j-th arm is ϵ-approximate top-m. In other words, any arm with a sample mean greater than the threshold $\mu_m - \epsilon/2$ must be ϵ-approximate top-m. Since there are at least m arms with sample means greater than $\mu_m - \epsilon/2$, the m selected arms must be ϵ-approximate top-m. □

Lemma B.2 (Restate Lemma 5.2). *For any* $\sigma^2_{1:n}$, *if* $\max_{i\in[n]}\sigma_i^2/\min_{j\in[n]}\sigma_j^2 \le 2$, *the* MedElim *algorithm has an expected stopping time*

$$O\left(\frac{\sum_{i\in[n]}\sigma_i^2}{\epsilon^2}\left(\ln\frac{1}{\delta} + \ln(m)\right)\right). \tag{39}$$

Moreover, for any $m' \le m$, *the MedElim algorithm satisfies the* $(\epsilon, \frac{m'}{m}\delta)$ *top-*m' *condition.*

Proof of Lemma 5.2. We study the stopping time and accuracy separately.

Stopping time analysis: Recall that $\overline{r} = \frac{\max_{i\in[n]}\sigma_i^2}{\min_{j\in[n]}\sigma_j^2}$. It is clear that the size of the candidate set $\mathcal{S}_\ell$ decreases as $|\mathcal{S}_\ell| \le \frac{n}{2^{\ell-1}}$. The sum of variances in the candidate set $\mathcal{S}_\ell$ decreases as follows

$$\frac{\sum_{i\in\mathcal{S}_\ell}\sigma_i^2}{\sum_{j\in[n]}\sigma_j^2} \le \frac{\sum_{i\in\mathcal{S}_\ell}\overline{r}\underline{\sigma}^2}{\sum_{j\in[n]}\underline{\sigma}^2} \le \overline{r}\frac{|\mathcal{S}_\ell|}{n} \le \frac{\overline{r}}{2^{\ell-1}}. \tag{40}$$

This implies that

$$\frac{\sum_{i\in S_\ell}\sigma_i^2}{(\epsilon_\ell/2)^2} = 36\frac{16^\ell}{9^\ell}\frac{\sum_{i\in S_\ell}\sigma_i^2}{\epsilon^2} \le 72\overline{r}\frac{8^\ell}{9^\ell}\frac{\sum_{i=1}^n\sigma_i^2}{\epsilon^2}. \tag{41}$$

The (random) total number of samples is thus upper bounded by

$$\sum_{\ell=1}^\infty\sum_{i\in S_\ell} t_{i,\ell} = \sum_{\ell=1}^\infty\frac{2\sum_{i\in S_\ell}\sigma_i^2}{(\epsilon_\ell/2)^2}\ln\left(\frac{m}{\delta_\ell}\right) \tag{42}$$

$$\le \bar{r}\frac{144\sum_{i=1}^{n}\sigma_i^2}{\epsilon^2}\sum_{\ell=1}^{\infty}\frac{8^\ell}{9^\ell}\left(\ell\ln(2)+\ln\frac{4m}{\delta}\right) \tag{43}$$

$$= O\left(\bar{r}\frac{\sum_{i=1}^{n}\sigma_i^2}{\epsilon^2}\left(\ln\frac{1}{\delta}+\ln(m)\right)\right), \tag{44}$$

with probability one. Thus the expected stopping time is of order $O\left(\tilde{r}\frac{\sum_{i\in[n]}\sigma_i^2}{\epsilon^2}\left(\ln\frac{1}{\delta}+\ln(m)\right)\right)$.

Accuracy analysis. Take an arbitrary $\ell \ge 1$ with $|\mathcal{S}_\ell| > 2m$. Fix some $m' \le m$. Let $1_\ell, 2_\ell, \dots, m'_\ell$ be the indices of the top-m' arms in S_ℓ obtained in iteration-$(\ell-1)$. For any $i \in [m']$, by Hoeffding's inequality (Lemma A.1), we have $\mathbb{P}(\hat{\mu}_{i_\ell,\ell} > \mu_{i_\ell} - \epsilon_\ell/2) \ge 1 - \frac{1}{m}\delta_\ell$. Define the event $\mathcal{E}_\ell = \{\forall i \in [m'],\ \hat{\mu}_{i_\ell,\ell} > \mu_{i_\ell} - \epsilon_\ell/2\}$. By applying the union bound over $i \in [m']$, it is straightforward to verify that $\mathbb{P}(\mathcal{E}_\ell) \ge 1 - \frac{m'}{m}\delta_\ell$.

Conditioned on event $\mathcal{E}_\ell$ occurring, consider a threshold $\mu_{m'_\ell} - \epsilon_\ell/2$. It is clear that for any $i \in [m']$, $\hat{\mu}_{i_\ell,\ell} > \mu_{i_\ell} - \epsilon/2 \ge \mu_{m'_\ell} - \epsilon/2$. Thus any arm in $\{1_\ell, \dots, m'_\ell\}$ has an empirical mean greater than the threshold $\mu_{m'_\ell} - \epsilon_\ell/2$. In iteration-$\ell$, $|\mathcal{S}_{\ell+1}|$ arms with the largest empirical means are selected from set $\mathcal{S}_\ell$.

- If the selected arm with the smallest sample mean $\min\{\hat{\mu}_{i,\ell} : \ i \in \mathcal{S}_{\ell+1}\}$ is less than or equal to the threshold, then all the arms in $\{1_\ell, \dots, m'_\ell\}$ must be selected and they are still the top-m' arms within $\mathcal{S}_{\ell+1}$. It implies that $\mu_{m'_{\ell+1}} = \mu_{m'_\ell} > \mu_{m'_\ell} - \epsilon_\ell$.
- On the other hand, if the selected arm with the smallest sample mean is greater than the threshold, some arms in $\{1_\ell, \dots, m'_\ell\}$ may not be selected. Define the set of bad arms $B_\ell := \{i \in \mathcal{S}_\ell : \ \mu_i < \mu_{m'_\ell} - \epsilon_\ell\}$. A bad arm will be selected only if its empirical mean is greater than the threshold. Denote the set of bad arms with such overestimated sample means as $N_{m',\ell} = \{j \in B_\ell : \ \hat{\mu}_{j,\ell} > \mu_{m'_\ell} - \epsilon_\ell/2\}$. Then there are at most $|N_{m',\ell}|$ bad arms in $\mathcal{S}_{\ell+1}$. If $|N_{m',\ell}| \le |\mathcal{S}_{\ell+1}| - m'$, at least m' good arms remain in $\mathcal{S}_{\ell+1}$, which guarantees $\mu_{m'_{\ell+1}} \ge \mu_{m'_\ell} - \epsilon_\ell$.

These two situations indicate that conditioned on $\mathcal{E}_\ell$, $|N_{m',\ell}| \le |\mathcal{S}_{\ell+1}| - m'$ implies $\mu_{m'_{\ell+1}} \ge \mu_{m'_\ell} - \epsilon_\ell$. It follows that

$$\begin{aligned}
\mathbb{P}\left(\mu_{m'_{\ell+1}} < \mu_{m'_\ell} - \epsilon_\ell | \mathcal{E}_\ell\right) &\le \mathbb{P}\left(|N_{m',\ell}| \ge |S_{\ell+1}| - m' + 1 | \mathcal{E}_\ell\right) \\
&\le \frac{\mathbb{E}[|N_{m',\ell}||\mathcal{E}_\ell]}{|S_{\ell+1}| - i + 1}.
\end{aligned}$$

where the second inequality is due to Markov inequality. The expectation can be bounded by

$$\begin{aligned}
\mathbb{E}[|N_{m',\ell}||\mathcal{E}_\ell] &= \sum_{j\in B_\ell}\mathbb{P}\left(\hat{\mu}_{j,\ell} > \mu_{m'_\ell} - \epsilon_\ell/2 | \mathcal{E}_\ell\right) \\
&= \sum_{j\in B_\ell}\mathbb{P}\left(\hat{\mu}_{j,\ell} > \mu_{m'_\ell} - \epsilon_\ell/2\right) \\
&\le \sum_{j\in B_\ell}\mathbb{P}\left(\hat{\mu}_{j,\ell} > \mu_j + \epsilon_\ell/2\right) \\
&\le (|S_\ell| - m')\frac{\delta_\ell}{m},
\end{aligned}$$

where the equality is because $\mathcal{E}_\ell$ is defined by the samples of arms in $[1_\ell, \dots, m'_\ell]$ which are independent from the samples of arms in B_ℓ, the first inequality is by $\mu_{m'_\ell} > \mu_j$ for $j \in B_\ell$, and the last inequality is by applying Hoeffding's inequality to each $\hat{\mu}_{j,l}, j \in B_\ell$ and $|B_\ell| \le |\mathcal{S}_\ell| - m'$. We thus have

$$\begin{aligned}
\mathbb{P}\left(\mu_{m'_{\ell+1}} < \mu_{m'_\ell} - \epsilon_\ell | \mathcal{E}_\ell\right) &\le \frac{\delta_\ell}{m}\frac{|S_\ell| - m'}{|S_{\ell+1}| - m' + 1} \\
&\le \frac{\delta_\ell}{m}\frac{|\mathcal{S}_\ell| - m}{|\mathcal{S}_{\ell+1}| - m + 1} && \text{by } m' \le m \\
&\le \frac{\delta_\ell}{m}\frac{2|\mathcal{S}_{\ell+1}| + 1 - m}{|\mathcal{S}_{\ell+1}| - m + 1} && \text{by } |\mathcal{S}_\ell| \le 2|\mathcal{S}_{\ell+1}| + 1
\end{aligned}$$

$$\begin{aligned}
&= \frac{\delta_\ell}{m}\left(2 + \frac{m-1}{|\mathcal{S}_{\ell+1}| - m + 1}\right) \\
&\le \frac{\delta_\ell}{m}\left(2 + \frac{m-1}{2m - m + 1}\right) \qquad \text{by } |\mathcal{S}_{\ell+1}| \ge 2m \\
&< \frac{3\delta_\ell}{m}.
\end{aligned}$$

It follows that

$$\begin{aligned}
\mathbb{P}\left(\mu_{m'_{\ell+1}} < \mu_{m'_\ell} - \epsilon_\ell\right) &= \mathbb{P}(\mathcal{E})\mathbb{P}\left(\mu_{m'_{\ell+1}} < \mu_{m'_\ell} - \epsilon_\ell | \mathcal{E}\right) + \mathbb{P}(\mathcal{E}^c)\mathbb{P}\left(\mu_{m'_{\ell+1}} < \mu_{m'_\ell} - \epsilon_\ell | \mathcal{E}^c\right) \\
&\le \mathbb{P}\left(\mu_{m'_{\ell+1}} < \mu_{m'_\ell} - \epsilon_\ell | \mathcal{E}\right) + \mathbb{P}(\mathcal{E}^c) \\
&\le \frac{3\delta_\ell}{m} + \frac{m'\delta_\ell}{m} \le \frac{4m'}{m}\delta_\ell.
\end{aligned}$$

The argument above holds for any $\ell \ge 1$ with $|S_\ell| > 2m$. The parameters satisfy

$$\sum_{\ell=1}^{\infty} \epsilon_\ell = \frac{\epsilon}{3}\sum_{\ell=1}^{\infty}(3/4)^\ell = \epsilon, \qquad \sum_{\ell=1}^{\infty} 4\delta_\ell = \delta \sum_{\ell=1}^{\infty}(1/2)^\ell = \delta.$$

The returned arm set is $R = \mathcal{S}_{\ell^*}$ for certain ℓ^*, and thus with probability at least $1 - \frac{m'}{m}\delta$, the final returned arm set R satisfies

$$\begin{aligned}
\max{}^{m'}_{i\in R}\mu_i &= \max{}^{m'}_{i\in \mathcal{S}_{\ell^*}}\mu_i \\
&\ge \max{}^{m'}_{i\in \mathcal{S}_{\ell^*-1}}\mu_i - \epsilon_{\ell^*-1} \\
&\ge \cdots \\
&\ge \max{}^{m'}_{i\in \mathcal{S}_1}\mu_i - \sum_{\ell=1}^{\ell^*-1}\epsilon_\ell \\
&> \max{}^{m'}_{i\in [n]}\mu_i - \epsilon.
\end{aligned}$$

The proof is thus complete. □

Calculation in the illustrative example Recall the illustrative example, where $\log(m) = k$ and $\log(n) = k^2$ for some integer $k \ge 2$ and $\ell = \lceil \log(k) \rceil$. Among these n arms, there are 2^i arms with the same variance 2^{-i} for each $i = 0, 1, \ldots, \ell - 1$, and the rest $n - \sum_{i=0}^{\ell-1} 2^i = 2^{k^2} - 2^\ell + 1$ arms have the same variance $2^{-k^2}\ell/k$. Then G^m is the set of arms with variances $2^{-k^2}\ell/k$, and G^l is the set of arms with variances 2^{-i} for $i = 0, 1, \ldots, \ell - 1$. It is seen that

$$\sum_{j\in G^m} \sigma_j^2 = (2^{k^2} - 2^\ell + 1)2^{-k^2}\ell/k = \Theta(\ell/k), \tag{45}$$

$$\sum_{j\in G^l} \sigma_j^2 = \sum_{i=0}^{\ell-1} 2^i 2^{-i} = \ell = \Theta(\log(k)), \tag{46}$$

which implies $\sum_{j\in[n]} \sigma_j^2 = \Theta(\log(k))$. Furthermore, we can calculate that

$$\mathrm{Ent}(\sigma^2_{G^l}) = \sum_{i=0}^{\ell-1} \frac{2^i 2^{-i}}{\ell} \ln(2^i) = \frac{\ln(2)}{2}(\ell - 1) = \Theta(\ell) = \Theta(\log(k)). \tag{47}$$

Furthermore, we can calculate that

$$\sum_{j\in G^r} \sigma_j^2 = 2m2^{-k^2}\ell/k + \sum_{j\in G^l}\sigma_j^2 = 2^{-k^2+1}\ell + \ell = \Theta(\ell) = \Theta(\log(k)). \tag{48}$$

Then the entropy values can be calculated as

$$\begin{aligned}\mathrm{Ent}(\sigma^2_{G^r}) &= \frac{\sum_{j\in G^r/G^l}\sigma_j^2}{\sum_{j\in G^r}\sigma_j^2}\mathrm{Ent}(\sigma^2_{G^r/G^l}) + \frac{\sum_{j\in G^l}\sigma_j^2}{\sum_{j\in G^r}\sigma_j^2}\mathrm{Ent}(\sigma^2_{G^l}) && (49)\\ &= \frac{2^{-k^2+1}\ell}{\sum_{j\in G^r}\sigma_j^2}\ln(2m) + \frac{\ell}{\sum_{j\in G^r}\sigma_j^2}\mathrm{Ent}(\sigma^2_{G^l}) && (50)\\ &= \Theta\left(2^{-k^2}k + \mathrm{Ent}(\sigma^2_{G^l})\right) && (51)\\ &= \Theta(\mathrm{Ent}(\sigma^2_{G^l})) = \Theta(\log(k)), && (52)\end{aligned}$$

and $\mathrm{Ent}(\sigma^2_{G^m}) = \Theta(k^2)$ implies

$$\begin{aligned}\mathrm{Ent}(\sigma^2_{1:n}) &= \frac{\sum_{j\in G^m}\sigma_j^2}{\sum_{j\in[n]}\sigma_j^2}\mathrm{Ent}(\sigma^2_{G^m}) + \frac{\sum_{j\in G^l}\sigma_j^2}{\sum_{j\in[n]}\sigma_j^2}\mathrm{Ent}(\sigma^2_{G^l}) && (53)\\ &= \Theta\left(\frac{\ell/k}{\log(k)}k^2 + \log(k)\right) = \Theta(k). && (54)\end{aligned}$$

C A More Adaptive Median Elimination Algorithm

Let us sort $\sigma^2_{1:n}$ in decreasing order, and denote the sorted variances as $\tilde{\sigma}^2_{1:n}$. For each $\ell \geq 1$, define $h_\ell := \max\{j \geq m : \sum_{i\in[j]}\tilde{\sigma}_i^2 \leq \frac{1}{2^{\ell-1}}\sum_{i\in[n]}\sigma_i^2\}$ if the set is not empty, otherwise $h_\ell = m$. Let $\ell^* := \min\{\ell \geq 1 : h_\ell = m\}$.

Define a ratio

$$\underline{r} := \min_{j\in[\ell^*-1]} \frac{h_{j+1}}{h_j} \tag{55}$$

Algorithm 3 Adapted-MedElim$(\sigma^2_{1:n}, m, [n], \epsilon, \delta)$

```
sInitialize $S_1 = [n]$, $\ell = 1$ and $\epsilon_\ell = (\epsilon/3)\frac{3^\ell}{4^\ell}$, $\delta_\ell = \frac{\underline{r}\delta}{2^\ell}$
for $\ell = 1, 2, \ldots, \ell^* - 1$ do
    Pull arm-$i$ $t_{i,\ell} = \frac{2\sigma_i^2}{(\epsilon_\ell/2)^2}\ln\frac{m}{\delta_\ell}$ times and calculate their sample mean $\hat{\mu}_{i,\ell}$ for each $i \in S_\ell$
    Update candidate set $S_{\ell+1} = \arg\max_{i\in S_\ell}^{1:h_{\ell+1}} \hat{\mu}_{i,\ell}$
Return: $S_{\ell^*}$
```

In the homogeneous setting, the MedElim algorithm halves the complexity of the problem if the candidate set is halved. However, it should be noted that in the heterogeneous setting, simply halving the candidate set may not be efficient since the complexity would depend on the sum of the variances, instead of the number of the candidate arms. We can instead aim to half the sum of the variances of the candidate set. This discrepancy is less pronounced when the heterogeneity is low, and thus the MedElim algorithm performs reasonably well in such cases.

Lemma C.1. *The algorithm is valid and has an expected stopping time*

$$O\left(\sum_{i\in[n]}\frac{\sigma_i^2}{\epsilon^2}\left(\ln\frac{1}{\delta} + \ln(m) + \ln\frac{1}{\underline{r}}\right)\right). \tag{56}$$

Proof of Lemma C.1. We study the stopping time and accuracy separately.

Stopping time analysis: First, notice the sum of variances in the candidate set decreases as follows:

$$\sum_{i\in\mathcal{S}_\ell}\sigma_i^2 = \frac{\sum_{i\in\mathcal{S}_\ell}\sigma_i^2}{\sum_{i\in[n]}\sigma_i^2}\sum_{i\in[n]}\sigma_i^2 \leq \frac{\sum_{i\in[h_\ell]}\tilde{\sigma}_i^2}{\sum_{i\in[n]}\sigma_i^2}\sum_{i\in[n]}\sigma_i^2 \leq \frac{1}{2^{\ell-1}}\sum_{i\in[n]}^{n}\sigma_i^2. \tag{57}$$

This implies that

$$\frac{\sum_{i\in S_\ell}\sigma_i^2}{(\epsilon_\ell/2)^2} = 36\frac{16^\ell}{9^\ell}\frac{\sum_{i\in S_\ell}\sigma_i^2}{\epsilon^2} \le 72\overline{r}\frac{8^\ell}{9^\ell}\frac{\sum_{i=1}^n\sigma_i^2}{\epsilon^2}. \tag{58}$$

The stopping time is thus upper bounded by

$$\begin{aligned}
\sum_{\ell=1}^{\infty}\sum_{i\in S_\ell} t_{i,\ell} &= \sum_{\ell=1}^{\infty}\frac{2\sum_{i\in S_\ell}\sigma_i^2}{(\epsilon_\ell/2)^2}\ln\left(\frac{m}{\delta_\ell}\right) && (59)\\
&\le \frac{144\sum_{i=1}^n\sigma_i^2}{\epsilon^2}\sum_{\ell=1}^{\infty}\frac{8^\ell}{9^\ell}\left(\ell\ln(2)+\ln\frac{m}{\delta}+\ln\frac{1}{\underline{r}}\right) && (60)\\
&= O\left(\frac{\sum_{i=1}^n\sigma_i^2}{\epsilon^2}\left(\ln\frac{1}{\delta}+\ln(m)+\ln\frac{1}{\underline{r}}\right)\right). && (61)
\end{aligned}$$

The expected stopping time is of order $O\left(\frac{\sum_{i=1}^n\sigma_i^2}{\epsilon^2}\left(\ln\frac{1}{\delta}+\ln(m)+\ln\frac{1}{\underline{r}}\right)\right)$.

Accuracy analysis. Take an arbitrary $\ell \in [\ell^*-1]$, and it is clear that $|\mathcal{S}_\ell| = h_\ell > m$. Fix some $m' \le m$. Let $1_\ell, 2_\ell, \dots, m'_\ell$ be the indices of the top-m' arms in S_ℓ, respectively. For any $i \in [m']$, by Hoeffding's inequality (Lemma A.1), we have $\mathbb{P}(\hat{\mu}_{i_\ell,\ell} > \mu_{i_\ell} - \epsilon_\ell/2) \ge 1 - \frac{1}{m}\delta_\ell$. Define the event $\mathcal{E}_\ell = \{\forall i \in [m'],\ \hat{\mu}_{i_\ell,\ell} > \mu_{i_\ell} - \epsilon_\ell/2\}$. By applying the union bound over $i \in [m']$, it is straightforward to verify that $\mathbb{P}(\mathcal{E}_\ell) \ge 1 - \frac{m'}{m}\delta_\ell$.

Conditioned on the event $\mathcal{E}_\ell$ occurring, consider a threshold $\mu_{m'_\ell} - \epsilon_\ell/2$. It is clear that for any $i \in [m']$, $\hat{\mu}_{i_\ell,\ell} > \mu_{i_\ell} - \epsilon/2 \ge \mu_{m'_\ell} - \epsilon/2$. Thus any arm in $\{1_\ell, \dots, m'_\ell\}$ has empirical mean greater than the threshold $\mu_{m'_\ell} - \epsilon_\ell/2$. $|\mathcal{S}_{\ell+1}| = h_{\ell+1}$ arms with the largest sample means are selected from set $\mathcal{S}_\ell$.

- If the smallest selected sample mean $\min\{\hat{\mu}_{i,\ell} : \ i \in \mathcal{S}_{\ell+1}\}$ is less or equal to the threshold, all arms in $\{1_\ell, \dots, m'_\ell\}$ must be selected and they are still top-m' arms within $\mathcal{S}_{\ell+1}$. It implies that $\mu_{m'_{\ell+1}} = \mu_{m'_\ell} > \mu_{m'_\ell} - \epsilon_\ell$.
- On the other hand, if the smallest selected sample mean is greater than the threshold, some arms in $\{1_\ell, \dots, m'_\ell\}$ may not be selected. Define the set of bad arms $B_\ell := \{i \in \mathcal{S}_\ell : \ \mu_i < \mu_{m'_\ell} - \epsilon_\ell\}$. A bad arm can be selected only if its empirical mean is greater than the threshold. Define the set of such overestimated bad arms as $N_{m',\ell} = \{j \in B_\ell : \ \hat{\mu}_{j,\ell} > \mu_{m'_\ell} - \epsilon_\ell/2\}$. Then there are at most $|N_{m',\ell}|$ bad arms in $\mathcal{S}_{\ell+1}$. If $|N_{m',\ell}| \le |\mathcal{S}_{\ell+1}| - m'$, at least m' good arms remain in $\mathcal{S}_{\ell+1}$, which guarantees $\mu_{m'_{\ell+1}} \ge \mu_{m'_\ell} - \epsilon_\ell$.

These two situations indicate that $|N_{m',\ell}| \le |\mathcal{S}_{\ell+1}| - m'$ implies $\mu_{m'_{\ell+1}} \ge \mu_{m'_\ell} - \epsilon_\ell$ conditioned on $\mathcal{E}_\ell$. It follows that

$$\begin{aligned}
\mathbb{P}\left(\mu_{m'_{\ell+1}} < \mu_{m'_\ell} - \epsilon_\ell | \mathcal{E}_\ell\right) &\le \mathbb{P}\left(|N_{m',\ell}| \ge |S_{\ell+1}| - m' + 1 | \mathcal{E}_\ell\right)\\
&\le \frac{\mathbb{E}[|N_{m',\ell}| | \mathcal{E}_\ell]}{|S_{\ell+1}| - i + 1}.
\end{aligned}$$

where the second inequality is by Markov inequality. The expectation can be bounded by

$$\mathbb{E}[|N_{m',\ell}| | \mathcal{E}_\ell] = \sum_{j\in B_\ell}\mathbb{P}\left(\hat{\mu}_{j,\ell} > \mu_{m'_\ell} - \epsilon_\ell/2 | \mathcal{E}_\ell\right) \le (|S_\ell| - m')\frac{\delta_\ell}{m},$$

where the inequality is by Hoeffding's inequality and $|B_\ell| \le |\mathcal{S}_\ell| - m'$. We thus have

$$\begin{aligned}
\mathbb{P}\left(\mu_{m'_{\ell+1}} < \mu_{m'_\ell} - \epsilon_\ell | \mathcal{E}_\ell\right) &\le \frac{\delta_\ell}{m}\frac{|S_\ell| - m'}{|S_{\ell+1}| - m' + 1}\\
&= \frac{\delta_\ell}{m}\frac{h_\ell - m'}{h_{\ell+1} - m' + 1}\\
&\le \frac{\delta_\ell}{m}\frac{h_{\ell+1}/\underline{r} - m'}{h_{\ell+1} - m' + 1} \qquad \text{by } h_\ell \le \frac{1}{\underline{r}}h_{\ell+1}
\end{aligned}$$

$$
\begin{aligned}
&= \frac{\delta_\ell}{m}\left(\frac{1}{\underline{r}} + \frac{(1/\underline{r}-1)m' - 1/\underline{r}}{h_{\ell+1} - m' + 1}\right) \\
&\le \frac{\delta_\ell}{m}\left(1/\underline{r} + (1/\underline{r}-1)m' - 1/\underline{r}\right) \qquad \text{by } h_{\ell+1} \ge m \ge m' \\
&= \frac{m'\delta_\ell}{m}(1/\underline{r}-1).
\end{aligned}
$$

It follows that

$$
\begin{aligned}
\mathbb{P}\left(\mu_{m'_{\ell+1}} < \mu_{m'_\ell} - \epsilon_\ell\right) &= \mathbb{P}(\mathcal{E})\mathbb{P}\left(\mu_{m'_{\ell+1}} < \mu_{m'_\ell} - \epsilon_\ell | \mathcal{E}\right) + \mathbb{P}(\mathcal{E}^c)\mathbb{P}\left(\mu_{m'_{\ell+1}} < \mu_{m'_\ell} - \epsilon_\ell | \mathcal{E}^c\right) \\
&\le \mathbb{P}\left(\mu_{m'_{\ell+1}} < \mu_{m'_\ell} - \epsilon_\ell | \mathcal{E}\right) + \mathbb{P}(\mathcal{E}^c) \\
&\le \frac{m'\delta_\ell}{m}(1/\underline{r}-1) + \frac{m'\delta_\ell}{m} = \frac{1}{\underline{r}}\frac{m'}{m}\delta_\ell.
\end{aligned}
$$

The argument above holds for any $\ell \ge 1$ with $|S_\ell| > 2m$. The parameters satisfy

$$
\sum_{\ell=1}^{\infty} \epsilon_\ell = \frac{\epsilon}{3}\sum_{\ell=1}^{\infty}(3/4)^\ell = \epsilon, \qquad \sum_{\ell=1}^{\infty}\frac{1}{\underline{r}}\delta_\ell = \delta\sum_{\ell=1}^{\infty}(1/2)^\ell = \delta.
$$

The returned arm set $R = \mathcal{S}_{\ell^*}$ for some ℓ^*. With probability at least $1 - \frac{m'}{m}\delta$, the final returned arm set R satisfies

$$
\begin{aligned}
\max{}^{m'}_{i\in R}\mu_i &= \max{}^{m'}_{i\in\mathcal{S}_{\ell^*}}\mu_i \\
&\ge \max{}^{m'}_{i\in\mathcal{S}_{\ell^*-1}}\mu_i - \epsilon_{\ell^*-1} \\
&\ge \cdots \\
&\ge \max{}^{m'}_{i\in\mathcal{S}_1}\mu_i - \sum_{\ell=1}^{\ell^*-1}\epsilon_\ell \\
&> \max{}^{m'}_{i\in[n]}\mu_i - \epsilon.
\end{aligned}
$$

□

D Proofs for Section 6

Define $\mathcal{I}(\sigma_{1:n}^2) := \{(\mu_{1:n}, \sigma_{1:n}^2) :\ \mu_{1:n} \in \mathbb{R}^n\}$. When $\sigma_{1:n}^2$ is obvious in the context, we simply write $\mathcal{I}(\sigma_{1:n}^2)$ as $\mathcal{I}$. The sample complexity of the approximate top-m identification problem under algorithm inputs $(\epsilon, \delta, m, [n], \sigma_{1:n}^2)$ is

$$
\mathrm{SC}(\epsilon, \delta, m, [n], \sigma_{1:n}^2) := \inf_{\mathsf{A}} \sup_{I\in\mathcal{I}(\sigma_{1:n}^2)} \mathbb{E}_I[T^{\mathsf{A}}], \tag{62}
$$

where the infimum is taken over all valid algorithms, the supreme is taken over the instance class $\mathcal{I}(\sigma_{1:n}^2) := \{(\mu_{1:n}, \sigma_{1:n}^2) :\ \mu_{1:n} \in \mathbb{R}^n\}$, and the subscript I in the expectation $\mathbb{E}_I[\cdot]$ indicates that it is with respect to bandit model I.

Lemma D.1 (Restate Lemma 6.2)**.** *For any two probability measure P, Q on the same measurable space $(\Omega, \mathcal{F})$, if $\mathcal{E} \in \mathcal{F}$ with $P(\mathcal{E}) \ge 1 - \delta > Q(\mathcal{E})$, we have*

$$
Q(\mathcal{E}) \ge B(\delta)e^{-\frac{D(P||Q)}{1-\delta}}, \tag{63}
$$

where $D(\cdot||\cdot)$ is the Kullback-Leibler divergence and $B(\delta) = e^{-\frac{\mathrm{Ent}(\delta,1-\delta)}{1-\delta}}$ is a strictly decreasing function with $B(0.1) > 0.69$.

Proof of Lemma 6.2. Let $D_b(p,q) = p\ln\frac{p}{q} + (1-p)\ln\frac{1-p}{1-q}$ be the binary KL-divergence. Since $P(\mathcal{E}) \ge 1-\delta$, by the data processing inequality for the KL-divergence, we have

$$D(P||Q) \ge D_b(P(\mathcal{E}), Q(\mathcal{E})) \ge D_b(1-\delta, Q(\mathcal{E})) \tag{64}$$

$$> (1-\delta)\ln\frac{1-\delta}{Q(\mathcal{E})} + \delta\ln\delta \ge (1-\delta)\ln\frac{B(\delta)}{Q(\mathcal{E})}, \tag{65}$$

where the second inequality is due to $P(\mathcal{E}) \ge 1-\delta > Q(\mathcal{E})$, and the fact that $D_b(p,q)$ is monotonically increasing in p in the range $[q,1]$ for any fixed q. We thus concludes that

$$Q(\mathcal{E}) \ge B(\delta)e^{-\frac{D(P||Q)}{1-\delta}}. \tag{66}$$

□

Lemma D.2 (Restate Lemma 6.3)**.** *For $\epsilon > 0$, $\delta < 0.25$, $m < n/2$, $(\sigma_i^2)_{i\in[n]}$,* $\mathrm{SC}(\epsilon, \delta, m, [n], \sigma^2_{1:n}) \ge \frac{1-\delta}{2\epsilon^2}v^*$, *where v^* is the optimal value of the following optimization problem:*

$$\textit{maximize:} \quad \sum_{M\subset[n]:|M|=m}\left(\sum_{l\in M}\eta_{M\setminus\{l\}}\sigma_l^2\right)\left(\ln\frac{B(\delta)}{\delta} + \mathrm{Ent}(\{\eta_{M\setminus\{l\}}\sigma_l^2\}_{l\in M})\right) \tag{67}$$

$$\textit{subject to:} \quad \sum_{F\subset[n]:|F|=m-1}\eta_F = 1, \quad \eta_F \ge 0,\ \forall F\subset[n], |F| = m-1. \tag{68}$$

Proof of Lemma 6.3. We have shown in Section 6 that $\mathrm{SC}(\epsilon, \delta, m, [n], \sigma^2_{1:n})$ is lower bounded by the optimal value of the following optimization problem:

$$\text{minimize:} \quad \max_{F\subset[n]:|F|=m-1,\ l\notin F}\ \sum_{j\notin F\cup\{l\}} t_{l,F,j} \tag{69}$$

$$\text{subject to:} \quad \sum_{i\in M}\exp\left(-t_{l,M\setminus\{i\},i}/\theta_i\right) \le \delta', \quad \forall M\subset[n], |M| = m,\ \forall l\notin M, \tag{70}$$

where $\theta_i = \frac{(1-\delta)\sigma_i^2}{2\epsilon^2}, \forall i\in[n]$ and $\delta' = \frac{\delta}{B(\delta)}$. This problem is equivalent to the following convex optimization.

$$\min_{t,\tau} \quad \tau \tag{71}$$

$$\text{s.t.} \quad \sum_{j\notin F\cup\{l\}} t_{l,F,j} \le \tau, \quad \forall F\subset[n]\setminus\{l\} : |F| = m-1, \forall l\in[n] \tag{72}$$

$$\sum_{i\in M}\exp\left(-t_{l,M\setminus\{i\},i}/\theta_i\right) \le \delta', \quad \forall M\subset[n]\setminus\{l\} : |M| = m, \forall l\in[n]. \tag{73}$$

For simplicity, we use notation $\sum_{l,F}$ and $\sum_{l,M}$ to indicate $\sum_{l\in[n]}\sum_{F\subset[n]\setminus\{i\}:|F|=m-1}$ and $\sum_{l\in[n]}\sum_{M\subset[n]\setminus\{i\}:|M|=m}$, respectively. The Lagrangian of the optimization problem above is

$$L(t,\tau,\eta,\lambda) = \tau + \sum_{l,F}\eta_{l,F}\left(\sum_{j\notin F\cup\{l\}} t_{l,F,j} - \tau\right) + \sum_{l,M}\lambda_{l,M}\left(\sum_{i\in M}\exp\left(-t_{l,M\setminus\{i\},i}/\theta_i\right) - \delta'\right) \tag{74}$$

It is straightforward to check the optimization problem satisfies Slater's condition by assigning large enough $t_{l,F,j}$ and τ values. Since the optimization problem is convex, the optimal value equals to $\sup_{\eta,\lambda}\inf_{t,\tau} L(t,\tau,\eta,\lambda)$ according to the strong duality. For the saddle point, we must have $\sum_{l,F}\eta_{l,F} = 1$, or else $\inf_{t,\tau} L(t,\tau,\eta,\lambda) = -\infty$. Decision variable τ can thus be omitted. Let $L(t,\eta,\lambda) = L(t,\tau,\eta,\lambda)$ by restricting $\sum_{l,F}\eta_{l,F} = 1$. The derivative can be calculated that

$$\frac{\mathrm{d}L(t,\eta,\lambda)}{\mathrm{d}t_{l,F,i}} = \eta_{l,F} - \frac{\lambda_{l,F\cup\{i\}}}{\theta_i}\exp(-t_{l,F,i}/\theta_i). \tag{75}$$

It implies that when $\eta_{l,F} > 0$ and $\lambda_{l,F\cup\{i\}} > 0$, $t_{l,F,i} = \theta_i \ln \frac{\lambda_{l,F\cup\{i\}}}{\eta_{l,F}\theta_i}$. Define $\ln(0) = -\infty$ and let $0 \cdot \infty = 0$. The extended real valued function $g(\eta, \lambda)$ for $\sum_{l,F} \eta_{l,F} = 1$, $\eta_{l,F} \geq 0$ and $\lambda_{l,M} \geq 0$, is

$$\begin{aligned} g(\eta,\lambda) := \inf_t L(t,\eta,\lambda) = \sum_{l,F} \eta_{l,F} \sum_{i\notin F\cup\{l\}} \theta_i \ln \lambda_{l,F\cup\{i\}} - \sum_{l,F} \eta_{l,F} \sum_{i\notin F\cup\{l\}} \theta_i \ln(\eta_{l,F}\theta_i) \\ + \sum_{l,M} \left(\sum_{i\in M} \eta_{l,M\setminus\{i\}}\theta_i - \delta' \lambda_{l,M} \right). \end{aligned} \tag{76}$$

This dual function has two set of variables, however one of them can be eliminated explicitly as follows. For fixed η's with $\sum_{l,F} \eta_{l,F} = 1$ and $\eta_{l,F} \geq 0$, the function is separable with respect to λ's, and thus we can maximize $g(\eta, \lambda)$ by optimizing each individual $\lambda_{l,M}$ separately. It is straightforward to verify that $\lambda_{l,M} = \left(\sum_{F,i:F\cup\{i\}=M} \eta_{l,F}\theta_i\right)/\delta'$.

Since η's, θ's and δ' are positive, the assignments of λ's are also positive, which satisfy the constraints in the dual program. Plug it into $g(\eta, \lambda)$, we have the induced objective as

$$\begin{aligned} g(\eta) &= \sum_{l,F} \eta_{l,F} \sum_{i\notin F\cup\{l\}} \theta_i \ln \frac{\sum_{F',i':F'\cup\{i'\}=F\cup\{i\}} \eta_{l,F}\theta_i}{\eta_{l,F}\theta_i\delta'} && (77) \\ &= \sum_{F} \sum_{i\notin F} \sum_{l\notin F\cup\{i\}} \eta_{l,F}\theta_i \ln \frac{\sum_{F',i':F'\cup\{i'\}=F\cup\{i\}} \eta_{l,F}\theta_i}{\eta_{l,F}\theta_i\delta'}. && (78) \end{aligned}$$

and the dual variables η's lie in a probability simplex.

Further constraining the problem by requiring $\eta_F := (n-m)\eta_{l,F}$ for all $l \notin F$ reduces the number of dual variables, but does not change the fact that any valid assignment of η_F's will provide a lower bound to the original primal problem. The following restricted objective will be considered:

$$\begin{aligned} g(\eta) &= \sum_{F} \sum_{i\notin F} \sum_{l\notin F\cup\{i\}} \frac{\eta_F}{n-m}\theta_i \ln \frac{\sum_{F',i':F'\cup\{i'\}=F\cup\{i\}} \eta_F\theta_i}{\eta_F\theta_i\delta'} && (79) \\ &= \sum_{F} \sum_{i\notin F} \eta_F\theta_i \ln \frac{\sum_{F',i':F'\cup\{i'\}=F\cup\{i\}} \eta_F\theta_i}{\eta_F\theta_i\delta'}. && (80) \end{aligned}$$

The optimal value of the optimization above is lower bounded by

$$\begin{aligned} \text{maximize} \quad & \sum_{M\subset[n],|M|=m} \left(\left(\sum_{j\in M} \eta_{M\setminus\{j\}}\theta_j \right) \left(\mathrm{Ent}(\{\eta_{M\setminus\{j\}}\sigma_j^2\}_{j\in M}) + \ln \frac{B(\delta)}{\delta} \right) \right) && (81) \\ \text{subject to} \quad & \sum_{F\subset[n]:|F|=m-1} \eta_F = 1, \quad \eta_F \geq 0, \ \forall F \subset [n], |F| = m-1. && (82) \end{aligned}$$

The lemma is proved. □

Recall that the optimization in (26) is

$$\begin{aligned} \text{maximize:} \quad & \sum_{M\subset[n]:|M|=m} \left(\sum_{l\in M} \eta_{M\setminus\{l\}}\sigma_l^2 \right) \mathrm{Ent}(\{\eta_{M\setminus\{l\}}\sigma_l^2\}_{l\in M}) && (83) \\ \text{subject to:} \quad & \sum_{F\subset[n]:|F|=m-1} \eta_F = 1, \quad \eta_F \geq 0, \ \forall F \subset [n], |F| = m-1. && (84) \end{aligned}$$

Lemma D.3 (Restate Lemma 6.4). *The optimal value of the optimization (83) is lower-bounded by* $\frac{1}{3}\sum_{j\in G^m} \sigma_j^2 \ln(m)$.

Proof of Lemma 6.4. The objective function of equation (83) can be written as

$$\begin{aligned}\sum_{F\subset[n]:|F|=m-1}\sum_{i\notin F}\eta_F\sigma_i^2\ln\left(\frac{\sum_{F'\cup\{j\}=F\cup\{i\}}\eta_{F'}\sigma_j^2}{\eta_F\sigma_i}\right)\\ =\sum_{i\in[n]}\sum_{F:i\notin F}\eta_F\sigma_i^2\ln\left(\frac{\sum_{F'\cup\{j\}=F\cup\{i\}}\eta_{F'}\sigma_j^2}{\eta_F\sigma_i}\right)\end{aligned}\tag{85}$$

For any $F\subset G^m$ with $|F|=m-1$, let $\eta_F=\frac{\prod_{i\in F}\sigma_i^2}{\sum_{F'\subset G^m:|F'|=m-1}\prod_{j\in F}\sigma_i^2}$; and for any $F\not\subset G^m$ with $|F|=m-1$, set $\eta_F=0$. In the following analysis F indicates subset of G^m with $|F|=m-1$ and E indicates subset of G^m with $|E|=m-2$. Items in (85) can be lower bounded as follows.

$$\ln(m)\sum_{i\in G^m}\sigma_i^2\sum_{F:i\notin F}\eta_F\tag{86}$$

$$=\ln(m)\sum_{i\in G^m}\sigma_i^2\frac{\sum_{F:i\notin F}\prod_{l\in F}\sigma_l^2}{\sum_{F:i\in F}\prod_{l\in F}\sigma_l^2+\sum_{F:i\notin F}\prod_{l\in F}\sigma_l^2}\tag{87}$$

$$=\ln(m)\sum_i\sigma_i^2\left(\frac{\sum_{F:i\in F}\prod_{l\in F}\sigma_l^2}{\sum_{F:i\notin F}\prod_{l\in F}\sigma_l^2}+1\right)^{-1}\tag{88}$$

$$=\ln(m)\sum_{i\in G^m}\sigma_i^2\left((m-1)\frac{\sum_{F:i\in F}\prod_{l\in F}\sigma_l^2}{(m-1)\sum_{F:i\notin F}\prod_{l\in F}\sigma_l^2}+1\right)^{-1}\tag{89}$$

$$=\ln(m)\sum_{i\in G^m}\sigma_i^2\left((m-1)\sigma_i^2\frac{\sum_{E:i\notin E}\prod_{l\in E}\sigma_l^2}{\sum_{E:i\notin E}\prod_{l\in E}\sigma_l^2\left(\sum_{j\in G^m\setminus E\setminus\{i\}}\sigma_j^2\right)}+1\right)^{-1}\tag{90}$$

$$\geq\ln(m)\sum_{i\in G^m}\sigma_i^2\left((m-1)\sigma_i^2\frac{\sum_E\prod_{l\in E}\sigma_l^2}{\sum_E\prod_{l\in E}\sigma_l^2\left(\min_{F:i\in F}\sum_{j\in G^m\setminus F}\sigma_j^2\right)}+1\right)^{-1}\tag{91}$$

$$=\ln(m)\sum_{i\in G^m}\sigma_i^2\left(\frac{(m-1)\sigma_i^2}{\sum_{j\in G^m}\sigma_j^2-\max_{F:i\in F}\sum_{l\in F}\sigma_l^2}+1\right)^{-1},\tag{92}$$

where the last inequality is by $\sum_{j\in G^m\setminus E\setminus\{i\}}\sigma_j^2\geq\min_{F:i\in F}\sum_{j\in G^m\setminus F}\sigma_j^2$ for any $\mathcal{E}$. Recall the definition of G^m: there are $G_{1:k}$ groups partitioning $[n]$ and $G^m=\cup_{j:|G_j|>2m}G_j$. Consider the group $G_{k'}\subset G^m$ with the largest index $k'\leq k$. Since the heterogeneity within group $G_{k'}$ is at most 2, we have $\max_{i\in G^m}\sigma_i^2\leq 2\sigma_j^2$ for any $j\in G_{k'}$. Then for any $F\subset G^m$ and any $i\in G^m$,

$$\frac{(m-1)\sigma_i^2}{\sum_{j\in G^m}\sigma_j^2-\sum_{l\in F}\sigma_l^2}=\frac{(m-1)\sigma_i^2}{\sum_{j\in G^m\setminus F}\sigma_j^2}\leq\frac{(m-1)\sigma_i^2}{\sum_{j\in G_{k'}\setminus F}\sigma_j^2}\leq\frac{2(m-1)}{|G_{k'}\setminus F|}\leq\frac{2(m-1)}{m+1}<2,\tag{93}$$

where the first inequality is by $G_{k'}\subset G^m$, the second inequality is by $\sigma_i^2\leq 2\sigma_j^2$ for any $j\in G_{k'}$, and the third inequality is by $|G_{k'}|>2m$. It follows that

$$\ln(m)\sum_i\sigma_i^2\left(\frac{(m-1)\sigma_i^2}{\sum_{j\in G^m}\sigma_j^2-\max_{F:i\in F}\sum_{l\in F}\sigma_l^2}+1\right)^{-1}\tag{94}$$

$$\geq\ln(m)\sum_i\sigma_i^2\,(2+1)^{-1}=\frac{1}{3}\ln(m)\sum_j\sigma_j^2.\tag{95}$$

□

Lemma D.4. *There exists some constant $0<c'<1$, that for any choices of $\sigma_{1:n}^2$, $\mathrm{Ent}(\sigma_L^2)\geq c'\mathrm{Ent}(\sigma_{G^r}^2)-c'\ln(2)$.*

Proof of Lemma D.4. By the grouping property of entropy, we have

$$\begin{aligned}
\mathrm{Ent}(\sigma^2_{G^r}) &= \mathrm{Ent}(\sum_{j\in L}\sigma_j^2, \sum_{i\in G^r\setminus L}\sigma_i^2) && (96)\\
&\quad + \frac{\sum_{i\in L}\sigma_i^2}{\sum_{j\in G^r}\sigma_j^2}\mathrm{Ent}(\sigma_L^2) + \left(1 - \frac{\sum_{i\in L}\sigma_i^2}{\sum_{j\in G^r}\sigma_j^2}\right)\mathrm{Ent}(\sigma^2_{G^r\setminus L}) && (97)\\
&< \ln(2) + \mathrm{Ent}(\sigma_L^2) + \left(1 - \frac{\sum_{i\in L}\sigma_i^2}{\sum_{j\in G^r}\sigma_j^2}\right) 8\ln(m) && (98)\\
&\le \ln(2) + 33\mathrm{Ent}(\sigma_L^2), && (99)
\end{aligned}$$

where the first inequality is due to the principal of maximum entropy and Lemma 4.2, and the last inequality is by Lemma E.2. □

Lemma D.5 (Retate Lemma 6.5). *Let $\eta_F = \binom{2m}{m-1}^{-1}$ for any $F \subset L$ with $|F| = m-1$ and $\eta_F = 0$ otherwise. There exists some constant $c' > 0.005$. The objective of optimization (26) is at least $c'\sum_{i\in G^l}\sigma_i^2\mathrm{Ent}(\sigma^2_{G^r}) - \ln(2)\sum_{i\in L}\sigma_i^2$.*

Proof. Recall $L \subset G^r$ with $|L| = 2m$ is the subset of arms with largest variances within G^r. For any $M \subset L$ with $|M| = m$, by the grouping property of entropy function we have

$$\begin{aligned}
\mathrm{Ent}(\sigma_L^2) &= \mathrm{Ent}(\sum_{i\in M}\sigma_i^2, \sum_{j\in L\setminus M}\sigma_j^2) + \frac{\sum_{i\in M}\sigma_i^2}{\sum_{j\in L}\sigma_j^2}\mathrm{Ent}(\sigma_M^2) + \frac{\sum_{i\in L\setminus M}\sigma_i^2}{\sum_{j\in L}\sigma_j^2}\mathrm{Ent}(\sigma^2_{L\setminus M}) && (100)\\
&\le \ln(2) + \frac{\sum_{i\in M}\sigma_i^2}{\sum_{j\in L}\sigma_j^2}\mathrm{Ent}(\sigma_M^2) + \frac{\sum_{i\in L\setminus M}\sigma_i^2}{\sum_{j\in L}\sigma_j^2}\mathrm{Ent}(\sigma^2_{L\setminus M}), && (101)
\end{aligned}$$

where the inequality is by the principal of maximum entropy. Multiply $\sum_{j\in L}\sigma_j^2$ on both side, and we have

$$\sum_{i\in M}\sigma_j^2\mathrm{Ent}(\sigma_M^2) + \sum_{i\in L\setminus M}\sigma_i^2\mathrm{Ent}(\sigma^2_{L\setminus M}) \ge \sum_{j\in L}\sigma_j^2(\mathrm{Ent}(\sigma_l^2) - \ln(2)). \tag{102}$$

Since $|M| = |L\setminus M| = m$, summing the inequality above for each $M \subset L$ with $|M| = m$ and multiplying by $\frac{1}{2\binom{2m}{m-1}}$ gives us

$$\begin{aligned}
\sum_{M\subset L:|M|=m}\frac{1}{\binom{2m}{m-1}}\sum_{i\in M}\sigma_i^2\mathrm{Ent}(\sigma_M^2) &\ge \frac{\binom{2m}{m}}{2\binom{2m}{m-1}}(\mathrm{Ent}(\sigma_L^2) - \ln(2))\sum_{i\in L}\sigma_i^2 && (103)\\
&\ge \frac{1}{2}(\mathrm{Ent}(\sigma_L^2) - \ln(2))\sum_{i\in L}\sigma_i^2 = \frac{1}{2}\mathrm{Ent}(\sigma_L^2)\sum_{i\in L}\sigma_i^2 - \frac{\ln(2)}{2}\sum_{i\in L}\sigma_i^2 && (104)\\
&\ge \frac{1}{2}\sum_{i\in L}\sigma_i^2\frac{\mathrm{Ent}(\sigma^2_{G^r}) - \ln(2)}{33} - \frac{\ln(2)}{2}\sum_{i\in L}\sigma_i^2 && (105)\\
&\ge \frac{1}{6}\sum_{i\in G^r}\sigma_i^2\frac{\mathrm{Ent}(\sigma^2_{G^r})}{33} - \frac{1}{2}\sum_{i\in L}\sigma_i^2\frac{\ln(2)}{33} - \frac{\ln(2)}{2}\sum_{i\in L}\sigma_i^2 && (106)\\
&\ge \frac{1}{174}\sum_{i\in G^r}\sigma_i^2\mathrm{Ent}(\sigma^2_{G^r}) - \ln(2)\sum_{i\in L}\sigma_i^2, && (107)
\end{aligned}$$

where the second inequality is by $\frac{\binom{2m}{m}}{\binom{2m}{m-1}} \ge 1$, the third and forth inequalities are by Lemma D.4. □

E Supporting Lemmas

Lemma E.1 (Lemma 5.1 in (Lattimore and Szepesvári, 2020)). *Given two bandit instances $I = (\mu_{1:n}, \sigma^2_{1:n})$ and $I' = (\mu'_{1:n}, \sigma'^2_{1:n})$, and let P_I and $P_{I'}$ be the probability measure associated with the bandit instances, respectively.*

Then for any algorithm A with the number of pulling for each arm-i as T_i^{A}, which is a random variable, let τ^{A} be the bandit process and let $\mathbb{P}_{I,\pi}$ and $\mathbb{P}_{I',\pi}$ be the probability measures induced by τ^{A} on instance I and I', respectively. We have

$$D(\mathbb{P}_{I,\mathsf{A}}||\mathbb{P}_{I',\mathsf{A}}) = \sum_{i=1}^{n} \mathbb{E}_I[T_i^{\mathsf{A}}] D\left(\mathcal{N}(\mu_i, \sigma_i^2)||\mathcal{N}(\mu_i', \sigma_i'^2)\right). \tag{108}$$

Lemma E.2. *For any $\sigma_{1:n}^2$, we have $\frac{\sum_{i\in L}\sigma_i^2}{\sum_{j\in G^r}\sigma_j^2} \geq \frac{1}{3}$. In addition,*

$$\left(1 - \frac{\sum_{i\in L}\sigma_i^2}{\sum_{j\in G^r}\sigma_j^2}\right)\ln(m) \leq 4\mathrm{Ent}(\sigma_L^2), \tag{109}$$

for some constant $c > 0$.

Proof of Lemma E.2. Suppose the minimum variance in σ_L^2 is $\tilde{\sigma}^2$. Let $\alpha = \frac{2m\tilde{\sigma}^2}{\sum_{i\in L}\sigma_i^2} \in (0, 1]$, which implies $\sum_{i\in L}\sigma_i^2 = 2m\tilde{\sigma}^2/\alpha$. In addition, $\sum_{j\in G^r\setminus L}\sigma_j^2 \leq 2m\tilde{\sigma}^2\sum_{i=0}^{\infty}2^{-i} = 4m\tilde{\sigma}^2$. It is straightforward to verify that

$$\frac{\sum_{i\in L}\sigma_i^2}{\sum_{j\in G^r}\sigma_j^2} = \frac{2m\tilde{\sigma}^2/\alpha}{\sum_{j\in G^r\setminus L}\sigma_j^2 + 2m\tilde{\sigma}^2/\alpha} \geq \frac{2m\tilde{\sigma}^2/\alpha}{4m\tilde{\sigma}^2 + 2m\tilde{\sigma}^2/\alpha} = \frac{1/\alpha}{2 + 1/\alpha} \geq \frac{1}{3}, \tag{110}$$

which proves the first statement. It follows that

$$1 - \frac{\sum_{i\in L}\sigma_i^2}{\sum_{j\in G^r}\sigma_j^2} \leq 1 - \frac{1/\alpha}{2 + 1/\alpha} = \frac{2}{2 + 1/\alpha} < \frac{2}{1 + 1/\alpha}. \tag{111}$$

By concavity of entropy function, $\mathrm{Ent}(\sigma_L^2) \geq \mathrm{Ent}\left(1 - \frac{2m-1}{2m}\alpha, \frac{\alpha}{2m}, \frac{\alpha}{2m}, \cdots, \frac{\alpha}{2m}\right)$. It implies that

$$\begin{aligned}
&(1 + 1/\alpha)\mathrm{Ent}(\sigma_L^2) && (112)\\
&\geq (1 + 1/\alpha)\left(-(1 - \frac{2m-1}{m}\alpha)\ln\left(1 - \frac{2m-1}{2m}\alpha\right) + \frac{2m-1}{2m}\alpha\ln\frac{2m}{\alpha}\right) && (113)\\
&\geq \frac{2m-1}{2m}\ln(2m) + \frac{2m-1}{2m}\ln\frac{1}{\alpha} - \left(\frac{1}{\alpha} - \frac{2m-1}{2m}\right)\ln\left(1 - \frac{2m-1}{2m}\alpha\right) && (114)\\
&\geq \frac{1}{2}\ln(2m) - \frac{1}{2}\ln(\alpha) - (1/\alpha - 1)\ln(1 - \alpha) && (115)\\
&\geq \frac{1}{2}\ln(2m) > \frac{1}{2}\ln(m). && (116)
\end{aligned}$$

We thus have

$$4\mathrm{Ent}(\sigma_L^2) > \frac{2}{1 + 1/\alpha}\ln(m) > \left(1 - \frac{\sum_{i\in L}\sigma_i^2}{\sum_{j\in G^r}\sigma_j^2}\right)\ln(m). \tag{117}$$

□