Almost Optimal Variance-Constrained Best Arm Identification

Yunlong Hou, *Graduate Student Member, IEEE*, Vincent Y. F. Tan, *Senior Member, IEEE*, and Zixin Zhong

***Abstract*— We design and analyze *Variance-Aware-Lower and Upper Confidence Bound* (VA-LUCB), a parameter-free algorithm, for identifying the best arm under the fixed-confidence setup and under a stringent constraint that the variance of the chosen arm is strictly smaller than a given threshold. An upper bound on VA-LUCB's sample complexity is shown to be characterized by a fundamental variance-aware hardness quantity H_{VA}. By proving an information-theoretic lower bound, we show that sample complexity of VA-LUCB is optimal up to a factor logarithmic in H_{VA}. Extensive experiments corroborate the dependence of the sample complexity on the various terms in H_{VA}. By comparing VA-LUCB's empirical performance to a close competitor RiskAverse-UCB-BAI by David et al. (2018) our experiments suggest that VA-LUCB has the lowest sample complexity for this class of risk-constrained best arm identification problems, especially for the riskiest instances.**

***Index Terms*— Stochastic multi-armed bandits, best arm identification, risk-aware bandits.**

I. Introduction

THE stochastic multi-armed bandit (MAB) problem [1] is a classical framework for online decision-making problems with extensive applications, e.g., clinical trials and financial portfolio. In a conventional stochastic MAB problem, given several arms with each of them associated with a fixed but unknown reward distribution, an agent selects an arm and observes a random reward returned from the corresponding distribution at each round. There are two complementary tasks in MAB problems. Firstly, the *regret minimization problem* aims to maximize the expected cumulative reward. The second task, the main focus of the present paper, is the *best arm identification* or BAI problem that aims to devise a strategy to identify the arm with the largest expected reward.

While the expected reward is a key indication of the quality of an arm, its *risk* should also be taken into consideration, e.g., in clinical trials where the effects of experimental drugs exhibit variability over different individuals and in financial portfolio where conservative investors seek a beneficial and also safe product. Instead of pursuing the highest payoff, one may wish to mitigate the underlying risks of certain arms by balancing between the reward and the potential risk. Various measures of risk [2], [3], [4] have been adopted, such as the variance, Value-at-Risk (VaR or α-quantile), and the Conditional Value-at-Risk (CVaR). We adopt the variance as the risk measure, but our techniques are *also* applicable to other risk measures if suitable concentration bounds are available. We design and analyze the VA-LUCB algorithm which is shown to be almost optimal in terms of the sample complexity, and we identify a key fundamental hardness quantity H_{VA}. VA-LUCB also significantly outperforms a suitably modified algorithm of [5].

A. Literature Review

There are three main families of algorithms for the standard fixed-confidence BAI problem—confidence bound-based (CBB) algorithms [6], [7], [8], [9], tracking-based (TB) algorithms [10], and Bayesian-style (BS) algorithms [11]. Jamieson and Nowak [12] provide a comprehensive survey for CBB algorithms, which includes the Action Elimination algorithm [6], the Upper Confidence Bound (UCB) algorithm [7] and the LUCB algorithm [9]. While LUCB [9] is originally designed for top-k arm identification, Jamieson and Nowak [12] claimed that LUCB-based methods perform well both theoretically and empirically for BAI task (thus we build our algorithm upon LUCB). LUCB samples the arm with the largest sample mean i_t and another arm with the largest upper confidence bound j_t within the remaining arms. It terminates when the lower confidence bound of i_t is greater than the upper confidence bound of j_t. The family of LIL techniques [8], [13] which provide uniform (in time) bounds on the deviation of an empirical statistic from the true quantity can boost the performance of these CBB methods. TB algorithms such as Track and Stop [10] track the proportion of arm pulls and achieves asymptotic optimality. BS algorithms such as Top-Two Thompson sampling [11] are easy to implement, asymptotically optimal, and yield good theoretical and empirical results.

For the risk-aware BAI problem, there is a large body of literature that measures the quality of an arm by *general functions of its distribution instead of the expectation*. The mean-variance paradigm is studied by [14], [15], and [16] under the regret minimization framework. Sani et al. [14] regarded the variance as the measure of risk and proposed the MV-LCB

Manuscript received 20 May 2022; revised 11 September 2022; accepted 7 November 2022. Date of publication 14 November 2022; date of current version 17 March 2023. This work was supported in part by the Singapore National Research Foundation (NRF) Fellowship under Grant A-0005077-01-00 and in part by the two Singapore Ministry of Education (MOE) Academic Research Fund (AcRF) Tier 1 under Grant A-0009042-01-00 and Grant A-8000189-01-00. *(Corresponding author: Zixin Zhong.)*

Yunlong Hou is with the Department of Mathematics, National University of Singapore, Singapore 119076 (e-mail: yhou@u.nus.edu).

Vincent Y. F. Tan is with the Department of Mathematics and the Department of Electrical and Computer Engineering, National University of Singapore, Singapore 119076 (e-mail: vtan@nus.edu.sg).

Zixin Zhong is with the Department of Computing Science, University of Alberta, Edmonton, AB T6G 2R3, Canada (e-mail: zixin.zhong@u.nus.edu).

Communicated by C. Tian, Associate Editor for Signal Processing and Source Coding.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIT.2022.3222231.

Digital Object Identifier 10.1109/TIT.2022.3222231

algorithm. The regret analysis of MV-LCB [14] was improved by [15]. Zhu and Tan [16] and Chang et al. [17] proposed Thompson sampling-based algorithms that are optimal under different regimes for the mean-variance and CVaR criteria respectively. The mean-variance paradigm was generalized by [18] where the quality of an arm is measured by some functions of the mean and the variance. Another class of risk measures that is widely studied consists of the VaR and CVaR. Under the BAI framework, Prashanth et al. [19] adapted the successive rejects algorithm of [7] for optimizing the CVaR. Kagrecha et al. [20] utilized a linear combination of the reward and the CVaR as the measure of quality of the arms and relaxed the prior knowledge of the reward distribution; this was generalized recently to general risk measures [21]. David and Shimkin [22] aimed at finding the arm with the maximum α-quantile. Under the regret minimization framework, Kagrecha et al. [23] and Baudry et al. [24] regarded the CVaR as a risk measure and proposed the RC-LCB algorithm and Thompson sampling-based algorithms respectively. Other risk measures have also been considered. For example, the *Sharpe ratio*, together with the mean-variance, was adopted by [25] to balance the tradeoff between return and risk. Maillard [26] proposed RA-UCB which considers the measure of *entropic risk* with a parameter λ. Cassel et al. [2] presented a general and systematic approach to analyzing risk-aware MABs. They adopted the *Empirical Distribution Performance Measure* and proposed the U-UCB algorithm to perform "proxy regret minimization".

Another approach casts the risk-aware MAB problem as a *constrained MAB problem*, i.e., the allowable risk that the agent can tolerate is formulated as a constraint in the online optimization problem. This is of practical interest in high-risk settings (such as clinical trials) in which the agent demands that the arm (treatment) to be eventually selected has a risk that is strictly below a permissible threshold. David et al. [5] focused on identifying an arm with *almost* the largest mean among those *almost* satisfying an α-quantile constraint under the fixed confidence setting. The authors presented a UCB-based algorithm named RiskAverse-UCB-m-best. Chang [27] considered an average cost constraint where each arm is associated with a cost variable that is independent of the reward and analyzed the probability of pulling optimal arms. This approach is also related to safe bandits [28], [29], where the arms are conservatively pulled to meet the safety constraint. However, safe bandits are often considered in a cumulative regret setting and the pulled arms should be safe with high probability (w.h.p.). A brief and current survey of taking risk into account in the study of multi-armed bandits is presented in [30].

The variance-constrained BAI problem consists of two distinct tasks—we seek *optimality* in the mean and *feasibility* in the variance. This is different from the Pareto-front identification with bandit feedback problem [31], [32], [33], which seeks optimality in *both* objectives, i.e., it seeks a solution/arm that has high mean and low variance simultaneously. While the best feasible arm, if it exists, belongs to the Pareto-Front, we still need to identify the best feasible arm among all arms on the Pareto-Front. These two problems are relevant but are essentially different. The problem is also related to identifying the best arm among the feasible arms. In [34], the arms follow multi-dimensional distributions and the feasible arms are defined to be arms whose mean vectors lie in a polyhedron. It only involves a single mean vector and its projection onto either a subspace (for the objective) and a polyhedron (for the feasibility constraint), while here we have to consider two different statistics—the mean and the variance.

There are works associated with the variance estimation [35], [36] in the BAI problem. However, the variance estimation is done to *improve* the algorithms for the *standard* BAI objective in both works. Our feasibility constraint in terms of the variance, in conjunction with the standard BAI objective, is a novel problem setting.

B. Contributions

We consider the *variance-constrained* BAI problem under the fixed confidence setting, i.e., we wish to identify the arm which satisfies a certain variance constraint and has the largest expectation w.h.p. Different from [5], we aim to identify the best arm strictly satisfying the risk constraint *without any slack* or *suboptimality*. We discuss more differences of our setting and our algorithm vis-à-vis [5] in Section IV-C.

We design VA-LUCB$(\bar{\sigma}^2, \delta)$ and derive an upper bound on its time or sample complexity. VA-LUCB is an LUCB-based [9] algorithm that is generally better than UCB-based algorithms for BAI problems [12]. It particularizes to LUCB when the constraint is inactive. A hardness parameter H_{VA} is identified as a fundamental limit; H_{VA} also reduces to H_1 [7] when the constraint is inactive. Furthermore, the framework and analysis of VA-LUCB can be extended to other risk measures as long as there are appropriate concentration bounds, e.g., Bhat and Prashanth [37] or Chang and Tan [4] enables us to use CVaR or certain continuous functions as risk measures within the generic VA-LUCB framework. Different from the work of [5] which addresses a similar problem, our algorithm is *completely parameter free*, in the sense that Algorithm 1 can output the best feasible arm $i^\star$ without knowledge of any parameters that define the instance (e.g., the suboptimality gaps).

To assess the optimality of VA-LUCB, we prove an accompanying information-theoretic lower bound on the optimal expected sample complexity of *any* variance-constrained BAI algorithm. We show that VA-LUCB's sample complexity is *optimal* up to a logarithmic factor in H_{VA}.

Lastly, we present extensive experiments in which we examine the effect of each term in H_{VA}. We compare VA-LUCB to a naïve algorithm based on uniform sampling and a variant of the algorithm in David et al. [5] which can only be applied if some unknown parameters (such as the suboptimality gaps) are known (see App. B). Our experiments suggest that VA-LUCB is the gold standard for this class of constrained BAI problems, reducing the sample complexity significantly, especially for the riskiest instances.

II. Problem Setup

Given a positive integer n, let $[n] = \{1, 2, \ldots, n\}$. We assume that there are N arms and arm $i \in [N]$ corresponds

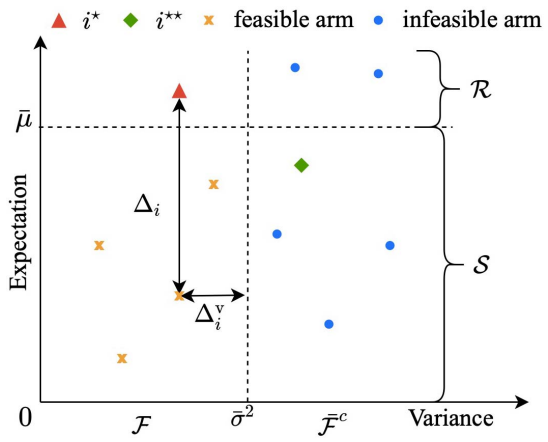


Fig. 1. A diagram of the arms. Each dot represents the expectation and variance of an arm.

to a reward distribution ν_i. For each i, the reward of arm i is denoted by X_i with $X_i \sim \nu_i$, which is independent of $X_j \sim \nu_j$ for all $j \in [N]\backslash\{i\}$. The expectation and variance of X_i are denoted by μ_i and σ_i^2 respectively. The permissible upper bound on the variance is denoted by $\bar{\sigma}^2 > 0$. An *instance* $(\nu = (\nu_1, \ldots, \nu_N), \bar{\sigma}^2)$, consists of N reward distributions and the upper bound on the variance $\bar{\sigma}^2$. Given any instance $(\nu, \bar{\sigma}^2)$, arm i is said to be *feasible* if $\sigma_i^2 \le \bar{\sigma}^2$. We define $\mathcal{F} := \{i \in [N] : \sigma_i^2 \le \bar{\sigma}^2\}$ to be the *feasible set* which contains all the feasible arms. Let $\bar{\mathcal{F}}^c := [N]\backslash\mathcal{F}$ be the set of all the *infeasible* arms. We say an instance is *feasible* if $\mathcal{F}$ is nonempty and we say it is *infeasible* otherwise. For a feasible instance, the feasibility flag $\mathsf{f} = 1$ and the best feasible arm $i^\star := \operatorname{argmax}\{\mu_i : i \in \mathcal{F}\}$, where argmax returns the smallest index that achieves the maximum. For an infeasible instance, the feasibility flag f is set to be 0.

An arm i is said to be *suboptimal* if $\mu_i < \mu_{i^\star}$ and *risky* otherwise. We define the *suboptimal set* $\mathcal{S} := \{i \in [N] : \mu_i < \mu_{i^\star}\}$ if $\mathcal{F} \neq \emptyset$ and $\mathcal{S} := \emptyset$ if $\mathcal{F} = \emptyset$. The *risky set* $\mathcal{R} := [N] \setminus \mathcal{S}$ consists of arms whose expectations are not smaller than $\mu_{i^\star}$. Define $i^{\star\star} := \operatorname{argmax}\{\mu_i : i \in \mathcal{S}\}$ to be the arm with greatest expectation among all the suboptimal arms if $\mathcal{S} \neq \emptyset$. Denote the *mean gap* for arms $i \in \mathcal{S}$ as $\Delta_i = \mu_{i^\star} - \mu_i$ if $\mathcal{F} \neq \emptyset$. Denote the mean gap for arm $i^\star$ as $\Delta_{i^{\star\star}}$ if $\mathcal{F} \neq \emptyset$ and $\mathcal{S} \neq \emptyset$ and $+\infty$ if $\mathcal{S} = \emptyset$. Let the *variance gaps* for all arms $i \in [N]$ be $\Delta_i^{\mathrm{v}} := |\sigma_i^2 - \bar{\sigma}^2|$. The *separator* between $i^\star$ and the suboptimal arms is denoted by $\bar{\mu} := (\mu_{i^\star} + \mu_{i^{\star\star}})/2$ if $\mathcal{F} \neq \emptyset$ and $\mathcal{S} \neq \emptyset$ and $\bar{\mu} := -\infty$ otherwise. These sets and quantities are illustrated in Figure 1.

At round r, the agent pulls an arm $i_r \in [N]$ based on the observation history $((i_1, X_{1,i_1}), \ldots, (i_{r-1}, X_{r-1,i_{r-1}}))$. The agent then observes $X_{r,i_r} \sim \nu_{i_r}$. The rewards sampled from the same arm at different rounds are i.i.d., i.e., $\{X_{r,i} : r \in \mathbb{N}\}$ are i.i.d. samples drawn from ν_i.

We assume that, if it exists, the best feasible arm is *unique* and the variance of the best feasible arm is strictly smaller than $\bar{\sigma}^2$, i.e., $\sigma_{i^\star}^2 < \bar{\sigma}^2$. We discuss the case $\sigma_{i^\star}^2 = \bar{\sigma}^2$ in App. A. For the sake of clarity, we consider bounded rewards, which are sub-Gaussian. Without loss of generality, the reward distributions are supported on $[0, 1]$. We describe extensions to sub-Gaussian rewards in App. E.

Given an instance $(\nu, \bar{\sigma}^2)$, we would like to design and analyze an algorithm that *succeeds* w.h.p., i.e., to identify whether the instance is feasible, and if so, identify the best feasible arm $i^\star$ in the fewest number of rounds. An algorithm $\pi := ((\pi_r)_{r\in\mathbb{N}}, (\Gamma_r^\pi)_{r\in\mathbb{N}}, \phi^\pi)$ determines which arm to pull, when to stop, whether the instance is feasible, and which arm to recommend. More precisely,

- The *sampling strategy* $\pi_r : ([N] \times [0, 1])^{r-1} \to [N]$ decides which arm to sample at round r based on the observation history, i.e.
$$\pi_r((i_1^\pi, X_{1,i_1^\pi}), \ldots, (i_{r-1}^\pi, X_{r-1,i_{r-1}^\pi})) = i_r^\pi.$$
Let $\mathcal{H}_r := \sigma(i_1^\pi, X_{1,i_1^\pi}, \ldots, i_t^\pi, X_{r,i_r^\pi})$ be the history of arm pulls and rewards. Then π_r is $\mathcal{H}_{r-1}$-measurable.
- The *stopping rule* Γ_r^π where Γ_r^π is $\mathcal{H}_r$-measurable and $\{\text{stop}, \text{continue}\}$-valued, decides whether to stop the algorithm at each round r. The stopping round is denoted by τ^π if the algorithm stops.
- The *recommendation rule* $\phi^\pi : ([N] \times [0, 1])^{\tau^\pi} \to \{0, 1\} \times ([N] \cup \{\emptyset\})$ finally gives an estimated flag $\hat{\mathsf{f}}^\pi \in \{0, 1\}$ and an arm $i_{\mathrm{out}}^\pi \in [N]$ if $\hat{\mathsf{f}}^\pi = 1$ based on the observation history (i.e. ϕ^π is $\mathcal{H}_{\tau^\pi}$-measurable):
$$\phi^\pi((i_1^\pi, X_{1,i_1^\pi}), \ldots, (i_{\tau^\pi}^\pi, X_{\tau^\pi,i_{\tau^\pi}^\pi})) = (\hat{\mathsf{f}}^\pi, i_{\mathrm{out}}^\pi),$$

The sample complexity of the algorithm π is denoted as τ^π. In the fixed confidence setting, we say that an algorithm π is δ-PAC if the following two conditions hold

$$\mathbb{P}_\nu[\hat{\mathsf{f}}^\pi = 1, i_{\mathrm{out}}^\pi = i^\star \mid \mathsf{f} = 1] \ge 1 - \delta \quad \text{and}$$
$$\mathbb{P}_\nu[\hat{\mathsf{f}}^\pi = 0 \mid \mathsf{f} = 0] \ge 1 - \delta.$$

The above conditions imply that π succeeds with probability at least $1-\delta$. Our aim is to design and analyze a δ-PAC algorithm π that minimizes the sample complexity τ^π in expectation and w.h.p. We define the optimal expected sample complexity as

$$\tau_\delta^\star = \tau_\delta^\star(\nu, \bar{\sigma}^2) := \inf\left\{\mathbb{E}[\tau^\pi] : \pi \text{ is } \delta\text{-PAC}\right\},$$

where the infimum is taken over all δ-PAC algorithms π (as defined above). For simplicity, we omit the superscripts π in τ^π, ϕ^π and $\hat{\mathsf{f}}^\pi$ if there is no risk of confusion.

III. The VA-LUCB Algorithm

We present our algorithm which is named *Variance-Aware-Lower and Upper Confidence Bound* (or *VA-LUCB*) in Algorithm 1. Given an instance $(\nu, \bar{\sigma}^2)$, the agent pulls each arm according to the VA-LUCB policy to ascertain whether the instance is feasible and to determine which arm is the best feasible arm if the instance is ascertained to be feasible.

Each *time step* (Lines 3 to 19) in our algorithm consists of one or two *rounds*, i.e., the agent may pull one or two arms at each time step. The algorithm warms up by pulling each of the arms twice (Line 2). At time step t, we first update the sample means, the sample variances and the confidence bounds of the arms that require exploration (Lines 4 and 5); these are the arms in the so-called possibly feasible set $\bar{\mathcal{F}}_{t-1}$, which will be defined formally in (7). Let $\mathcal{J}_t$ denote the set of arms sampled at time step t. Define $T_i(t) := \sum_{s=1}^{t-1} \mathbb{1}\{i \in \mathcal{J}_s\}$ to

Algorithm 1 Variance-Aware LUCB (VA-LUCB)

1: **Input**: threshold $\bar{\sigma}^2 > 0$ and confidence parameter $\delta \in (0,1)$.
2: Sample each of the N arms twice and set $\bar{\mathcal{F}}_N = [N]$.
3: **for** time step $t = N+1, N+2 \ldots$ **do**
4: Compute the sample mean using (1) and sample variance using (2) for $i \in \bar{\mathcal{F}}_{t-1}$.
5: Update the confidence bounds for the mean and variance by (4) and (5) for $i \in \bar{\mathcal{F}}_{t-1}$.
6: Update $\mathcal{F}_t$ and $\bar{\mathcal{F}}_t$ (see (6) and (7)).
7: Find $i_t^\star := \operatorname{argmax}\{\hat{\mu}_i(t) : i \in \mathcal{F}_t\}$ if $\mathcal{F}_t \neq \emptyset$.
8: Update $\mathcal{P}_t$ according to (8).
9: **if** $\bar{\mathcal{F}}_t \cap \mathcal{P}_t = \emptyset$ **then**
10: **if** $\mathcal{F}_t \neq \emptyset$ **then** Set $i_{\text{out}} = i_t$ using (9) and $\hat{\mathsf{f}} = 1$. **else** Set $\hat{\mathsf{f}} = 0$. **end if**
11: **break**
12: **end if**
13: **if** $|\bar{\mathcal{F}}_t| = 1$ **then**
14: Sample arm i_t using (9) (in one round).
15: **else**
16: Find i_t and competitor arm c_t according to (10).
17: **if** $U_{c_t}^{\mu}(t) \geq L_{i_t}^{\mu}(t)$ **then** Sample arms i_t and c_t (in two rounds).
18: **else** Sample arm i_t (in one round). **end if**
19: **end if**
20: **end for**

be the number of times arm i is pulled before time step t. For arm i that requires exploration, the sample mean and sample variance before time step t are

$$\hat{\mu}_i(t) := \frac{1}{T_i(t)} \sum_{s=1}^{t-1} X_{s,i} \mathbb{1}\{i \in \mathcal{J}_s\}, \quad \text{and} \tag{1}$$

$$\hat{\sigma}_i^2(t) := \frac{\sum_{s=1}^{t-1} \left(X_{s,i} - \hat{\mu}_i(t)\right)^2 \mathbb{1}\{i \in \mathcal{J}_s\}}{T_i(t) - 1}. \tag{2}$$

We define the *confidence radii* for the mean and variance as

$$\alpha(t,T) = \beta(t,T) := \sqrt{\frac{1}{2T} \ln\left(\frac{2Nt^4}{\delta}\right)}. \tag{3}$$

We denote the *lower* and *upper confidence bounds* (LCB and UCB) for the empirical mean of arm i as

$$\begin{aligned} L_i^{\mu}(t) &:= \hat{\mu}_i(t) - \alpha(t, T_i(t)) \quad \text{and} \\ U_i^{\mu}(t) &:= \hat{\mu}_i(t) + \alpha(t, T_i(t)) \end{aligned} \tag{4}$$

respectively, as well as the LCB and UCB for the empirical variance respectively as

$$\begin{aligned} L_i^{\mathrm{v}}(t) &:= \hat{\sigma}_i^2(t) - \beta(t, T_i(t)) \quad \text{and} \\ U_i^{\mathrm{v}}(t) &:= \hat{\sigma}_i^2(t) + \beta(t, T_i(t)). \end{aligned} \tag{5}$$

A. Partition of the Arms

Based on the empirical variances, at each time step t, we partition the arms into three disjoint subsets based on the confidence bounds on the variance (Line 6 of Algorithm 1). The first set is the *empirically feasible set* at time step t,

$$\mathcal{F}_t := \{i : U_i^{\mathrm{v}}(t) \leq \bar{\sigma}^2\}. \tag{6}$$

The second set is the *empirically almost feasible set*,

$$\partial\mathcal{F}_t := \{i : L_i^{\mathrm{v}}(t) \leq \bar{\sigma}^2 < U_i^{\mathrm{v}}(t)\}.$$

We define the union of the above two sets as the *possibly feasible set*,

$$\bar{\mathcal{F}}_t := \mathcal{F}_t \cup \partial\mathcal{F}_t. \tag{7}$$

The *empirically infeasible set* at time step t is

$$\bar{\mathcal{F}}_t^c := \{i : L_i^{\mathrm{v}}(t) > \bar{\sigma}^2\}.$$

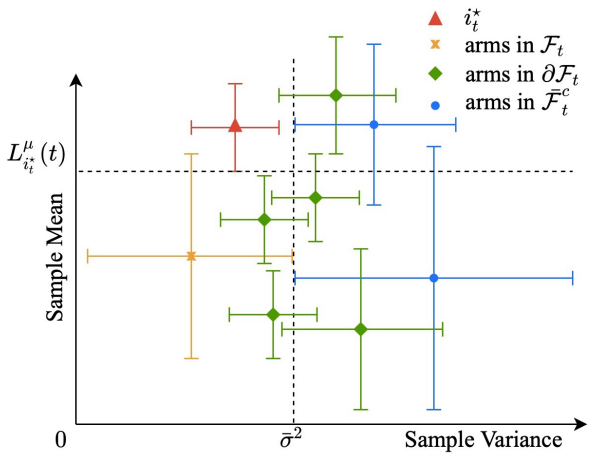


Fig. 2. Illustration of the empirical sets. Each dot represents the estimated mean and variance of each arm at time step t. The horizontal (resp. vertical) component and the horizontal (resp. vertical) crossbar indicates the sample variance (resp. mean) and the confidence interval for the true variance (resp. mean). i_t is the green arm at the top and $c_t = i_t^\star$. Arms whose UCBs of the empirical means $U_i^{\mu}(t)$ are greater than $L_{i_t^\star}^{\mu}(t)$ are in $\mathcal{P}_t$.

These sets are illustrated in Figure 2. The arms that require exploration are the arms in the possibly feasible set $\bar{\mathcal{F}}_t$. Our intuition is that w.h.p., the true variance of each arm is bounded by the corresponding LCB and UCB, i.e., $\sigma_i^2 \in [L_i^{\mathrm{v}}(t), U_i^{\mathrm{v}}(t)]$; this is stated precisely in Lemma 1. If one arm lies in $\mathcal{F}_t$, it is feasible ($\sigma_i^2 \leq \bar{\sigma}^2$) w.h.p. Thus only its sample mean needs to be further examined. If one arm lies in $\partial\mathcal{F}_t$, its true feasibility remains unclear, which indicates that this arm needs to be pulled more. If one arm lies in $\bar{\mathcal{F}}_t^c$, it is infeasible ($\sigma_i^2 > \bar{\sigma}^2$) w.h.p. Hence, it will not be pulled in future. In summary, only the arms in the possibly feasible set $\bar{\mathcal{F}}_t$ need to be explored more. This justifies the update rules in Lines 4 and 5 of Algorithm 1.

In terms of the sample mean, if $\mathcal{F}_t \neq \emptyset$, there is an *empirically best feasible arm* at time step t (Line 7)

$$i_t^\star := \operatorname{argmax}\{\hat{\mu}_i(t) : i \in \mathcal{F}_t\}.$$

Define the *potential set* at time step t (Line 8) as:

$$\mathcal{P}_t := \begin{cases} \{i : L_{i_t^\star}^{\mu}(t) \leq U_i^{\mu}(t), i \neq i_t^\star\}, & \mathcal{F}_t \neq \emptyset \\ [N], & \mathcal{F}_t = \emptyset \end{cases}. \tag{8}$$

The potential set contains those arms which potentially have greater expectations than $i_t^\star$, regardless of their feasibility.

Considering both the sample variance and sample mean, arms in $\bar{\mathcal{F}}_t \cap \mathcal{P}_t$ are said to be *competitor arms*, in the sense that they are possibly feasible and potentially have greater means than the empirically best feasible arm $i_t^\star$. In conclusion, only the competitor arms in the set $\bar{\mathcal{F}}_t \cap \mathcal{P}_t$ need to be pulled more, which also motivates our stopping rule.

B. Stopping Rule

The intuition for the stopping rule in Lines 9 to 12 of VA-LUCB is straightforward. If the given instance is infeasible, after pulling all arms sufficiently many times, we have $\mathcal{F}_t = \bar{\mathcal{F}}_t = \emptyset$ and $\mathcal{P}_t = [N]$ (i.e., all the arms are deemed to be infeasible) w.h.p. Thus we set the flag $\hat{\mathsf{f}} = 0$ for this instance. If the given instance is feasible, after sufficiently many arm pulls, the arms in $\mathcal{R}\backslash\mathcal{F}$, the best feasible arm $i^\star$ and the arms in $\mathcal{S}$ will be ascertained to be infeasible, feasible, and suboptimal respectively. At the stopping time step τ, $\mathcal{F}_\tau \neq \emptyset$, and we expect that $i_\tau = i_\tau^\star = i^\star$, where

$$i_t := \operatorname{argmax} \left\{ \hat{\mu}_i(t) : i \in \bar{\mathcal{F}}_t \right\}. \tag{9}$$

We formalize this intuition in Lemma 2 in Section VI. When $\bar{\mathcal{F}}_t \cap \mathcal{P}_t = \emptyset$, there are no competitor arms and we are confident in asserting that the instance is feasible, i.e., $\hat{\mathsf{f}} = 1$ and i_t is the best feasible arm.

C. Sampling Strategy

When the algorithm has not terminated, $\mathcal{P}_t$ and $\bar{\mathcal{F}}_t$ are not empty. The intuition for the sampling strategy in Lines 13 to 20 of VA-LUCB can be justified as follows. If the given instance is feasible, firstly, when i_t is a truly infeasible arm, its infeasibility needs to be ascertained, and secondly, when i_t is a truly feasible arm, we need to check both of its feasibility and optimality. Thus, in either case, arm i_t requires more pulls. When $|\bar{\mathcal{F}}_t| > 1$, define the *best competitor arm* to i_t as[1]

$$c_t := \operatorname{argmax}\{U_i^\mu(t) : i \in \bar{\mathcal{F}}_t, i \neq i_t\}. \tag{10}$$

The fact that $c_t \in \mathcal{P}_t \cup \{i_t^\star\}$ can be justified by Lemma 7 in App. D. Thus, more pulls of c_t are needed to ascertain which of i_t and c_t has a larger true mean. If the given instance is infeasible, all the arms in $\bar{\mathcal{F}}_t$, including i_t and c_t, need to be sampled more times to assert they are indeed infeasible.

We remark that in VA-LUCB, we are *interleaving* the verification of optimality (in the mean aspect) and feasibility (in the variance aspect). This is in stark contrast to a naïve but suboptimal strategy in which one uses a two-phase strategy to first identify the feasible arms, then search among these arms for the one with the largest mean.

IV. Bounds on the Time Complexity

We state an upper bound on the sample complexity of our VA-LUCB algorithm and a lower bound on the optimal expected sample complexity over all algorithms.

[1]Note the *competitor arms* are defined for $i_t^\star$ and the *best competitor arm* is defined for i_t. However, when the given instance is feasible and the arms in $\bar{\mathcal{F}}^c \cap \mathcal{R}$ are identified as infeasible w.h.p., i_t will likely be $i_t^\star$ and c_t will likely be an suboptimal arm in $\mathcal{S}$. Thus $c_t \in \mathcal{P}_t$ and is a competitor arm to $i_t^\star$ w.h.p.

A. Time Complexity of VA-LUCB

Given an instance $(\nu, \bar{\sigma}^2)$, define the *variance-aware hardness parameter*

$$H_{\mathrm{VA}} := \frac{1}{\min\{\frac{\Delta_{i^\star}}{2}, \Delta_{i^\star}^{\mathrm{v}}\}^2} + \sum_{i \in \mathcal{F} \cap \mathcal{S}} \frac{1}{(\frac{\Delta_i}{2})^2} + \sum_{i \in \bar{\mathcal{F}}^c \cap \mathcal{R}} \frac{1}{(\Delta_i^{\mathrm{v}})^2} + \sum_{i \in \bar{\mathcal{F}}^c \cap \mathcal{S}} \frac{1}{\max\{\frac{\Delta_i}{2}, \Delta_i^{\mathrm{v}}\}^2}. \tag{11}$$

Our main result is stated as follows.

Theorem 1 (Upper Bound): Given an instance $(\nu, \bar{\sigma}^2)$ and confidence parameter δ, with probability at least $1 - \delta$, VA-LUCB succeeds and terminates in

$$O\left(H_{\mathrm{VA}} \ln \frac{H_{\mathrm{VA}}}{\delta} \right) \quad \text{time steps.}$$

The implied constant in the O-notation can be taken to be no more than 304. The mean gap Δ_i and variance gap Δ_i^{v} of arm i are indicative of the hardness of ascertaining its optimality and feasibility respectively. It is easy to see that when the threshold $\bar{\sigma}^2 \to \infty$ (in fact, $\bar{\sigma}^2 \geq 3/4$ suffices), H_{VA} reduces to the hardness parameter $H_1 := \sum_{i \neq i^\star} \Delta_i^{-2}$ in the conventional (unconstrained) BAI problem [7].

The intuitions for the four terms in H_{VA} are as follows: Firstly, to identify the best feasible arm $i^\star$, both of its feasibility and optimality need to be ascertained, which leads to the first term. Secondly, for the arms in $\mathcal{F} \cap \mathcal{S}$, we can identify them once we have established that they are indeed suboptimal, explaining the dependence on $\Delta_i^{-2}, i \in \mathcal{F} \cap \mathcal{S}$. Thirdly, since the arms in $\bar{\mathcal{F}}^c \cap \mathcal{R}$ have larger means than the best feasible arm, the algorithm needs to sample them sufficiently many times to learn they are infeasible, which contributes to the third term in H_{VA}. Finally, when either the suboptimality or the infeasibility of the arms in $\bar{\mathcal{F}}^c \cap \mathcal{S}$ is ascertained, we can eliminate them, which explains the last term in H_{VA}. The proof of Theorem 1 is presented in App. D.

Remark 1: We highlight that Algorithm 1 constitutes a convenient framework to tackle any risk-aware BAI problem in the sense that it is compatible with other concentration bounds. For example, one can define alternative confidence radii, different from those specified in (3), based on the (non-asymptotic) Law of the Iterated Logarithms (LIL) [8], [13], [38], [39]. We adopt a simple non-asymptotic LIL concentration bound from Jamieson et al. [8] to show that different confidence bounds utilized in VA-LUCB (Algorithm 1) can lead to slightly different upper bounds on the stopping time with high-probability. First, we replace the unbiased sample variance $\hat{\sigma}^2(t)$ by a biased counterpart

$$\tilde{\sigma}_i^2(t) := \frac{\sum_{s=1}^{t-1} \left(X_{s,i} - \hat{\mu}_i(t)\right)^2 \mathbb{1}\{i \in \mathcal{J}_s\}}{T_i(t)} = \frac{1}{T_i(t)} \sum_{s=1}^{t-1} X_{s,i}^2 \mathbb{1}\{i \in \mathcal{J}_s\} - \left(\frac{1}{T_i(t)} \sum_{s=1}^{t-1} X_{s,i} \mathbb{1}\{i \in \mathcal{J}_s\} \right)^2.$$

Next, we redefine the confidence radii $\alpha(t,T)$ and $\beta(t,T)$ (originally defined in (3)) by

$$\tilde{\alpha}(t) := \left(1+\sqrt{\epsilon}\right)\sqrt{\frac{1+\epsilon}{2t}\ln\left(\frac{4N\ln\left((1+\epsilon)t\right)}{\delta}\right)}, \quad \text{and}$$
$$\tilde{\beta}(t) := 3\tilde{\alpha}(t), \tag{12}$$

respectively, where $\epsilon \in (0,1)$ is a fixed constant. These choices of the confidence radii allow us to avoid using a union bound to bound the probability of the complement of the "good" event E in (18). With the above modifications, we show in App. F that VA-LUCB is δ-PAC (for $\epsilon = 0.9$ and $\delta < 0.1$) and succeeds in

$$O\left(H_{\mathrm{VA}}^{(1)}\ln\frac{N}{\delta} + H_{\mathrm{VA}}^{(3)}\right) \quad \text{time steps}, \tag{13}$$

where

$$H_{\mathrm{VA}}^{(1)} := \frac{1}{\min\{\Delta_{i^\star}, \frac{2}{3}\Delta_{i^\star}^{\mathrm{v}}\}^2} + \sum_{i\in\mathcal{F}\cap\mathcal{S}}\frac{1}{\Delta_i^2} + \sum_{i\in\bar{\mathcal{F}}^c\cap\mathcal{R}}\frac{1}{(\frac{2}{3}\Delta_i^{\mathrm{v}})^2} + \sum_{i\in\bar{\mathcal{F}}^c\cap\mathcal{S}}\frac{1}{\max\{\Delta_i, \frac{2}{3}\Delta_i^{\mathrm{v}}\}^2}, \tag{14}$$

$$H_{\mathrm{VA}}^{(3)} := \psi\Big(\frac{1}{\min\{\Delta_{i^\star}, \frac{2}{3}\Delta_{i^\star}^{\mathrm{v}}\}^2}\Big) + \sum_{i\in\mathcal{F}\cap\mathcal{S}}\psi\Big(\frac{1}{\Delta_i^2}\Big) + \sum_{i\in\bar{\mathcal{F}}^c\cap\mathcal{R}}\psi\Big(\frac{1}{(\frac{2}{3}\Delta_i^{\mathrm{v}})^2}\Big) + \sum_{i\in\bar{\mathcal{F}}^c\cap\mathcal{S}}\psi\Big(\frac{1}{\max\{\Delta_i, \frac{2}{3}\Delta_i^{\mathrm{v}}\}^2}\Big) \tag{15}$$

and $\psi : x \in \mathbb{R}_+ \mapsto x\ln\ln_+(x)$ with $\ln\ln_+ : x \in \mathbb{R}_+ \mapsto \ln\ln(\max\{e,x\})$. Note that $H_{\mathrm{VA}}^{(1)}$ and H_{VA} are order-wise equal and $H_{\mathrm{VA}}^{(3)}$ is also of the same order as H_{VA} up to double logarithmic terms in the gaps $\{(\Delta_i, \Delta_i^{\mathrm{v}})\}_{i\in[N]}$.

Remark 2: While we only study a generalization of the LUCB-based method [9] for this variance-constrained BAI problem, other confidence bound-based strategies, e.g., Successive Elimination [6] and lil'UCB [8], also have the potential to be generalized to solve this problem. We provide some intuitions in the following.

- **Successive Elimination**: Denote the set of active arms as $\mathcal{A}_t$ and initialize $\mathcal{A}_0 = [N]$. At each time step t, the algorithm first pulls all active arms once and updates the sample means and sample variances. It then updates the confidence bounds $\alpha(t)$ and $\beta(t)$ for the means and variances (which are similar to the α_t in [6, Alg. 3]). Next, it identifies the empirically best feasible arm $i_t^\star = \arg\max\{\hat{\mu}_i(t) : i \in \mathcal{A}_t, U_i^{\mathrm{v}}(t) \le \bar{\sigma}^2\}$. Finally, it updates the active arm set to be $\mathcal{A}_{t+1} = \{i \in \mathcal{A}_t : \hat{\mu}_{i_t^\star}(t) - \hat{\mu}_i(t) < 2\alpha(t), \hat{\sigma}_i^2(t) - \bar{\sigma}^2 < \beta(t)\}$. The algorithm terminates when the active set is empty (which indicates that the instance is infeasible) or the active set contains only the empirically best feasible arm (which is then declared to be the best feasible arm).
- **lil'UCB**: The sampling strategy for this algorithm when there is a constraint on variance (or risk) of the arms is obvious. In particular, the algorithm samples arm $i_t = \arg\max\{U_i^{\mu}(t) : L_i^{\mathrm{v}}(t) \le \bar{\sigma}^2\}$ where $U_i^{\mu}(t)$ and $L_i^{\mathrm{v}}(t)$ are constructed in view of the LIL. However, the stopping criterion is not straightforward, since the LIL-based stopping rule [8] cannot be directly utilized. This is an interesting direction for future research.

B. Lower Bound

A natural question is whether the upper bound stated in Theorem 1 (or the number of time steps of the LIL version of VA-LUCB in (13)) is tight and whether the quantity H_{VA} is *fundamental*. This is addressed in this section via an information-theoretic lower bound which indicates the expected sample complexity of VA-LUCB is optimal up to $\ln H_{\mathrm{VA}}$.

Since the rewards are bounded in $[0,1]$, the variance of each arm is at most $1/4$. Therefore, when $\bar{\sigma}^2 \in [1/4, \infty)$, all arms are feasible and there exists a generic lower bound [40]. When $\bar{\sigma}^2 \in (0, 1/4)$, let

$$\bar{a} := \frac{1+\sqrt{1-\bar{\sigma}^2}}{2} \quad \text{and} \quad \underline{a} := \frac{1-\sqrt{1-\bar{\sigma}^2}}{2}.$$

These quantities are the solutions to the quadratic equation $a(1-a) = \bar{\sigma}^2$.

Theorem 2 (Lower bound): Given any instance $(\nu, \bar{\sigma}^2)$ with $\bar{\sigma}^2 \in (0, 1/4)$, define the constant $c(\nu,\bar{\sigma}^2) := \min\left\{\underline{a}(1/4 - \bar{\sigma}^2),\ \underline{a}/8,\ (1-\mu_{i^\star})/8\right\}$,

$$\tau_\delta^\star \ge c(\nu,\bar{\sigma}^2)\, H_{\mathrm{VA}} \ln\left(\frac{1}{2.4\delta}\right). \tag{16}$$

The proof is in App. G. Based on Theorems 1 and 2, we have the following corollary whose proof is also provided in App. G. This almost conclusive result says that we have characterized $\tau_\delta^\star$ up to a (small) factor logarithmic in H_{VA}.

Corollary 1 (Almost Optimality of VA-LUCB): Given any instance $(\nu, \bar{\sigma}^2)$ and confidence parameter $\bar{\sigma}^2 \in (0, 1/4)$, the optimal expected sample complexity is

$$\tau_\delta^\star = O\Big(H_{\mathrm{VA}}\ln\frac{H_{\mathrm{VA}}}{\delta}\Big) \bigcap \Omega\Big(H_{\mathrm{VA}}\ln\frac{1}{\delta}\Big).$$

The bounds can also be expressed as

$$\lim_{\delta\downarrow 0}\frac{\tau_\delta^\star}{\log\frac{1}{\delta}} = \Theta(H_{\mathrm{VA}}),$$

and VA-LUCB achieves the upper bounds.

Corollary 1 says that H_{VA} is the fundamental limit for the problem of variance-constrained BAI.

C. Comparison to David et al. [5]

We adopt the variance as the risk measure and focus on the *(strict) best feasible* arm identification problem under the δ-PAC framework, while David et al. [5] uses the α-quantile as the risk metric and consider ϵ_ρ*-approximately feasible and* ϵ_μ*-approximately optimal* arms. We consider a variant of their algorithm, named *RiskAverse-UCB-BAI* (See App. B) that is tailored to our variance-constrained problem in which the best feasible arm must be produced w.h.p.

- **Parameters**: The most important difference is that VA-LUCB is *parameter free*. In contrast, RiskAverse-UCB-BAI heavily relies on knowledge of the hardness parameter H (which appears in the confidence radii),

and the accuracy parameters ϵ_μ and $\epsilon_{\rm v}$ (of the mean and variance respectively), which determine when it terminates. To output the best feasible arm w.h.p., one needs to set the accuracy parameters to be some functions of the unknown mean gaps and variance gaps such that the only $\epsilon_{\rm v}$-*approximately feasible and* ϵ_μ-*approximately optimal* arm is exactly the (strict) best feasible arm. Thus, if we want to output the best feasible arm, RiskAverse-UCB-BAI is *not* parameter free.

- **Upper Bounds**: The hardness parameters $\sum_{i\in[N]} C_i$ and H, defined in (S.5) and (S.6) respectively, are used to characterize the upper bound (on the sample complexity of RiskAverse-UCB-BAI) in [5, Theorem 3] and are lower bounded by $H_{\rm VA}$ (see App. C-A). Intuitively, since H is only a function of the accuracy parameters $(\epsilon_{\rm v}, \epsilon_\mu)$, but $H_{\rm VA}$ takes the means and variances of all arms into account, the latter is smaller (hence better). We formalize this intuition in App. C-A. Even disregarding these constants, the additional $\ln$ term in N and $\ln\ln$ term in N/δ in the upper bound of RiskAverse-UCB-BAI (see Eqn. (S.7)) indicates that its sample complexity is strictly larger than that of VA-LUCB (see App. C-A for details).
- **Lower Bounds**: By comparing terms involving arm i in both lower bounds, we deduce that our lower bound is strictly larger than that in [5, Theorem 2] for most ($\geq$ 99.9% of) $(\mu_{i^\star}, \bar{\sigma}^2)$ pairs (see App. C-B). Corollary 1 states that $H_{\rm VA}$ is fundamental in characterizing the hardness of the instance. This also implies the lower bound of [5] is, in general, not tight in our variance-constrained BAI setting. Due to the choice of confidence radius in (3), we also claim that VA-LUCB identifies risky arms faster than RiskAverse-UCB-BAI (see App. C-C).

V. Experiments

We design experiments to illustrate the empirical performance of VA-LUCB. We compare VA-LUCB to RiskAverse-UCB-BAI [5] and a naïve baseline algorithm VA-Uniform (described in Section V-C). The code to reproduce all the figures is available at https://github.com/Y-Hou/VA-BAI.git.

A. Experimental Design

By Theorem 1, the sample complexity of VA-LUCB is upper bounded by $O\left(H_{\rm VA}\ln(H_{\rm VA}/\delta)\right)$ w.h.p. We design four sets of test cases to empirically demonstrate the impact of the mean gaps Δ_i and the variance gap $\Delta_i^{\rm v}$ in $H_{\rm VA}$ on the sample complexity, in particular the smaller one of $\Delta_{i^\star}/2, \Delta_{i^\star}^{\rm v}$ will dominate the best feasible arm term and the greater one of $\Delta_i/2, \Delta_i^{\rm v}$ will dominate the suboptimal and infeasible arm term. The parameters that are varied in each test case are described below.

1. For the first term $\min\{\Delta_{i^\star}/2, \Delta_{i^\star}^{\rm v}\}^{-2}$,

(a). Under the condition that $\Delta_{i^\star}/2 \leq \Delta_{i^\star}^{\rm v}$, when $\Delta_{i^\star}$ and $\Delta_{i^{\star\star}}$ increase with the rest of the arms kept the same, $H_{\rm VA}$ and the sample complexity will decrease.

TABLE I

Parameter Settings for Instance $j \in [10]$. The Variance Gaps for the Infeasible Arms $\Delta_i^{\rm v} = \epsilon_j^{\rm v} = 0.233 - 0.003 \cdot j$ in Instance $j \in [10]$

$\bar{\sigma}^2 = 0.2, N = 10$		
arm	μ_i	σ_i^2
1	0.1	0.08
2	0.15	0.1
3	0.2	0.12
4	0.25	0.14
5	0.3	0.16
6	0.4	$\epsilon_j^{\rm v}$
7	0.45	$\epsilon_j^{\rm v}$
8	0.5	$\epsilon_j^{\rm v}$
9	0.55	$\epsilon_j^{\rm v}$
10	0.6	$\epsilon_j^{\rm v}$

(b). Under the condition that $\Delta_{i^\star}/2 \leq \Delta_{i^\star}^{\rm v}$, when $\Delta_{i^\star}^{\rm v}$ increases, $H_{\rm VA}$ and the sample complexity will be kept the same.

(c). Under the condition that $\Delta_{i^\star}/2 \geq \Delta_{i^\star}^{\rm v}$, as $\Delta_{i^\star}^{\rm v}$ increases, $H_{\rm VA}$ and the sample complexity will decrease.

(d). Under the condition that $\Delta_{i^\star}/2 \geq \Delta_{i^\star}^{\rm v}$, as $\Delta_{i^\star}$ and $\Delta_{i^{\star\star}}$ increase, $H_{\rm VA}$ and the sample complexity will decrease.

2. For the second term $\sum_{i\in\mathcal{F}\cap\mathcal{S}} 4\Delta_i^{-2}$, when $\Delta_{i^\star}$ and Δ_i for all $i \in \mathcal{F}\cap\mathcal{S}$ increase, $H_{\rm VA}$ and the sample complexity will decrease.

3. For the third term $\sum_{i\in\bar{\mathcal{F}}^c\cap\mathcal{R}}(\Delta_i^{\rm v})^{-2}$, when $\Delta_i^{\rm v}$ for all $i \in \bar{\mathcal{F}}^c\cap\mathcal{R}$ increase, $H_{\rm VA}$ and the sample complexity will decrease.

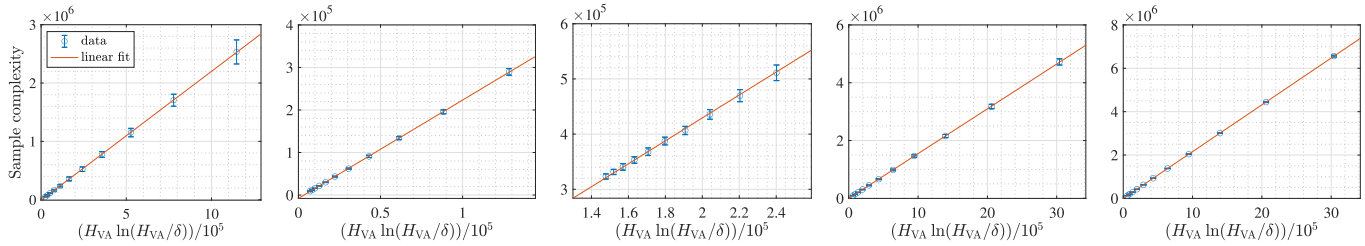
4. For the fourth term $\sum_{i\in\bar{\mathcal{F}}^c\cap\mathcal{S}} \max\{\Delta_i/2, \Delta_i^{\rm v}\}^{-2}$, the design is quite similar to Case 1, and thus the details are omitted here and presented in App. H-A.

The confidence parameter δ is set to be 0.05. In each case, there are 11 instances with $N = 20$ arms. The specific instances are described in detail in App. H-B. For each algorithm and instance, we run 20 independent trials to estimate the average time complexities and their standard deviations.

Note that there are 4 cases for the first term as we wish to elucidate that the smaller quantity between $\Delta_{i^\star}/2$ and $\Delta_{i^\star}^{\rm v}$ dominates the sample complexity of $i^\star$. The same experimental design applies to the study of the fourth term.

B. Performance of VA-LUCB

We plot the time complexities of Cases 1–3 with respect to $H_{\rm VA}\ln(H_{\rm VA}/\delta)$ in Figure 3; the rest of the figures are relegated to App. H-C. Although we do not prove the sample complexity grows linearly with $H_{\rm VA}\ln(H_{\rm VA}/\delta)$, this phenomenon can indeed be observed in our experiments. All the experimental results indicate the true sample complexity of VA-LUCB appears to be linear in $H_{\rm VA}\ln(H_{\rm VA}/\delta)$ (showing the tightness of our analyses) and is also bounded by $H_{\rm VA}\ln(H_{\rm VA}/\delta)$ and $3\,H_{\rm VA}\ln(H_{\rm VA}/\delta)$. The upper bound of $3\,H_{\rm VA}\ln(H_{\rm VA}/\delta)$ is usually sufficient for VA-LUCB to succeed.



(a) Case 1(a) (b) Case 1(c) (c) Case 1(d) (d) Case 2 (e) Case 3

Fig. 3. The time complexities for various cases with respect to $H_{\mathrm{VA}} \ln(H_{\mathrm{VA}}/\delta)$ with $\delta = 0.05$.

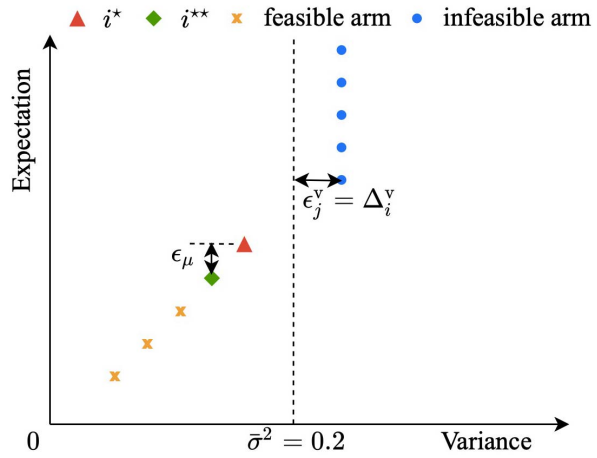


Fig. 4. An illustration of the instances.

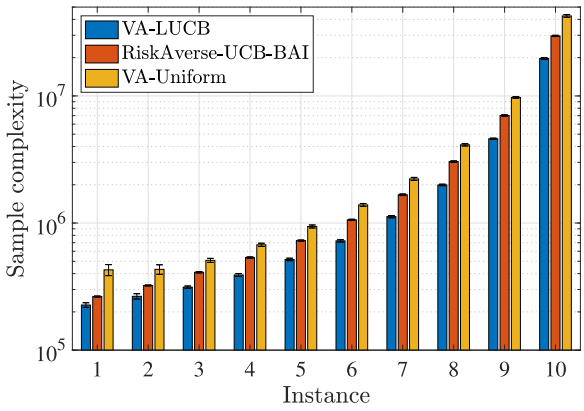


Fig. 5. Comparison among the time complexities of VA-LUCB, RiskAverse-UCB-BAI, and VA-Uniform (error bars denote 1 standard deviation across 20 runs). As the index of the instance increases, the instance becomes riskier.

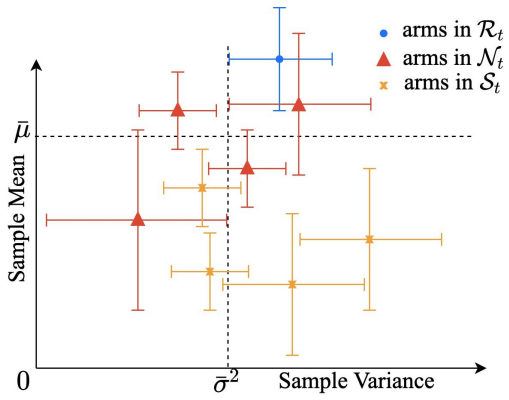


Fig. 6. Illustration of the confidence intervals of the empirically suboptimal and the empirically risky sets.

C. Comparison of VA-LUCB to RiskAverse-UCB-BAI [5] and VA-Uniform

We compare VA-LUCB to its closest competitor RiskAverse-UCB-BAI and VA-Uniform, which differs from VA-LUCB only in the sampling strategy. VA-Uniform uniformly samples two out of $N = 10$ arms at each time step. For comparison among the three algorithms, we construct 10 high-risk, high-reward instances with $N = 10$ arms in each instance to demonstrate that VA-LUCB outperforms a variant of RiskAverse-UCB-m-best [5] (named RiskAverse-UCB-BAI) and VA-Uniform in identifying the risky arms and the optimal feasible arm. We fix the feasible arms and the threshold $\bar{\sigma}^2$ and vary the variance gaps Δ_i^{v} of the infeasible arms. The accuracy parameters $\epsilon_\mu = \Delta_{i^\star}$ and $\epsilon_j^{\mathrm{v}} := \min_{i \in \mathcal{R} \setminus \{i^\star\}} \Delta_i^{\mathrm{v}}$ in instance j.[2] An illustration of the parameter setting of the arms is in Figure 4 and the specific parameters for the arms in instance $j \in [10]$ are presented in Table I. Note that the larger the index j, the riskier the instance as the true variances of the infeasible arms is closer to $\bar{\sigma}^2$ but their means are higher than that of the optimal feasible arm. This instance is apt for modeling real-world investment settings in which there may be several high-reward but risky options such as mini-bonds or cryptoassets, and several other low-reward but less risky options such as real estate (which appreciates with time with high probability).

[2]Our notation ϵ_j^{v} corresponds to ϵ_{v} in [5] under the variance-constrained setup (for the j^{th} instance).

The results are presented in Figure 5. VA-LUCB outperforms RiskAverse-UCB-BAI and VA-Uniform in all instances. In the riskiest instance considered (i.e., the one with the smallest ϵ_j^{v}), VA-LUCB requires $\approx 32\%$ fewer arm pulls compared to RiskAverse-UCB-BAI.

VI. Sketch of the Proof of Theorem 1

We extend the techniques used in the analysis of LUCB [9] to derive an upper bound on the sample complexity of VA-LUCB. To facilitate the analysis, define the *empirically suboptimal set*, *empirically risky set* and the complement of

their union respectively as

$$\mathcal{S}_t := \{i : U_i^{\mu}(t) < \bar{\mu}\}, \quad \mathcal{R}_t := \{i : L_i^{\mu}(t) > \bar{\mu}\}, \quad \text{and}$$
$$\mathcal{N}_t := [N] \setminus (\mathcal{S}_t \cup \mathcal{R}_t) = \{i : L_i^{\mu}(t) \le \bar{\mu} \le U_i^{\mu}(t)\}.$$

Note that $\mathcal{S}_t$ and $\mathcal{R}_t$ can be regarded as the empirical versions of $\mathcal{S}$ and $\mathcal{R}$ respectively. Intuitively, when t is large enough, $\mathcal{S}_t = \mathcal{S}$ and $\mathcal{R}_t = \mathcal{R}$. We illustrate these sets in Figure 6. Define the events

$$E_i^{\mu}(t) := \{|\hat{\mu}_i(t) - \mu_i| \le \alpha(t, T_i(t))\},$$
$$E_i^{\mathrm{v}}(t) := \{|\hat{\sigma}_i^2(t) - \sigma_i^2| \le \beta(t, T_i(t))\}, \quad \text{and}$$
$$E_i(t) := E_i^{\mu}(t) \bigcap E_i^{\mathrm{v}}(t), \quad \forall\, i \in [N]. \tag{17}$$

Finally, for $t \ge 2$, define

$$E(t) := \bigcap_{i \in [N]} E_i(t) \quad \text{and} \quad E := \bigcap_{t \ge 2} E(t). \tag{18}$$

Conditioned on E, we can show that the empirical mean and variance are accurate estimates of the true mean and variance respectively, in the sense that $\mu_i \in [L_i^{\mu}(t), U_i^{\mu}(t)]$ and $\sigma_i^2 \in [L_i^{\mathrm{v}}(t), U_i^{\mathrm{v}}(t)]$ for all $i \in [N]$ and $t \in \mathbb{N}$.

Lemma 1: Define E as in (18) with $\alpha(t,T)$ and $\beta(t,T)$ as in (3). Then E occurs with probability at least $1 - \delta/2$.

Lemma 2: Given an instance $(\nu, \bar{\sigma}^2)$ with confidence parameter δ, on the event $E(\tau)$, and the termination of VA-LUCB,

- if the instance is infeasible, $\hat{\mathsf{f}} = \mathsf{f} = 0$.
- if the instance is feasible, $i_{\mathrm{out}} = i_\tau = i_\tau^\star = i^\star, \hat{\mathsf{f}} = \mathsf{f} = 1$.

The proofs of the above lemmas are provided in App. D. Lemma 2 also justifies our stopping criterion.

What is left to do is to prove that VA-LUCB terminates at some finite time. We first state a useful core lemma, which constitutes the main workhorse of the entire argument that VA-LUCB succeeds upon termination.

Lemma 3: On the event $E(t)$, if VA-LUCB does not terminate, then at least one of the following statements holds:

- $i_t \in (\partial\mathcal{F}_t \backslash \mathcal{S}_t) \cup (\mathcal{F}_t \cap \mathcal{N}_t)$.
- $c_t \in (\partial\mathcal{F}_t \backslash \mathcal{S}_t) \cup (\mathcal{F}_t \cap \mathcal{N}_t)$.

The proof is presented in App. D. When VA-LUCB has not terminated, there are three possible scenarios. Firstly, the feasibility of the instance remains uncertain, i.e., $\mathcal{F}_t = \emptyset, \partial\mathcal{F}_t \ne \emptyset$. Secondly, the feasibility of i_t has not been confirmed, i.e., $i_t \in \partial\mathcal{F}_t$ and $i_t \ne i_t^\star$ (if $i_t^\star$ exists). Thirdly, the optimality of i_t has not been ascertained, i.e., $U_{c_t}^{\mu} \ge L_{i_t}^{\mu}$. Note that when only arm i_t is sampled, i.e., $|\bar{\mathcal{F}}_t| = 1$ or $U_{c_t}^{\mu} < L_{i_t}^{\mu}$, the optimality of i_t is guaranteed and we prove $i_t \in (\partial\mathcal{F}_t \backslash \mathcal{S}_t) \cup (\mathcal{F}_t \cap \mathcal{N}_t)$. Thus c_t does not need to be pulled at this time step. This strategy is essential in practice when the variances of the arms in $\mathcal{R}$ are much closer to the threshold $\bar{\sigma}^2$ compared to the arms in $\mathcal{S}$.

Lemma 3 indicates a sufficient condition for the termination of the algorithm. Namely, when neither of the arms i_t and c_t belongs to $(\partial\mathcal{F}_t \backslash \mathcal{S}_t) \cup (\mathcal{F}_t \cap \mathcal{N}_t)$, the algorithm must have terminated.

Next, we show that after sufficiently many pulls of each arm, the set $(\partial\mathcal{F}_t \backslash \mathcal{S}_t) \cup (\mathcal{F}_t \cap \mathcal{N}_t)$ remains nonempty with small probability. For a sufficient large t, let $u_i(t)$ be the smallest number of pulls of a suboptimal arm i such that $\alpha(t, u_i(t))$ is no greater than Δ_i, i.e.,

$$u_i(t) := \left\lceil \frac{1}{2\Delta_i^2} \ln\left(\frac{2Nt^4}{\delta}\right) \right\rceil$$

and $v_i(t)$ be the smallest number of pulls of an arm i such that $\beta(t, u_i(t))$ is no greater than Δ_i^{v}, i.e.,

$$v_i(t) := \left\lceil \frac{1}{2(\Delta_i^{\mathrm{v}})^2} \ln\left(\frac{2Nt^4}{\delta}\right) \right\rceil$$

Here we follow the convention: $\frac{1}{0} = +\infty$ and $\frac{1}{+\infty} = 0$, which may occur when $\mathcal{F} = \emptyset$ or $\mathcal{S} = \emptyset$.

Lemma 4: Using VA-LUCB, then 1) for $i^\star$,

$$\mathbb{P}[T_{i^\star}(t) > 16u_{i^\star}(t), i^\star \notin \mathcal{R}_t] \le \frac{\delta}{2(\frac{\Delta_{i^\star}}{2})^2 Nt^4} =: A_1(i^\star)$$

2) for any suboptimal arm $i \in \mathcal{S}$,

$$\mathbb{P}[T_i(t) > 16u_i(t), i \notin \mathcal{S}_t] \le \frac{\delta}{2(\frac{\Delta_i}{2})^2 Nt^4} =: A_2(i)$$

3) for any feasible arm $i \in \mathcal{F}$,

$$\mathbb{P}[T_i(t) > 4v_i(t), i \notin \mathcal{F}_t] \le \frac{\delta}{2(\Delta_i^{\mathrm{v}})^2 Nt^4} =: A_3(i)$$

4) for any infeasible arm $i \in \bar{\mathcal{F}}^c$,

$$\mathbb{P}[T_i(t) > 4v_i(t), i \notin \bar{\mathcal{F}}_t^c] \le \frac{\delta}{2(\Delta_i^{\mathrm{v}})^2 Nt^4} =: A_4(i)$$

For a suboptimal arm i, note that $\Delta_i/2 \le \bar{\mu} - \mu_i \le \Delta_i$. We compute $\mathbb{P}[T_i(t) > 16u_i(t), i \notin \mathcal{S}_t]$ in the same approach as [9]. This method is also utilized to analyze the variances.

Lemma 4 indicates the following:

- For the best feasible arm $i^\star$ (if it exists), after sampling it $\max\{16u_{i^\star}(t), 4v_{i^\star}(t)\}$ times, by using a union bound, $i^\star \in \mathcal{F}_t \cap \mathcal{R}_t$ with failure probability at most $B_1(i^\star) := A_1(i^\star) + A_3(i^\star)$. Therefore $i^\star \notin (\mathcal{F}_t \cap \mathcal{N}_t) \cup (\partial\mathcal{F}_t \backslash \mathcal{S}_t)$.
- For any feasible and suboptimal arm $i \in \mathcal{F} \cap \mathcal{S}$, when $T_i(t) > 16u_i(t)$, $i \in \mathcal{S}_t$ with failure probability at most $B_2(i) := A_2(i)$.
- For any arms $i \in \bar{\mathcal{F}}^c \cap \mathcal{R}$, when $T_i(t) > 4v_i(t)$, $i \in \bar{\mathcal{F}}_t^c$ with failure probability at most $B_3(i) := A_4(i)$.
- For arm $i \in \bar{\mathcal{F}}^c \cap \mathcal{S}$, if it has been pulled more than $\min\{16u_i(t), 4v_i(t)\}$ times, $i \in \bar{\mathcal{F}}_t^c \cup \mathcal{S}_t$ with failure probability at most $B_4(i) := A_2(i) \cdot \mathbb{1}\{16u_i(t) < 4v_i(t)\} + A_4(i) \cdot \mathbb{1}\{16u_i(t) \ge 4v_i(t)\}$.

In conclusion, if all arms are pulled sufficiently many times, the probability that any of them stays in the set $(\mathcal{F}_t \cap \mathcal{N}_t) \cup (\partial\mathcal{F}_t \backslash \mathcal{S}_t)$ is upper bounded by

$$B_1(i^\star) + \sum_{i \in \mathcal{F} \cap \mathcal{S}} B_2(i) + \sum_{i \in \bar{\mathcal{F}}^c \cap \mathcal{R}} B_3(i) + \sum_{i \in \bar{\mathcal{F}}^c \cap \mathcal{S}} B_4(i). \tag{19}$$

Finally, based on the above lemmas, we show that the algorithm dose not terminate with small probability after time $t^\star = O(H_{\mathrm{VA}} \ln(H_{\mathrm{VA}}/\delta))$.

Lemma 5: Let $t^\star = 152 H_{\mathrm{VA}} \ln(H_{\mathrm{VA}}/\delta)$. At any time step $t > t^\star$, the probability that Algorithm 1 does not terminate is at most $5\delta/t^2$.

According to Lemma 3, when neither arm i_t nor c_t belongs to $(\partial\mathcal{F}_t\backslash\mathcal{S}_t) \cup (\mathcal{F}_t \cap \mathcal{N}_t)$, the algorithm stops. In particular, if none of arms in $[N]$ is in $(\partial\mathcal{F}_t\backslash\mathcal{S}_t)\cup(\mathcal{F}_t\cap\mathcal{N}_t)$, the algorithm must terminate, which can be guaranteed by Lemma 4 with failure probability at most (19). The complete proof involves counting the numbers of pulls of the arms and estimating $t^\star$. This is presented in App. D.

VII. Conclusion and Future Direction

We proposed framework for the risk-constrained Best Arm Identification problem and also developed an algorithm VA-LUCB whose sample complexity is almost optimal in the sense that its upper bound almost matches the information-theoretic lower bound. We highlight the VA-LUCB Algorithm constitutes a convenient framework to tackle any risk-aware BAI problem in the sense that it is compatible with other concentration bounds, including LIL bounds.

However, we believe it is hard to derive an exact sample complexity using confidence bound-based algorithms, in the sense of nailing down the exact number

$$\liminf_{\delta\downarrow 0} \frac{\tau_\delta^\star}{\log\frac{1}{\delta}},$$

where $\tau_\delta^\star$ is the minimum expectation of the stopping time for an algorithm to be δ-PAC.

To characterize the exact asymptotic sample complexity, we have explored adapting tracking-based algorithms such as Track and Stop (T&S) from [10] to the variance-constrained BAI problem. A lower bound similar to [10] can be derived. For the corresponding algorithm, since the variances and the bound on the variance complicate the alternative instances $\mathrm{Alt}(\nu,\bar\sigma^2)$ for a given instance $(\nu,\bar\sigma^2)$, the optimization to obtain the optimal proportion of the arm pulls is difficult. In particular, the (allocation vector) w that attains the supremum in

$$\sup_{w:w_i>0\,\forall\, i\in[K],\sum_{i=1}^K w_i=1} \inf_{(\nu',\bar\sigma^2)\in\mathrm{Alt}(\nu,\bar\sigma^2)} \sum_{i=1}^N w_i \mathrm{KL}(\nu_i,\nu_i')$$

is difficult to characterize even for Gaussians because the variances (in addition to the means) are now *variables* in the inner optimization. This complicates the design and analysis of a *constrained* T&S-like algorithm, especially the sampling strategy. This is an promising direction for future research.

Appendices

In Appendix A, we discuss the necessity of the assumption $\sigma_{i^\star}^2 < \bar\sigma^2$. In Appendix B, a variant of RiskAverse-UCB-m-best [5] is presented. In Appendix C, we systematically compare the bounds on the sample complexity presented in this paper to those in [5]. We also compare the assumptions needed to output the best feasible arm. In Appendix D, we provide the detailed proofs of the lemmas used to prove Theorem 1. In Appendix E, VA-LUCB is extended to VA-LUCB-sub-Gaussian, which deals with arms following σ-sub-Gaussian distributions. In Appendix F, we discuss how to modify the analysis of VA-LUCB when the confidence radii are designed based on the non-asymptotic LIL (cf. Remark 1). In Appendix G, the complete proofs of Theorem 1 and Corollary 1 are presented. In Appendix H, specific parameter settings and additional numerical results are presented.

Appendix A
Discussion of the Case $\sigma_{i^\star}^2 = \bar\sigma^2$

We assume $\sigma_{i^\star}^2 < \bar\sigma^2$ in Section II such that the problem is solvable by applying confidence-bound techniques without knowledge of any unknown parameter. We provide an explanation in this section. Given a permissible bound on the variance $\bar\sigma^2$, it is natural to define the feasible set as

$$\begin{aligned}&\mathcal{F} := \{i\in[N] : \sigma_i^2 < \bar\sigma^2\} \quad \text{or}\\ &\mathcal{F} := \{i\in[N] : \sigma_i^2 \le \bar\sigma^2\}.\end{aligned} \tag{S.1}$$

First, with either choice of definition of $\mathcal{F}$, inspired by Lemma 6, to ascertain there is no feasible arm and to terminate, an algorithm needs to check either

$$\begin{aligned}&\{i\in[N] : U_i^{\mathrm{v}}(t) \le \bar\sigma^2\} = \emptyset \quad \text{or}\\ &\{i\in[N] : L_i^{\mathrm{v}}(t) \le \bar\sigma^2\} = \emptyset.\end{aligned}$$

Since the feasible arms do not satisfy $U_i^{\mathrm{v}}(t) \le \bar\sigma^2$ w.h.p. in the beginning, it is only reasonable to ascertain there is no feasible arm and terminate the algorithm when $\{i\in[N] : L_i^{\mathrm{v}}(t) \le \bar\sigma^2\} = \emptyset$ as in our algorithm.

Note that with either choice of $\mathcal{F}$ in (S.1), we are confident ascertaining that an arm is feasible if $U_i^{\mathrm{v}}(t) \le \bar\sigma^2$ and is infeasible if $L_i^{\mathrm{v}}(t) > \bar\sigma^2$. We can only say an arm is possibly feasible with only $L_i^{\mathrm{v}}(t) \le \bar\sigma^2$. Our termination rule is $\bar{\mathcal{F}}_t \cap \mathcal{P}_t = \emptyset$, where $\mathcal{F}_t$, $\bar{\mathcal{F}}_t$, and $\mathcal{P}_t$ are defined as in (6)–(8) and repeated here for easy reference:

$$\begin{aligned}&\mathcal{F}_t = \{i : U_i^{\mathrm{v}}(t) \le \bar\sigma^2\},\\ &\bar{\mathcal{F}}_t = \{i : L_i^{\mathrm{v}}(t) \le \bar\sigma^2\},\\ &\mathcal{P}_t = \begin{cases}\{i : U_i^{\mu}(t) \ge L_{i_t^\star}^{\mu}(t), i\ne i_t^\star\}, & \mathcal{F}_t\ne\emptyset\\ [N], & \mathcal{F}_t = \emptyset\end{cases},\end{aligned}$$

where $i_t^\star = \arg\max_{i\in\mathcal{F}_t}\hat\mu_i(t)$.

Next, we discuss each possible choice of $\mathcal{F}$ in (S.1) individually.

Choice 1: $\mathcal{F} = \{i\in[N] : \sigma_i^2 < \bar\sigma^2\}$. Consider a case where there is an infeasible arm j with $\sigma_j^2 = \bar\sigma^2$ and $\mu_j > \mu_{i^\star}$. After pulling arms for a large number of times, w.h.p., we have

$$\begin{aligned}&L_j^{\mathrm{v}}(t) < \bar\sigma^2 < U_j^{\mathrm{v}}(t) \implies j\in\bar{\mathcal{F}}_t\setminus\mathcal{F}_t \quad \text{and}\\ &L_j^{\mu}(t) > U_{i^\star}^{\mu}(t) > L_{i^\star}^{\mu}(t),\quad U_i^{\mathrm{v}}(t)\le\bar\sigma^2 \quad \forall\, i\in\mathcal{F}\\ &\implies i^\star\in\mathcal{F}\subset\mathcal{F}_t,\quad j\in\mathcal{P}_t,\end{aligned}$$

which implies that $j\in\bar{\mathcal{F}}_t\cap\mathcal{P}_t$, and hence $\bar{\mathcal{F}}_t\cap\mathcal{P}_t\ne\emptyset$. In other words, the algorithm will never terminate w.h.p.

Choice 2: $\mathcal{F} = \{i\in[N] : \sigma_i^2 \le \bar\sigma^2\}$. Consider a case where $\sigma_{i*}^2 = \bar\sigma^2$. Similar to the discussion above, we can see that $i^\star\notin\mathcal{F}_t, i^\star\in\bar{\mathcal{F}}_t\cap\mathcal{P}_t$, and hence $\bar{\mathcal{F}}_t\cap\mathcal{P}_t\ne\emptyset$ w.h.p. Therefore, the algorithm will not terminate w.h.p.

Altogether, under either choice of the definition of the feasible set, any algorithm using UCB- and LCB-based termination rule will not terminate w.h.p. when there exists an

arm i with high expectation and $\sigma_i^2 = \bar{\sigma}^2$. Thus, we define $\mathcal{F} = \{i \in [N] : \sigma_i^2 \leq \bar{\sigma}^2\}$ and assume $\sigma_{i^\star}^2 < \bar{\sigma}^2$ so that the algorithm will terminate in a finite number of time steps w.h.p. when a confidence bound-based algorithm is employed.

Additional Prior Knowledge: We note that the variant of RiskAverse-UCB-m-best algorithm proposed by David et al. [5], RiskAverse-UCB-BAI, can also be applied to identify the best feasible arm (without any suboptimality or subfeasibility) under the δ-PAC framework only when $\min_{i\in\mathcal{R}\setminus\{i^\star\}} \Delta_i^{\mathrm{v}}$ is known (see Appendix C for detailed discussion). Though it can be applied when $\sigma_{i^\star}^2 = \bar{\sigma}^2$, we remark that our algorithm can also handle this case ($\sigma_{i^\star}^2 = \bar{\sigma}^2$) with such additional prior knowledge on the parameters. In detail, we regard ϵ_{v} be an optional parameter of our algorithm (set as it to be 0 if $\min_{i\in\mathcal{R}\setminus\{i^\star\}} \Delta_i^{\mathrm{v}}$ is unknown) and define

$$\bar{\mathcal{F}}_t := \{i : L_i^{\mathrm{v}}(t) \leq \bar{\sigma}^2\}, \quad \mathcal{F}_t := \{i \in \bar{\mathcal{F}}_t : U_i^{\mathrm{v}}(t) \leq \bar{\sigma}^2 + \epsilon_{\mathrm{v}}\}.$$

We set $\epsilon_{\mathrm{v}} := \min_{i\in\mathcal{R}\setminus\{i^\star\}} \Delta_i^{\mathrm{v}}$ when the quantity is known and we are not sure if $\sigma_{i^\star}^2 < \bar{\sigma}^2$ (i.e., it is possible that $\sigma_{i^\star}^2 = \bar{\sigma}^2$). With the prior knowledge of $\epsilon_{\mathrm{v}} = \min_{i\in\mathcal{R}\setminus\{i^\star\}} \Delta_i^{\mathrm{v}}$, the upper bound of the sample complexity of VA-LUCB can be improved to $O\big(\widetilde{H}_{\mathrm{VA}} \ln \frac{\widetilde{H}_{\mathrm{VA}}}{\delta}\big)$ w.h.p., where

$$\widetilde{H}_{\mathrm{VA}} := \frac{1}{\min\{\frac{\Delta_{i^\star}}{2}, \Delta_{i^\star}^{\mathrm{v}} + \epsilon_v\}^2} + \sum_{i\in\mathcal{F}\cap\mathcal{S}} \frac{1}{(\frac{\Delta_i}{2})^2} + \sum_{i\in\bar{\mathcal{F}}^c\cap\mathcal{R}} \frac{1}{(\Delta_i^{\mathrm{v}})^2} + \sum_{i\in\bar{\mathcal{F}}^c\cap\mathcal{S}} \frac{1}{\max\{\frac{\Delta_i}{2}, \Delta_i^{\mathrm{v}}\}^2} < H_{\mathrm{VA}}.$$

This enables us not only to deal with the case $\sigma_{i^\star} = \bar{\sigma}^2$, but also facilitates in ascertaining the feasibility of the best feasible arm in the usual instance in which $\sigma_{i^\star}^2 < \bar{\sigma}^2$. Therefore, the sample complexity is better than the current VA-LUCB algorithm, as well as RiskAverse-UCB-BAI algorithm to be discussed extensively in Appendix C. The proof just follows the same procedure as in Section VI.

APPENDIX B
RISKAVERSE-UCB-BAI

We present a variant of RiskAverse-UCB-m-best algorithm from [5], named RiskAverse-UCB-BAI, which is adapted to our variance-constrained BAI setup. To avoid any confusion, we redefine the sample mean, sample variance and confidence bounds, which are consistent with the notations in [5]. For arm $i \in [N]$, define

- the counter:
$$T_i'(t) := \sum_{s=1}^{t} \mathbb{1}\{i = i_s^\dagger\}; \tag{S.2}$$
- the sample mean and the sample variance respectively as:
$$\hat{\mu}_i'(t) := \frac{1}{T_i'(t)} \sum_{s=1}^{t} X_{s,i_s^\dagger} \mathbb{1}\{i = i_s^\dagger\}, \quad \text{and}$$
$$(\hat{\sigma}_i')^2(t) := \frac{\sum_{s=1}^{t} \big(X_{s,i_s^\dagger} - \hat{\mu}_i(t)\big)^2 \mathbb{1}\{i = i_s^\dagger\}}{T_i'(t) - 1};$$

Algorithm 2 RiskAverse-UCB-BAI (A Variant of RiskAverse-UCB-m-Best [5])

Input: threshold $\bar{\sigma}^2 > 0$, confidence parameter $\delta \in (0,1)$ and accuracy parameters $\epsilon_\mu, \epsilon_{\mathrm{v}} \in (0, +\infty)$.
for each $i \in [N]$ **do**
  Sample arm i twice, update the counter $T_i'(N)$, the sample mean $\hat{\mu}_i'(N)$ and sample variance $(\hat{\sigma}_i')^2(N)$.
end for
Set $H := 3N\left(\frac{1}{2\epsilon_\mu^2} + \frac{4}{\epsilon_{\mathrm{v}}^2}\right) \ln\left(\frac{6N}{\delta} N \left(\frac{1}{2\epsilon_\mu^2} + \frac{4}{\epsilon_{\mathrm{v}}^2}\right)\right)$ and $t := 2N$.
repeat
  Set $\bar{\mathcal{F}}_t' := \{i : L_i^{\mathrm{v}'}(t) \leq \bar{\sigma}^2\}$.
  Select an optimistic arm $i_{t+1}^\dagger := \mathrm{argmax}_{i\in\bar{\mathcal{F}}_t'} U_i^{\mu'}(t)$.
  Draw a sample from the selected arm $i_t^\dagger$.
  Set $t = t + 1$.
  Update the counter $T_i'(t)$ and estimates $\hat{\mu}_i'(t)$ and $(\hat{\sigma}_i')^2(t)$ for all arm $i \in [N]$ accordingly.
until ([$f_\mu(T_{i_t^\dagger}'(t)) \leq \epsilon_\mu/2$ and $(\hat{\sigma}_{i_t^\dagger}')^2(T_i'(t)) - \epsilon_{\mathrm{v}} \leq \bar{\sigma}^2$] or $t \geq H$).
Return $i_t^\dagger$

- the confidence radii for the mean and variance respectively as
$$f_\mu(T) := \sqrt{\frac{1}{2T} \ln\left(\frac{6HN}{\delta}\right)}, \quad \text{and} \tag{S.3}$$
$$f_{\mathrm{v}}(T) := \sqrt{\frac{2}{T} \ln\left(\frac{6HN}{\delta}\right)};$$
- the confidence bounds for the mean:
$$L_i^{\mu'}(t) := \hat{\mu}_i'(t) - f_\mu(T_i'(t)), \quad \text{and}$$
$$U_i^{\mu'}(t) := \hat{\mu}_i'(t) + f_\mu(T_i'(t));$$
- the confidence bounds for the variance:
$$L_i^{\mathrm{v}'}(t) := (\hat{\sigma}_i')^2(t) - f_{\mathrm{v}}(T_i'(t)), \quad \text{and}$$
$$U_i^{\mathrm{v}'}(t) := (\hat{\sigma}_i')^2(t) + f_{\mathrm{v}}(T_i'(t)).$$

The algorithm, RiskAverse-UCB-BAI, an adaptation of RiskAverse-UCB-m-best [5] to our variance-constrained setting, is presented in Algorithm 2. Since Algorithm 2 only guarantees to output an ϵ_{v}-*approximately feasible and* ϵ_μ-*approximately optimal* arm, in order to output the best feasible arm, the accuracy parameters have to be sufficiently small such that the only ϵ_{v}-*approximately feasible and* ϵ_μ-*approximately optimal* arm is the best feasible arm. See Appendix C for details.

Remark 3: We remark that Algorithm 1 can also be adapted to the BAI problem with an α-quantile constraint by replacing the sample variance and its associated confidence bound by the sample α-quantile and the corresponding concentration bound (see [5, Lemma 6]). The modified Algorithm 1 is completely parameter-free, whereas RiskAverse-UCB-m-best [5] is not. The sample complexity of the modified Algorithm 1 can be derived in a similar procedure as in this paper.

APPENDIX C
DISCUSSION OF THE BOUNDS IN DAVID ET AL. [5]

For RiskAverse-UCB-BAI to identify the best feasible arm (without any suboptimality or subfeasibility) under the δ-PAC framework, it needs to ensure that parameters $\epsilon_{\rm v}$ and ϵ_μ are set sufficiently small so that the $\epsilon_{\rm v}$-approximately feasible and ϵ_μ-approximately optimal arm is exactly the best feasible arm. A sufficient condition is $\epsilon_\mu < \Delta_{i^\star}$ and $\epsilon_{\rm v} < \min_{i\in\mathcal{R}\setminus\{i^\star\}} \Delta_i^{\rm v}$. Without the former/latter condition, a suboptimal/risky arm maybe produced by the RiskAverse-UCB-BAI. However, even we relax the accuracy parameters by allowing them to assume equality, i.e., $\epsilon_\mu = \Delta_{i^\star}$ and $\epsilon_{\rm v} = \min_{i\in\mathcal{R}} \Delta_i^{\rm v}$, as well as that H is given, we can still assert that VA-LUCB is superior in terms of the sample complexity; this is what we do in Section C-A. In addition, the confidence radii of the mean and variance (S.3) contain H, the hardness parameter that depends on the instance which is not known in practice, further underscoring that RiskAverse-UCB-BAI is not parameter free. In the following discussion, we recall that random variables bounded in $[0,1]$ are $1/2$-subgaussian.

A. Discussion of the Upper Bounds

The upper bound of the sample complexity of a variant of RiskAverse-UCB-m-best presented in [5, Theorem 3], which we call RiskAverse-UCB-BAI, and analyze using techniques along the same lines is

$$\sum_{i\in[N]} C_i \ln\left(\frac{6NH}{\delta}\right) \tag{S.4}$$

where

$$H := 3N\left(\frac{1}{2\epsilon_\mu^2}+\frac{4}{\epsilon_{\rm v}^2}\right)\ln\left(\frac{6N}{\delta}N\left(\frac{1}{2\epsilon_\mu^2}+\frac{4}{\epsilon_{\rm v}^2}\right)\right) \tag{S.5}$$

and

$$C_i := \min\left\{\frac{1}{\max\{0,\mu_{i^\star}-\mu_i\}^2}, \frac{4}{\max\{0,\sigma_i^2-\bar\sigma^2\}^2}, \max\left\{\frac{1}{\epsilon_\mu^2}, \frac{4}{\max\{0,\epsilon_{\rm v}-(\sigma_i^2-\bar\sigma^2)\}^2}\right\}\right\} \tag{S.6}$$

for all $i\in[N]$. For the sake of brevity, define $H' := 3N\left(\frac{1}{2\epsilon_\mu^2}+\frac{4}{\epsilon_{\rm v}^2}\right)$. Then the upper bound in (S.4) can be rewritten as $\sum_{i\in[N]} C_i \ln\left(\frac{6NH'\ln(\frac{2NH'}{\delta})}{\delta}\right)$. Given the similar roles of $H_{\rm VA}$ and H' in the upper bounds, we can also regard H' as another hardness parameter in [5] (in addition to $\{C_i\}_{i\in[N]}$). Since both $\sum_{i\in[N]} C_i$ and H appear in the upper bound (S.4), we carefully compare both of them to $H_{\rm VA}$. We firstly compare the terms in $H_{\rm VA}$ with C_i:

- For arm $i^\star$,
$$C_{i^\star} = \max\left\{\frac{1}{\epsilon_\mu^2}, \frac{4}{(\epsilon_{\rm v}-(\sigma_{i^\star}^2-\bar\sigma^2))^2}\right\} \geq \frac{1}{\min\left\{\Delta_{i^\star}, \Delta_{i^\star}^{\rm v}\right\}^2}$$
where equality holds if $\Delta_{i^\star}^{\rm v} = \epsilon_{\rm v}$.
- For any feasible and suboptimal arm i, $C_i = \frac{1}{\Delta_i^2}$.
- For any risky arm $i\neq i^\star$, $C_i = \frac{4}{(\Delta_i^{\rm v})^2}$.
- For any infeasible and suboptimal arm i,
$$C_i = \min\left\{\frac{1}{\Delta_i^2}, \frac{4}{(\Delta_i^{\rm v})^2}\right\} = \frac{1}{\max\left\{\Delta_i, \Delta_i^{\rm v}/2\right\}^2}.$$

This trivially leads to $\sum_{i\in[N]} C_i \geq H_{\rm VA}/4$. In terms of H, note that $\epsilon_\mu = \Delta_{i^\star}, \epsilon_{\rm v} = \min_{i\in\mathcal{R}} \Delta_i^{\rm v}$, so $H' > 3H_{\rm VA}/8$ trivially holds. However, in a practical instance where the means and variances of the arms are diverse, $H' > H_{\rm VA}$, e.g., when $\Delta_{i^\star} \geq \epsilon_{\rm v}/2$ (this can be interpreted as the scenario in which identifying the risky arms is more difficult than ascertaining the optimality of the best feasible arm) or there are at most $\lceil N/6\rceil$ suboptimal arms with $\Delta_i \leq 2\Delta_{i^\star}$, $H' > H_{\rm VA}$ holds. Therefore, the upper bound in [5] is

$$\begin{aligned}
&\underbrace{\sum_{i\in[N]} C_i \ln\left(\frac{6NH'\ln(\frac{2NH'}{\delta})}{\delta}\right)}_{\text{Upper bound (UB) in [5]}} \\
&= \underbrace{\Omega\left(H_{\rm VA}\ln\left(\frac{NH_{\rm VA}\ln(\frac{NH_{\rm VA}}{\delta})}{\delta}\right)\right)}_{\text{Order-wise result of UB in [5]}} \\
&= \omega\left(\underbrace{H_{\rm VA}\ln\left(\frac{H_{\rm VA}}{\delta}\right)}_{\text{Our upper bound up to constants}}\right).
\end{aligned} \tag{S.7}$$

Even disregarding constants and the fact that RiskAverse-UCB-BAI is not parameter free if we demand that the (strictly) best feasible arm is output by the algorithm, we note the presence of the additional $\ln$ term in N and the $\ln\ln$ term in $NH_{\rm VA}/\delta$ in the order-wise result of the upper bound in [5]. We conclude that the upper bound of the sample complexity of RiskAverse-UCB-BAI in (S.4) [5] is strictly larger in order than ours.

B. Discussion of the Lower Bounds

While the lower bound of [5, Theorem 2] holds under a set of assumptions, we assume that these assumptions are generally not needed and the only assumption made here is that $\delta\in(0,0.01)$. The lower bound in [5] is

$$\min_{i'\in[N]} \sum_{i\in[N]\setminus\{i'\}} \frac{\ln\left(\frac{1}{9\delta}\right)}{900} \max\left\{(16\Delta_i^{\rm v})^2, (5\Delta_{i^\star}+4\max\{0,\mu_{i^\star}-\mu_i\})^2\right\}^{-1} - N. \tag{S.8}$$

We also compare the denominator term-by-term:

- For arm $i^\star$,
$$\begin{aligned}
&\max\left\{(5\Delta_{i^\star}+4\max\{0,\mu_{i^\star}-\mu_{i^\star}\})^2, (16\Delta_{i^\star}^{\rm v})^2\right\} \\
&= \max\left\{5\Delta_{i^\star}, 16\Delta_{i^\star}^{\rm v}\right\}^2 \geq \min\left\{5\Delta_{i^\star}, 16\Delta_{i^\star}^{\rm v}\right\}^2
\end{aligned}$$
where equality holds if and only if $5\Delta_{i^\star} = 16\Delta_{i^\star}^{\rm v}$.
- For any feasible and suboptimal arm i,
$$\begin{aligned}
&\max\left\{(5\Delta_{i^\star}+4\max\{0,\mu_{i^\star}-\mu_i\})^2, (16\Delta_i^{\rm v})^2\right\} \\
&= \max\left\{5\Delta_{i^\star}+4\Delta_i, 16\Delta_i^{\rm v}\right\}^2 > 16\Delta_i^2.
\end{aligned}$$

- For any risky arm $i \neq i^\star$,
$$\max\left\{\left(5\Delta_{i^\star}+4\max\{0,\mu_{i^\star}-\mu_i\}\right)^2,\left(16\Delta_i^{\mathrm{v}}\right)^2\right\} = \max\left\{5\Delta_{i^\star},16\Delta_i^{\mathrm{v}}\right\}^2 \geq \left(16\Delta_i^{\mathrm{v}}\right)^2.$$
- For any infeasible and suboptimal arm i,
$$\max\left\{\left(5\Delta_{i^\star}+4\max\{0,\mu_{i^\star}-\mu_i\}\right)^2,\left(16\Delta_i^{\mathrm{v}}\right)^2\right\} = \max\left\{5\Delta_{i^\star},16\Delta_i^{\mathrm{v}}\right\}^2.$$

Therefore, the lower bound (S.8) is strictly smaller than

$$\left(\frac{1}{\min\left\{5\Delta_{i^\star},16\Delta_{i^\star}^{\mathrm{v}}\right\}^2}+\sum_{i\in\mathcal{F}\cap\mathcal{S}}\frac{1}{16\Delta_i^2}+\sum_{i\in\bar{\mathcal{F}}^c\cap\mathcal{R}}\frac{1}{(16\Delta_i^{\mathrm{v}})^2}+\sum_{i\in\bar{\mathcal{F}}^c\cap\mathcal{S}}\frac{1}{\max\left\{5\Delta_{i^\star},16\Delta_i^{\mathrm{v}}\right\}^2}\right)\frac{\ln\left(\frac{1}{9\delta}\right)}{900}.$$

This is strictly smaller than our lower bound in (16) (see Theorem 2) when $\bar{\sigma}^2 \geq 9\cdot 10^{-5}$ and $\mu_{i^\star} \leq 1-9\cdot 10^{-5}$, which is a large subset of practical instances (recalling that the rewards are bounded in $[0,1]$). In fact, the space of $(\mu,\bar{\sigma}^2)$ for which our lower bound is strictly better than that in [5] is $> 1-7.2\times 10^{-4} > 99.9\%$ times of the total area of the permissible parameter space of $(\mu,\bar{\sigma}^2)\in[0,1]\times[0,1/4)$.

Considering both the upper and lower bounds, as well as Corollary 1, even though we have relaxed several assumptions in [5], the bounds in [5] (performance upper bound on the sample complexity of RiskAverse-UCB-BAI and lower bound) are looser than ours. Furthermore, the term of arm i in the lower bound (S.8) does not match the corresponding term in the upper bound (S.4) or H', showing that the terms $\sum_{i\in[N]}C_i$ and H' do not characterize the inherent difficulty of identifying the best feasible arm. In contrast, we have shown that the optimal sample complexity of identifying the best feasible arm is characterized exactly by H_{VA}.

C. Discussion of the Complexity of Identifying Risky Arms

Since we are considering risk-constrained bandits, an important task is to identify the risky arms as quickly as possible. Based on the discussion of the accuracy parameters in the previous sections, we investigate the convergence speed of the confidence radius of the variance, which is essential in eliminating the risky arms. The confidence radius of risky arm $i\neq i^\star$ at round t in [5] is

$$f_{\mathrm{v}}(T_i'(t)) = \sqrt{\frac{2}{T_i'(t)}\ln\frac{6HN}{\delta}} \tag{S.9}$$

where $T_i'(t)$ is defined in (S.2). We claim that (S.9) is strictly greater than β in (3) in a generic case as follows:[3]

$$f_{\mathrm{v}}(T_i(t)) = \sqrt{\frac{2}{T_i(t)}\ln\frac{6HN}{\delta}} \geq \beta(t,T_i(t)) = \sqrt{\frac{1}{2T_i(t)}\ln\left(\frac{2Nt^4}{\delta}\right)} \tag{S.10}$$

This is equivalent to

$$4\ln\frac{6HN}{\delta}\geq\ln\frac{2Nt^4}{\delta} \iff \left(\frac{6HN}{\delta}\right)^4\geq\frac{2Nt^4}{\delta} \iff \frac{6}{2^{1/4}}\left(\frac{N}{\delta}\right)^{3/4}H'\ln\frac{2NH'}{\delta}\geq 152H_{\mathrm{VA}}\ln\frac{H_{\mathrm{VA}}}{\delta},$$

if $N\geq 10$ and $\delta\leq 0.05$, $\frac{6}{2^{1/4}}\left(\frac{N}{\delta}\right)^{3/4}>152$. According to the comparison between H' and H_{VA} in Section C-A, $H'>H_{\mathrm{VA}}$ in a general instance, thus the last inequality holds even we ignore the logarithmic term in N. In particular, the greater N is and the more risky arms, the larger $f_{\mathrm{v}}(T_i(t))$ is. Hence, the convergence speed of the confidence radius in β will be faster than that of (S.9), resulting fewer arm pulls of the risky arms of VA-LUCB. This indicates VA-LUCB is more efficient in the risk-constrained setup. This is also corroborated by our experiments in Section V-C.

From the experimental/practical point of view, the constant 152 on the right-hand side of (S.10) can essentially be replaced by 3 as our experiments indicate; see Section V-B. Furthermore, both algorithms will identify the risky arms first, thus t in β is small at the beginning. We refer to Section V-C for further experimental validations.

Appendix D
Proof of Upper Bound

Lemma 6 (Implication of Hoeffding's and McDiarmid's Inequalities): Given an instance $(\nu,\bar{\sigma}^2)$, for any arm i with $T_i(t)\geq 2$ and $\varepsilon>0$ we have

$$\begin{aligned}\mathbb{P}[\hat{\mu}_i(t)-\mu_i\geq\varepsilon]&\leq\exp(-2T_i(t)\varepsilon^2),\\ \mathbb{P}[\hat{\mu}_i(t)-\mu_i\leq-\varepsilon]&\leq\exp(-2T_i(t)\varepsilon^2),\end{aligned} \tag{S.11}$$

and

$$\begin{aligned}\mathbb{P}[\hat{\sigma}_i^2(t)-\sigma_i^2\geq\varepsilon]&\leq\exp(-2T_i(t)\varepsilon^2),\\ \mathbb{P}[\hat{\sigma}_i^2(t)-\sigma_i^2\leq-\varepsilon]&\leq\exp(-2T_i(t)\varepsilon^2).\end{aligned} \tag{S.12}$$

Proof: Note that since the reward distribution is bounded in $[0,1]$, (S.11) can be derived by a straightforward application of Hoeffding's inequality. As for the sample variance, note that for i.i.d. random variables $X_1,\ldots,X_n$, supported on $[0,1]$ with sample mean $\hat{\mu}$, the unbiased sample variance can be written as

$$\hat{\sigma}^2(n)=\frac{1}{n-1}\sum_{i=1}^n(X_i-\hat{\mu})^2 = \frac{1}{n(n-1)}\sum_{i<j}(X_i-X_j)^2=:f(X_1,X_2,\ldots,X_n)$$

Note by the unbiasedness of the sample variance that $\mathbb{E}[\hat{\sigma}^2(n)]=\sigma^2=\mathrm{Var}(X_i)$ and

$$\left|f(x_1,\ldots,x_{i-1},x_i,x_{i+1},\ldots,x_n) - f(x_1,\ldots,x_{i-1},x_i',x_{i+1},\ldots,x_n)\right|\leq\frac{1}{n}$$

[3]Due to the difference in algorithms (VA-LUCB vs. RiskAverse-UCB-BAI) and the definitions of $T_i(t)$ (in this paper and in [5]), one should use $T_i'(2t-2)$ in f_{v} in (S.10) to be consistent with Algorithm 2. However, in order to be fair when comparing the confidence radii, we assume there are two identical risky arms for the two algorithms. Thus both algorithm will pull the two arms approximately the same number of times. Hence, the denominator in the definition of f_{v} is roughly $T_i(t)$.

for any $x_1, \ldots, x_{i-1}, x_i, x_i', x_{i+1}, \ldots, x_n \in [0,1]$. Applying McDiarmid's inequality [41] to f, we get

$$\begin{aligned} &\mathbb{P}[\hat{\sigma}^2(n) - \sigma^2 \geq \varepsilon] \leq \exp(-2n\varepsilon^2), \\ &\mathbb{P}[\hat{\sigma}^2(n) - \sigma^2 \leq -\varepsilon] \leq \exp(-2n\varepsilon^2). \end{aligned} \tag{S.13}$$

Apply (S.13) to the arms, we obtain (S.12). □

Lemma 7: Conditioned on E, if $|\bar{\mathcal{F}}_t| > 1$ and Algorithm 1 has not terminated, then $c_t \in \mathcal{P}_t \cup \{i_t^\star\}$.

Proof: There are two trivial scenarios. Firstly, when $\mathcal{F}_t = \emptyset$, $\mathcal{P}_t = [N]$. Secondly, $\mathcal{F}_t \neq \emptyset$ and $c_t = i_t^\star$. The result is straightforward.

Consider the case where $\mathcal{F}_t \neq \emptyset$ and $c_t \neq i_t^\star$, when $|\bar{\mathcal{F}}_t| > 1$ and the algorithm has not terminated, we must have $c_t \in \mathcal{P}_t$, i.e.,

$$U_{c_t}^\mu(t) \geq L_{i_t^\star}^\mu(t)$$

Otherwise, conditioned on the event E,

$$\begin{aligned} \hat{\mu}_i(t) &< U_i^\mu(t) \leq U_{c_t}^\mu(t) < L_{i_t^\star}^\mu(t) \\ &< \hat{\mu}_{i_t^\star}(t) \leq \hat{\mu}_{i_t}(t) < U_{i_t}^\mu(t), \quad \forall i \in \bar{\mathcal{F}}_t \setminus \{i_t\}. \end{aligned}$$

If $i_t \neq i_t^\star$, we have $U_{c_t}^\mu(t) < L_{i_t^\star}^\mu(t) < \hat{\mu}_{i_t^\star}(t) \leq U_{i_t^\star}^\mu(t)$ which contradicts the definition of c_t. Thus $i_t = i_t^\star$ must hold. In this case, $U_i^\mu(t) \leq U_{c_t}^\mu(t) < L_{i_t^\star}^\mu(t)$ for all $i \in \bar{\mathcal{F}}_t \backslash \{i_t^\star\}$, i.e., $\bar{\mathcal{F}}_t \cap \mathcal{P}_t = \emptyset$. This contradicts the assumption that the algorithm does not terminate. Therefore, $c_t \in \mathcal{P}_t$. □

Lemma 8: The function $h : \mathbb{R} \to \mathbb{R}$ defined by

$$h(x) = \exp\left(-2(\Delta_i^{\mathrm{v}})^2\left(\sqrt{x} - \sqrt{v_i(t)}\right)^2\right)$$

is convex and decreasing on $(4v_i(t), \infty)$ for all $i \in [N]$. Furthermore,

$$\int_{4v_i(t)}^{\infty} h(x)\,\mathrm{d}x \leq \frac{\delta}{2(\Delta_i^{\mathrm{v}})^2 N t^4}.$$

Proof: For simplicity, fix any $i \in [N]$ and let $a = (\Delta_i^{\mathrm{v}})^2, b = \sqrt{v_i(t)}$, then $h(x) = \exp(-2a(\sqrt{x} - b)^2)$. By simple algebra, when $x > 4b^2$

$$\begin{aligned} h'(x) &= \frac{-2a(\sqrt{x} - b)}{\sqrt{x}} \exp(-2a(\sqrt{x} - b)^2) < 0, \\ h''(x) &= \frac{4a^2\sqrt{x}(\sqrt{x} - b)^2 - ab}{x^{\frac{3}{2}}} \exp(-2a(\sqrt{x} - b)^2). \end{aligned}$$

and

$$\begin{aligned} &4a^2\sqrt{x}(\sqrt{x} - b)^2 - ab \geq 8a^2b^3 - ab \\ &\quad = ab(8ab^2 - 1) = ab(8(\Delta_i^{\mathrm{v}})^2 v_i(t) - 1) \\ &\quad \geq ab\left(4\ln\left(\frac{2Nt^4}{\delta}\right) - 1\right) > 0 \end{aligned}$$

Thus $h''(x) > 0$ on $(4v_i(t), \infty)$. The integral can be estimated as

$$\begin{aligned} &\int_{4v_i(t)}^{\infty} h(x)\,\mathrm{d}x \\ &= \int_{4b^2}^{\infty} \exp(-2a(\sqrt{x} - b)^2)\,\mathrm{d}x \\ &= \int_b^{\infty} 2y\exp(-2ay^2)\,\mathrm{d}y + \int_b^{\infty} 2b\exp(-2ay^2)\,\mathrm{d}y \\ &= \frac{1}{2a}\exp(-2ab^2) + 2b\int_{b^2}^{\infty} \frac{1}{2\sqrt{z}}\exp(-2az)\,\mathrm{d}z \\ &\leq \frac{1}{2a}\exp(-2ab^2) + \int_{b^2}^{\infty} \exp(-2az)\,\mathrm{d}z \\ &= \frac{1}{a}\exp(-2ab^2) \leq \frac{\delta}{2(\Delta_i^{\mathrm{v}})^2 N t^4}, \end{aligned}$$

as desired. □

Proof of Lemma 1: Note that our choice of α and β in (3) satisfies

$$\sum_{t=2}^{\infty}\sum_{T=1}^{t-1} \exp\left(-2T\theta(t,T)^2\right) \leq C_\theta \frac{\delta}{N}, \quad \theta = \alpha, \beta \tag{S.14}$$

where $C_\alpha = C_\beta = 1/8$. Lemma 6, the definition of E in (18), and the properties of α, β in (S.14) imply a lower bound on the probability of E

$$\mathbb{P}[E] \geq 1 - 2(C_\alpha + C_\beta)\delta = 1 - \frac{\delta}{2}.$$

This completes the proof. □

Proof of Lemma 2: Conditioned on $E(\tau)$, where τ denotes the stopping time step, i.e., $\bar{\mathcal{F}}_\tau \cap \mathcal{P}_\tau = \emptyset$.

When the input instance is infeasible but $i_\tau \in \mathcal{F}_\tau$ is returned, then $U_{i_\tau}^{\mathrm{v}}(\tau) \leq \bar{\sigma}^2 < \sigma_{i_\tau}^2$ must hold at time step τ, which contradicts the event $E_{i_\tau}(\tau)$. Therefore, $\hat{\mathsf{f}} = \mathsf{f} = 0$.

When the input instance is feasible,

(i) If the instance is deemed to be infeasible, there exists arm $i \in \mathcal{F}$ such that $i \in \bar{\mathcal{F}}_\tau^c$, which violates $E_i(\tau)$.

So $\bar{\mathcal{F}}_\tau \neq \emptyset$. To ease the proof of the rest cases, we prove $i_\tau = i_\tau^\star$. If $\mathcal{F}_\tau = \emptyset$ and $i_\tau \in \partial\mathcal{F}_\tau$, according to the definition of $\mathcal{P}_t$ in (8), $\mathcal{P}_\tau = [N]$. Thus $\bar{\mathcal{F}}_\tau \cap \mathcal{P}_\tau = \bar{\mathcal{F}}_\tau \neq \emptyset$, which contradicts the stopping criterion. Therefore, $\mathcal{F}_\tau \neq \emptyset$ and $i_\tau^\star$ exists. By the definition of i_τ, we have $U_{i_\tau}^\mu(\tau) > \hat{\mu}_{i_\tau}(\tau) \geq \hat{\mu}_{i_\tau^\star}(\tau) > L_{i_\tau^\star}^\mu(\tau)$. This indicates $i_\tau = i_\tau^\star$ or $i_\tau \in \mathcal{P}_\tau$. If $i_\tau \in \mathcal{P}_\tau$, we have $\bar{\mathcal{F}}_\tau \cap \mathcal{P}_\tau \supset \{i_\tau\} \neq \emptyset$, which contradicts the stopping criterion. Therefore, $i_\tau = i_\tau^\star \in \mathcal{F}_\tau$.

(ii) If the instance is evaluated as feasible but the returned arm $i_\tau \in \mathcal{F}_\tau$ is a truly infeasible arm, then it must violate $E_{i_\tau}(\tau)$.

(iii) If the instance is evaluated as feasible but the returned arm i_τ is a truly feasible arm but not $i^\star$. Conditioned on $E_{i^\star}(\tau)$, the arm $i^\star$ belongs to $\bar{\mathcal{F}}_\tau$. Thus the stopping criterion yields $L_{i_\tau}^\mu(\tau) > U_{i^\star}^\mu(\tau)$. Together with $\mu_{i_\tau} < \mu_{i^\star}$, we have $L_{i_\tau}^\mu(\tau) > \mu_{i_\tau}$ or $U_{i^\star}^\mu(\tau) < \mu_{i^\star}$. This violates either $E_{i_\tau}(\tau)$ or $E_{i^\star}(\tau)$.

Hence, $i_{\mathrm{out}} = i_\tau = i_\tau^\star = i^\star, \hat{\mathsf{f}} = \mathsf{f} = 1$. □

Proof of Lemma 3: According to the termination condition, if the algorithm does not terminate, then $\bar{\mathcal{F}}_t \cap \mathcal{P}_t \neq \emptyset$.

To commence our discussion, we state an obvious case before proceeding. Note that if i_t, c_t and $i^\star$ exist with $i_t, c_t \in \mathcal{S}_t$, conditioned on $E(t)$, we have $i^\star \in \bar{\mathcal{F}}_t$ and

$$\bar{\mu} > \max\{U_{i_t}^\mu(t), U_{c_t}^\mu(t)\} \geq U_{i^\star}^\mu(t) > \mu_{i^\star} > \bar{\mu} \tag{S.15}$$

which constitutes a contradiction. So we cannot have i_t and c_t belong to $\mathcal{S}_t$ at the same time. The following discussions will heavily depend on $E(t)$, which guarantees $\mathcal{F}_t \subset \mathcal{F}$ and $\bar{\mathcal{F}}_t^c \subset \bar{\mathcal{F}}^c$.

Case One: Only one arm is sampled.

1) $|\bar{\mathcal{F}}_t| = 1$: in this case only arm $i_t = \text{argmax}\{\hat{\mu}_i(t)) : i \in \bar{\mathcal{F}}_t\}$ is sampled while the rest of the arms are in $\bar{\mathcal{F}}_t^c \subset \bar{\mathcal{F}}^c$.
   a) If $i_t \in \mathcal{F}_t$, by the definition of $i_t^\star$, $i_t = i_t^\star$. Therefore we have $i_t \notin \mathcal{P}_t$ and $\bar{\mathcal{F}}_t = \{i_t\}$, leading to $\bar{\mathcal{F}}_t \cap \mathcal{P}_t = \emptyset$. This contradicts the assumption that the algorithm does not terminate.
   b) If $i_t \in \partial\mathcal{F}_t$, since $|\bar{\mathcal{F}}_t| = 1$ and $i_t^\star$ does not exist, there are two cases:
      i) $i_t \in \mathcal{F}$, i.e., i_t is a truly feasible arm and $i_t = i^\star$. According to (17),
      $$\bar{\mu} < \mu_{i_t} < U_{i_t}^\mu(t)$$
      which indicates $i_t \notin \mathcal{S}_t$. Thus $i_t \in \partial\mathcal{F}_t \backslash \mathcal{S}_t$.
      ii) $i_t \in \bar{\mathcal{F}}^c$, i.e., i_t is a truly infeasible arm and the instance infeasible. By the definition of $\bar{\mu}$, $\bar{\mu} = -\infty$ which makes $\mathcal{S}_t = \emptyset$. So $i_t \in \partial\mathcal{F}_t = \partial\mathcal{F}_t \backslash \mathcal{S}_t$.
2) $|\bar{\mathcal{F}}_t| > 1$: in this case $U_{c_t}^\mu < L_{i_t}^\mu$ and only i_t is sampled.
   a) When $i_t \in \mathcal{F}_t$, we have $\mathcal{F}_t \neq \emptyset$ and $i_t^\star$ exists. We assert that $i_t = i_t^\star$. Assume that i_t and $i_t^\star$ are two different arms, note $i_t = \text{argmax}\{\hat{\mu}_i(t) : i \in \bar{\mathcal{F}}_t\}$, so $\hat{\mu}_{i_t}(t) \geq \hat{\mu}_{i_t^\star}(t)$. And $i_t \in \mathcal{F}_t$, $i_t^\star = \text{argmax}\{\hat{\mu}_i(t) : i \in \mathcal{F}_t\}$, so $\hat{\mu}_{i_t}(t) \leq \hat{\mu}_{i_t^\star}(t)$. We have
   $$\hat{\mu}_{i_t}(t) = \hat{\mu}_{i_t^\star}(t) \tag{S.16}$$
   By the definition of c_t,
   $$\begin{aligned} &U_{c_t}^\mu(t) \geq U_{i_t^\star}^\mu(t) > \hat{\mu}_{i_t^\star}(t) = \hat{\mu}_{i_t}(t) > L_{i_t}^\mu(t) \\ &\Rightarrow U_{c_t}^\mu(t) > L_{i_t}^\mu(t) \end{aligned}$$
   which contradicts $U_{c_t}^\mu < L_{i_t}^\mu$. Hence, we must have $i_t = i_t^\star$. In this case, $U_{c_t}^\mu < L_{i_t}^\mu$ would indicate
   $$U_i^\mu \leq U_{c_t}^\mu < L_{i_t}^\mu(t) = L_{i_t^\star}^\mu(t), \quad \forall\, i \in \bar{\mathcal{F}}_t \backslash \{i_t\}$$
   Thus $\mathcal{P}_t \cap \bar{\mathcal{F}}_t = \emptyset$. This contradicts the assumption that the algorithm does not terminate.
   b) When $i_t \in \partial\mathcal{F}_t$, we assert that $i_t \notin \mathcal{S}_t$. If $i_t \in \mathcal{S}_t$, then
   $$U_{c_t}^\mu < L_{i_t}^\mu < U_{i_t}^\mu < \bar{\mu}$$
   which indicates $c_t \in \mathcal{S}_t$. This contradicts (S.15). Thus $i_t \in \partial\mathcal{F}_t \backslash \mathcal{S}_t$.

   We conclude that when only arm i_t is pulled, we have $i_t \in \partial\mathcal{F}_t \backslash \mathcal{S}_t$.

Case Two: Both i_t and c_t are sampled, i.e., $|\bar{\mathcal{F}}_t| > 1$ and $U_{c_t}^\mu \geq L_{i_t}^\mu$.

1) If $\mathcal{F}_t = \emptyset$, $\partial\mathcal{F}_t$ cannot be empty, otherwise the algorithm terminates. According to (S.15), at least one of i_t or c_t locates in $\partial\mathcal{F}_t \backslash \mathcal{S}_t$.
2) If $\mathcal{F}_t \neq \emptyset$, by Lemma 7, we have $c_t \in \mathcal{P}_t \cup \{i_t^\star\}$.
   a) $i_t \in \partial\mathcal{F}_t$: when $i_t \notin \mathcal{S}_t$, thus $i_t \in \partial\mathcal{F}_t \backslash \mathcal{S}_t$. When $i_t \in \mathcal{S}_t$, according to (S.15), $c_t \notin \mathcal{S}_t$:
      i) if $c_t \in \partial\mathcal{F}_t$, we have $c_t \in \partial\mathcal{F}_t \backslash \mathcal{S}_t$.
      ii) if $c_t \in \mathcal{F}_t$, note that
      $$\begin{aligned} &U_{c_t}^\mu(t) \geq \bar{\mu} > U_{i_t}^\mu(t) > \hat{\mu}_{i_t}(t) \geq \hat{\mu}_{c_t}(t) > L_{c_t}^\mu(t) \\ &\Rightarrow U_{c_t}^\mu(t) \geq \bar{\mu} > L_{c_t}^\mu(t) \end{aligned}$$
      So $c_t \in \mathcal{N}_t$. This gives $c_t \in \mathcal{F}_t \cap \mathcal{N}_t$.
   b) $i_t \in \mathcal{F}_t$: by the same reasoning as (S.16), we obtain $\hat{\mu}_{i_t}(t) = \hat{\mu}_{i_t^\star}(t)$. Firstly, if i_t and $i_t^\star$ are two different arms, we have $U_{c_t}^\mu(t) \geq U_{i_t^\star}^\mu(t) > \hat{\mu}_{i_t^\star}(t) = \hat{\mu}_{i_t}(t) > L_{i_t}^\mu(t)$. Secondly, if i_t and $i_t^\star$ are the same arm, since $c_t \in \mathcal{P}_t \cup \{i_t^\star\}$, we have $U_{c_t}^\mu(t) > L_{i_t}^\mu(t)$. In either case,
   $$U_{c_t}^\mu(t) > L_{i_t}^\mu(t) \tag{S.17}$$
   holds. Note $i_t \in \mathcal{F}$, conditioned on $E(t)$.
      i) When $c_t \in \mathcal{F}_t \subset \mathcal{F}$, if $i_t, c_t \notin \mathcal{N}_t$, we have the following:
         - $i_t \in \mathcal{R}_t, c_t \in \mathcal{R}_t$: since we assumed the optimal arm is unique, then at least one of the two arms has expectation smaller than $\bar{\mu}$. Denote this arm by j. We have $\bar{\mu} < L_j^\mu(t) < \mu_j < \bar{\mu}$, which is a contradiction.
         - $i_t \in \mathcal{S}_t, c_t \in \mathcal{S}_t$: this contradicts (S.15).
         - $i_t \in \mathcal{R}_t, c_t \in \mathcal{S}_t$: we have $U_{c_t}^\mu(t) < \bar{\mu} < L_{i_t}^\mu(t)$. This contradicts (S.17).
         - $i_t \in \mathcal{S}_t, c_t \in \mathcal{R}_t$: we have $\hat{\mu}_{i_t}(t) < U_{i_t}^\mu(t) < \bar{\mu} < L_{c_t}^\mu(t) < \hat{\mu}_{c_t}(t)$. This contradicts the definition of i_t.

      So at least one of i_t and c_t lies in $\mathcal{N}_t$. This gives $i_t \in \mathcal{F}_t \cap \mathcal{N}_t$ or $c_t \in \mathcal{F}_t \cap \mathcal{N}_t$.
      ii) When $c_t \in \partial\mathcal{F}_t$, if $c_t \notin \mathcal{S}_t$, this gives $c_t \in \partial\mathcal{F}_t \backslash \mathcal{S}_t$. If $c_t \in \mathcal{S}_t$, according to (S.15), $i_t \notin \mathcal{S}_t$.
         - if $i_t \in \mathcal{R}_t$, we have $U_{c_t}^\mu(t) < \bar{\mu} < L_{i_t}^\mu(t)$. This contradicts (S.17).
         - if $i_t \in \mathcal{N}_t$, then $i_t \in \mathcal{F}_t \cap \mathcal{N}_t$.

In conclusion, we have

$$i_t \in (\partial\mathcal{F}_t \backslash \mathcal{S}_t) \cup (\mathcal{F}_t \cap \mathcal{N}_t) \quad \text{or} \quad c_t \in (\partial\mathcal{F}_t \backslash \mathcal{S}_t) \cup (\mathcal{F}_t \cap \mathcal{N}_t)$$

as desired. □

Proof of Lemma 4: The techniques used in the analysis of LUCB [9] are adapted in this proof. For a suboptimal arm i, note that $\frac{\Delta_i}{2} \leq \bar{\mu} - \mu_i \leq \Delta_i$.

$$\begin{aligned} &\mathbb{P}[T_i(t) > 16u_i(t), i \in \mathcal{N}_t] \\ &\leq \mathbb{P}[T_i(t) > 16u_i(t), i \notin \mathcal{S}_t] \\ &= \mathbb{P}[T_i(t) > 16u_i(t), U_i^\mu(t) \geq \bar{\mu}] \\ &\leq \sum_{T=16u_i(t)+1}^{\infty} \mathbb{P}[T_i(t) = T, \hat{\mu}_i(t) - \mu_i \geq \bar{\mu} - \alpha(t,T) - \mu_i] \\ &\leq \sum_{T=16u_i(t)+1}^{\infty} \exp\left(-2T\left(\bar{\mu} - \alpha(t,T) - \mu_i\right)^2\right) \\ &\leq \sum_{T=16u_i(t)+1}^{\infty} \exp\left(-2T\left(\frac{\Delta_i}{2} - \sqrt{\frac{1}{2T}\ln\left(\frac{2Nt^4}{\delta}\right)}\right)^2\right) \end{aligned}$$

$$
\begin{aligned}
&\leq \sum_{T=16u_i(t)+1}^{\infty} \exp\left(-2\Delta_i^2\left(\frac{\sqrt{T}}{2}-\sqrt{u_i(t)}\right)^2\right)\\
&\leq \int_{16u_i(t)}^{\infty} \exp\left(-2\Delta_i^2\left(\frac{\sqrt{x}}{2}-\sqrt{u_i(t)}\right)^2\right)\mathrm{d}x\\
&= \int_{4u_i(t)}^{\infty} 4\exp\left(-2\Delta_i^2\left(\sqrt{x}-\sqrt{u_i(t)}\right)^2\right)\mathrm{d}x\\
&\leq \frac{\delta}{2(\frac{\Delta_i}{2})^2 N t^4}
\end{aligned}
$$

The third inequality results from Hoeffding's inequality. The summation can be upper bounded by the integral is due to the fact that the integrand is convex and decreasing within the range of integration, which can be derived by using similar techniques as in Lemma 8. Similarly, for $i = i^\star$, we have

$$
\begin{aligned}
\mathbb{P}[T_{i^\star}(t) > 16u_{i^\star}(t), i^\star \in \mathcal{N}_t] &\leq \mathbb{P}[T_{i^\star}(t) > 16u_{i^\star}(t), i^\star \notin \mathcal{R}_t]\\
&\leq \frac{\delta}{2(\frac{\Delta_{i^\star}}{2})^2 N t^4}
\end{aligned}
$$

For any arm $i \in \mathcal{F}$

$$
\begin{aligned}
&\mathbb{P}[T_i(t) > 4v_i(t), i \in \partial\mathcal{F}_t]\\
&\leq \mathbb{P}[T_i(t) > 4v_i(t), i \notin \mathcal{F}_t]\\
&= \mathbb{P}[T_i(t) > 4v_i(t), U_i^{\mathrm{v}}(t) > \bar{\sigma}^2]\\
&\leq \sum_{T=4v_i(t)+1}^{\infty} \mathbb{P}[T_i(t) = T, \hat{\sigma}_i^2(t) - \sigma_i^2 > \bar{\sigma}^2 - \sigma_i^2 - \beta(t,T)]\\
&\leq \sum_{T=4v_i(t)+1}^{\infty} \exp\left(-2T\left(\bar{\sigma}^2 - \sigma_i^2 - \beta(t,T)\right)^2\right)\\
&\leq \sum_{T=4v_i(t)+1}^{\infty} \exp\left(-2T\left(\Delta_i^{\mathrm{v}} - \sqrt{\frac{1}{2T}\ln\left(\frac{2Nt^4}{\delta}\right)}\right)^2\right)\\
&\leq \sum_{T=4v_i(t)+1}^{\infty} \exp\left(-2(\Delta_i^{\mathrm{v}})^2\left(\sqrt{T}-\sqrt{v_i(t)}\right)^2\right)\\
&\leq \int_{T=4v_i(t)}^{\infty} \exp\left(-2(\Delta_i^{\mathrm{v}})^2\left(\sqrt{x}-\sqrt{v_i(t)}\right)^2\right)\mathrm{d}x\\
&\leq \frac{\delta}{2(\Delta_i^{\mathrm{v}})^2 N t^4}.
\end{aligned}
$$

The third inequality utilizes McDiarmid's inequality. The last two steps are due to Lemma 8. The same holds for $i \in \bar{\mathcal{F}}^c$:

$$
\begin{aligned}
\mathbb{P}[T_i(t) > 4v_i(t), i \in \partial\mathcal{F}_t] &\leq \mathbb{P}[T_i(t) > 4v_i(t), i \notin \bar{\mathcal{F}}_t^c]\\
&\leq \frac{\delta}{2(\Delta_i^{\mathrm{v}})^2 N t^4}.
\end{aligned}
$$

This completes the proof. □

Proof of Lemma 5: The proof improves the techniques used in the analysis of the original LUCB algorithm [9] in order to analyze the effect of the empirical variances on the sample complexity.

Let t be a sufficiently large integer and $\mathcal{T} := \{\lceil t/2\rceil, \ldots, t-1\}$. Define events

- $\mathcal{I}_1$: $\exists s \in \mathcal{T}$ such that $i^\star \notin \mathcal{F}_s \cap \mathcal{R}_s$ and $T_{i^\star}(s) > \max\{16u_{i^\star}(s), 4v_{i^\star}(s)\}$.
- $\mathcal{I}_2$: $\exists i \in \mathcal{F} \cap \mathcal{S}, s \in \mathcal{T}$ such that $i \notin \mathcal{S}_s$ and $T_i(s) > 16u_i(s)$.
- $\mathcal{I}_3$: $\exists i \in \bar{\mathcal{F}}^c \cap \mathcal{R}, s \in \mathcal{T}$ such that $i \notin \bar{\mathcal{F}}_s^c$ and $T_i(s) > 4v_i(s)$
- $\mathcal{I}_4$: $\exists i \in \bar{\mathcal{F}}^c \cap \mathcal{S}, s \in \mathcal{T}$ such that $i \notin \bar{\mathcal{F}}_s^c \cup \mathcal{S}_s$ and $T_i(s) > \min\{16u_i(s), 4v_i(s)\}$.
- $\mathcal{I}_5$: $\exists s \in \mathcal{T}$ such that $E(s)$ does not occur.

If VA-LUCB terminates before $\lceil t/2\rceil$, the statement is definitely right. If not, we assume the above five events do not occur. Based on Lemma 3, the additional time steps after $\lceil t/2\rceil - 1$ for VA-LUCB can be upper bounded as (S.18) in Derivation 1, shown at the next page. On the other hand, we observe that if the time step $t = CH_{\mathrm{VA}}\ln(H_{\mathrm{VA}}/\delta)$ with $C \geq 152$, we have (S.19) as in Derivation 2, shown at the next page. So the total number of time steps is bounded by

$$
\begin{aligned}
&\left\lceil \frac{t}{2}\right\rceil - 1 + \max\{16u_{i^\star}(t), 4v_{i^\star}(t)\} + \sum_{i\in\mathcal{F}\cap\mathcal{S}} 16u_i(t)\\
&+ \sum_{i\in\bar{\mathcal{F}}^c\cap\mathcal{R}} 4v_i(t) + \sum_{i\in\bar{\mathcal{F}}^c\cap\mathcal{S}} \min\{16u_i(t), 4v_i(t)\} \leq t.
\end{aligned}
$$

We then compute the probability of the event $\bigcup_{i\in[5]} \mathcal{I}_i$. To simplify notations used in the following, we define the hardness quantity

$$
\begin{aligned}
\widetilde{H} :=& \frac{1}{(\Delta_{i^\star}^{\mathrm{v}})^2} + \sum_{i\in\mathcal{F}} \frac{1}{(\frac{\Delta_i}{2})^2}\\
&+ \sum_{i\in\bar{\mathcal{F}}^c\cap\mathcal{R}} \frac{1}{(\Delta_i^{\mathrm{v}})^2} + \sum_{i\in\bar{\mathcal{F}}^c\cap\mathcal{S}} \frac{1}{f(u_i(t), v_i(t))},
\end{aligned}
$$

where $f(u_i(t), v_i(t)) = (\frac{\Delta_i}{2})^2 \cdot \mathbb{1}\{16u_i(t) < 4v_i(t)\} + (\Delta_i^{\mathrm{v}})^2 \cdot \mathbb{1}\{16u_i(t) \geq 4v_i(t)\}$ which is exactly $\max\{\frac{\Delta_i}{2}, \Delta_i^{\mathrm{v}}\}^2$ if we ignore the ceiling operators in $u_i(t)$ and $v_i(t)$. By Lemma 6 and Lemma 4,

$$
\begin{aligned}
\mathbb{P}[\mathcal{I}_1] &\leq \sum_{s\in\mathcal{T}}\left(\frac{\delta}{2(\frac{\Delta_{i^\star}}{2})^2 N s^4} + \frac{\delta}{2(\Delta_{i^\star}^{\mathrm{v}})^2 N s^4}\right),\\
\mathbb{P}[\mathcal{I}_2] &\leq \sum_{i\in\mathcal{F}\cap\mathcal{S}}\sum_{s\in\mathcal{T}} \frac{\delta}{2(\frac{\Delta_i}{2})^2 N s^4},\\
\mathbb{P}[\mathcal{I}_3] &\leq \sum_{i\in\bar{\mathcal{F}}^c\cap\mathcal{R}}\sum_{s\in\mathcal{T}} \frac{\delta}{2(\Delta_i^{\mathrm{v}})^2 N s^4},\\
\mathbb{P}[\mathcal{I}_4] &\leq \sum_{i\in\bar{\mathcal{F}}^c\cap\mathcal{S}}\sum_{s\in\mathcal{T}} \frac{\delta}{2f(u_i(t), v_i(t)) N s^4},\\
\mathbb{P}[\mathcal{I}_5] &\leq \sum_{s\in\mathcal{T}} \frac{2\delta}{s^3} \leq \frac{3\delta}{t^2},
\end{aligned}
$$

which implies that

$$
\mathbb{P}\left[\bigcup_{i\in[5]} \mathcal{I}_i\right] \leq \widetilde{H}\sum_{s\in\mathcal{T}} \frac{\delta}{2Ns^4} + \frac{3\delta}{t^2} \leq \frac{5\delta}{t^2},
$$

where the last inequality utilizes the fact that $\widetilde{H} < 2H_{\mathrm{VA}} < t$ and $N \geq 2$. This yields the upper bound of the probability that the algorithm does not terminate at time step t. □

Proof of Theorem 1: By Lemmas 1 and 2, if it terminates, Algorithm 1 succeeds on event E, which occurs with probability at least $1 - \delta/2$. According to Lemma 5, Algorithm 1

Derivation 1:

$$
\begin{aligned}
&\sum_{s\in\mathcal{T}} \mathbb{1}\{i_s \in (\mathcal{F}_s \cap \mathcal{N}_s) \cup (\partial\mathcal{F}_s \backslash \mathcal{S}_s) \text{ or } c_s \in (\mathcal{F}_s \cap \mathcal{N}_s) \cup (\partial\mathcal{F}_s \backslash \mathcal{S}_s)\} \\
&\le \sum_{s\in\mathcal{T}} \sum_{i\in[N]} \mathbb{1}\{i = i_s \text{ or } c_s,\, i \in (\mathcal{F}_s \cap \mathcal{N}_s) \cup (\partial\mathcal{F}_s \backslash \mathcal{S}_s)\} \\
&= \sum_{s\in\mathcal{T}} \Bigg(\mathbb{1}\{i^\star = i_s \text{ or } c_s,\, i^\star \in (\mathcal{F}_s \cap \mathcal{N}_s) \cup (\partial\mathcal{F}_s \backslash \mathcal{S}_s)\} + \sum_{i\in\mathcal{F}\cap\mathcal{S}} \mathbb{1}\{i = i_s \text{ or } c_s,\, i \in (\mathcal{F}_s \cap \mathcal{N}_s) \cup (\partial\mathcal{F}_s \backslash \mathcal{S}_s)\} \\
&\quad + \sum_{i\in\bar{\mathcal{F}}^c\cap\mathcal{R}} \mathbb{1}\{i = i_s \text{ or } c_s,\, i \in (\mathcal{F}_s \cap \mathcal{N}_s) \cup (\partial\mathcal{F}_s \backslash \mathcal{S}_s)\} + \sum_{i\in\bar{\mathcal{F}}^c\cap\mathcal{S}} \mathbb{1}\{i = i_s \text{ or } c_s,\, i \in (\mathcal{F}_s \cap \mathcal{N}_s) \cup (\partial\mathcal{F}_s \backslash \mathcal{S}_s)\} \Bigg) \\
&\le \sum_{s\in\mathcal{T}} \Bigg(\mathbb{1}\{i^\star = i_s \text{ or } c_s,\, i^\star \notin \mathcal{F}_s \cap \mathcal{R}_s\} + \sum_{i\in\mathcal{F}\cap\mathcal{S}} \mathbb{1}\{i = i_s \text{ or } c_s,\, i \notin \mathcal{S}_s\} \\
&\quad + \sum_{i\in\bar{\mathcal{F}}^c\cap\mathcal{R}} \mathbb{1}\{i = i_s \text{ or } c_s,\, i \notin \bar{\mathcal{F}}_s^c\} + \sum_{i\in\bar{\mathcal{F}}^c\cap\mathcal{S}} \mathbb{1}\{i = i_s \text{ or } c_s,\, i \notin \bar{\mathcal{F}}_s^c \cup \mathcal{S}_s\} \Bigg) \\
&\le \sum_{s\in\mathcal{T}} \Bigg(\mathbb{1}\{i^\star = i_s \text{ or } c_s,\, T_{i^\star}(s) \le \max\{16u_{i^\star}(s), 4v_{i^\star}(s)\}\} + \sum_{i\in\mathcal{F}\cap\mathcal{S}} \mathbb{1}\{i = i_s \text{ or } c_s,\, T_i(s) \le 16u_i(s)\} \\
&\quad + \sum_{i\in\bar{\mathcal{F}}^c\cap\mathcal{R}} \mathbb{1}\{i = i_s \text{ or } c_s,\, T_i(s) \le 4v_i(s)\} + \sum_{i\in\bar{\mathcal{F}}^c\cap\mathcal{S}} \mathbb{1}\{i = i_s \text{ or } c_s,\, T_i(s) \le \min\{16u_i(s), 4v_i(s)\}\} \Bigg) \\
&\le \sum_{s\in\mathcal{T}} \mathbb{1}\{i^\star = i_s \text{ or } c_s,\, T_{i^\star}(s) \le \max\{16u_{i^\star}(s), 4v_{i^\star}(s)\}\} + \sum_{i\in\mathcal{F}\cap\mathcal{S}} \sum_{s\in\mathcal{T}} \mathbb{1}\{i = i_s \text{ or } c_s,\, T_i(s) \le 16u_i(s)\} \\
&\quad + \sum_{i\in\bar{\mathcal{F}}^c\cap\mathcal{R}} \sum_{s\in\mathcal{T}} \mathbb{1}\{i = i_s \text{ or } c_s,\, T_i(s) \le 4v_i(s)\} + \sum_{i\in\bar{\mathcal{F}}^c\cap\mathcal{S}} \sum_{s\in\mathcal{T}} \mathbb{1}\{i = i_s \text{ or } c_s,\, T_i(s) \le \min\{16u_i(s), 4v_i(s)\}\} \\
&\le \max\{16u_{i^\star}(t), 4v_{i^\star}(t)\} + \sum_{i\in\mathcal{F}\cap\mathcal{S}} 16u_i(t) + \sum_{i\in\bar{\mathcal{F}}^c\cap\mathcal{R}} 4v_i(t) + \sum_{i\in\bar{\mathcal{F}}^c\cap\mathcal{S}} \min\{16u_i(t), 4v_i(t)\}. \qquad \text{(S.18)}
\end{aligned}
$$

Derivation 2:

$$
\begin{aligned}
&\max\{16u_{i^\star}(t), 4v_{i^\star}(t)\} + \sum_{i\in\mathcal{F}\cap\mathcal{S}} 16u_i(t) + \sum_{i\in\bar{\mathcal{F}}^c\cap\mathcal{R}} 4v_i(t) + \sum_{i\in\bar{\mathcal{F}}^c\cap\mathcal{S}} \min\{16u_i(t), 4v_i(t)\} \\
&= \max\left\{16\left\lceil\frac{1}{2\Delta_{i^\star}^2}\ln\left(\frac{2Nt^4}{\delta}\right)\right\rceil, 4\left\lceil\frac{1}{2(\Delta_{i^\star}^{\mathrm{v}})^2}\ln\left(\frac{2Nt^4}{\delta}\right)\right\rceil\right\} + 16\sum_{i\in\mathcal{F}\cap\mathcal{S}}\left\lceil\frac{1}{2\Delta_i^2}\ln\left(\frac{2Nt^4}{\delta}\right)\right\rceil \\
&\quad + 4\sum_{i\in\bar{\mathcal{F}}^c\cap\mathcal{R}}\left\lceil\frac{1}{2(\Delta_i^{\mathrm{v}})^2}\ln\left(\frac{2Nt^4}{\delta}\right)\right\rceil + \sum_{i\in\bar{\mathcal{F}}^c\cap\mathcal{S}} \min\left\{16\left\lceil\frac{1}{2\Delta_i^2}\ln\left(\frac{2Nt^4}{\delta}\right)\right\rceil, 4\left\lceil\frac{1}{2(\Delta_i^{\mathrm{v}})^2}\ln\left(\frac{2Nt^4}{\delta}\right)\right\rceil\right\} \\
&\le 16N + \max\left\{16\frac{1}{2\Delta_{i^\star}^2}\ln\left(\frac{2Nt^4}{\delta}\right), 4\frac{1}{2(\Delta_{i^\star}^{\mathrm{v}})^2}\ln\left(\frac{2Nt^4}{\delta}\right)\right\} + 16\sum_{i\in\mathcal{F}\cap\mathcal{S}}\frac{1}{2\Delta_i^2}\ln\left(\frac{2Nt^4}{\delta}\right) \\
&\quad + 4\sum_{i\in\bar{\mathcal{F}}^c\cap\mathcal{R}}\frac{1}{2(\Delta_i^{\mathrm{v}})^2}\ln\left(\frac{2Nt^4}{\delta}\right) + \sum_{i\in\bar{\mathcal{F}}^c\cap\mathcal{S}} \min\left\{16\frac{1}{2\Delta_i^2}\ln\left(\frac{2Nt^4}{\delta}\right), 4\frac{1}{2(\Delta_i^{\mathrm{v}})^2}\ln\left(\frac{2Nt^4}{\delta}\right)\right\} \\
&= 16N + 2H_{\mathrm{VA}}\ln\left(\frac{2Nt^4}{\delta}\right) \\
&= (16N + 2H_{\mathrm{VA}}\ln 2) + 2H_{\mathrm{VA}}\ln\frac{N}{\delta} + 8H_{\mathrm{VA}}\ln\left(CH_{\mathrm{VA}}\ln\frac{H_{\mathrm{VA}}}{\delta}\right) \\
&\le (16N + 2H_{\mathrm{VA}}\ln 2 + 8H_{\mathrm{VA}}\ln C) + 2H_{\mathrm{VA}}\ln\frac{N}{\delta} + 16H_{\mathrm{VA}}\ln\frac{H_{\mathrm{VA}}}{\delta} \\
&\le \frac{1}{2}CH_{\mathrm{VA}}\ln\frac{H_{\mathrm{VA}}}{\delta}. \qquad \text{(S.19)}
\end{aligned}
$$

terminates at time $t > t^\star > 152 > 5$ with probability at least $5\delta/t^2$. So Algorithm 1 succeeds after $O(H_{\mathrm{VA}} \ln(H_{\mathrm{VA}}/\delta))$ time steps with probability at least $1-(\delta/2+5\delta/5^2) \geq 1-\delta$.[4] Note that the sample complexity is at most twice of number of time steps. This completes the proof of Theorem 1. □

APPENDIX E
THE SUB-GAUSSIAN CASE

In this section, we extend the utility and analysis of VA-LUCB to the case in which the rewards are sub-Gaussian. We see that the main difficulty lies in the fact that the empirical variance is sub-Exponential and its concentration bound (see Lemma 9) is not as convenient as that for the bounded rewards case (in Lemma 6). Thus the main change of VA-LUCB is the inclusion of a warm-up phase in which we pull each arm a fixed number of times and a forced-sampling procedure in the following time steps. We specify precisely in the following how many we need to pull each arm in the initial forced exploration phase and in the forced-sampling procedure.

Recall (see, for example, Duchi [42, Chapter 3]) that a random variable X is *σ-sub-Gaussian* (or *sub-Gaussian* with variance proxy σ^2) if for all $s \in \mathbb{R}$,

$$\ln \mathbb{E}[\exp(s(X - \mathbb{E}X))] \leq \frac{s^2\sigma^2}{2}.$$

Additionally, Y is *sub-Exponential* with parameters (τ^2, b) (also written as $Y \sim \mathrm{SE}(\tau^2, b)$) if for all $s \in \mathbb{R}$ such that $|s| \leq 1/b$,

$$\ln \mathbb{E}[\exp(s(Y - \mathbb{E}Y))] \leq \frac{s^2\tau^2}{2}.$$

For brevity, let c denote the absolute constant 64 from now on. Given an instance $(\nu, \bar{\sigma}^2)$, where $\nu_i, i \in [N]$ are independent σ-sub-Gaussian distributions, we define the *hardness parameter* for this σ-sub-Gaussian instance as

$$H^{(\sigma)}_{\mathrm{VA},N} := \max\{H^{(\sigma)}_{\mathrm{VA}}, N\}, \quad \text{where}$$
$$H^{(\sigma)}_{\mathrm{VA}} := \max\left\{\frac{2\sigma^2}{(\frac{\Delta_{i^\star}}{2})^2}, \frac{2c\sigma^4}{(\Delta^{\mathrm{v}}_{i^\star})^2}\right\} + \sum_{i\in\mathcal{F}\cap\mathcal{S}} \frac{2\sigma^2}{(\frac{\Delta_i}{2})^2} + \sum_{i\in\bar{\mathcal{F}}^c\cap\mathcal{R}} \frac{2c\sigma^4}{(\Delta^{\mathrm{v}}_i)^2} + \sum_{i\in\bar{\mathcal{F}}^c\cap\mathcal{S}} \min\left\{\frac{2\sigma^2}{(\frac{\Delta_i}{2})^2}, \frac{2c\sigma^4}{(\Delta^{\mathrm{v}}_i)^2}\right\}.$$

Algorithm 3, designed for sub-Gaussian random rewards, is a slight extension of VA-LUCB (Algorithm 1) and has the following guarantee.

Theorem 3 (Upper Bound for σ-sub-Gaussian Case): Given an instance $(\nu, \bar{\sigma}^2)$ with $\bar{\sigma}^2 < \sigma^2$ and confidence parameter $\delta \leq 0.05$, with probability at least $1-\delta$, Algorithm 3 succeeds and terminates in

$$O\left(H^{(\sigma)}_{\mathrm{VA},N} \ln \frac{H^{(\sigma)}_{\mathrm{VA},N}}{\delta}\right) \tag{S.20}$$

time steps. Furthermore, the expected sample complexity is as in (S.20).

[4]The reader may notice that our estimate of t is rather coarse here. When t is large enough, the probability that the algorithm does not stop is negligible.

Algorithm 3 Variance-Aware LUCB for σ-Sub-Gaussian Distributions (VA-LUCB-σ-Sub-Gaussian)

Input: threshold $\bar{\sigma}^2 > 0$, sub-Gaussian parameter σ, and confidence parameter $\delta \in (0,1)$.
Sample each of the N arms T_0 (see (S.25)) times and set $\bar{\mathcal{F}}_{T_0} = [N]$.
for time step $t = T_0+1, T_0+2 \ldots$ **do**
  Compute the sample mean and sample variance using (1) and (2) for $i \in \bar{\mathcal{F}}_{t-1}$.
  Update the confidence bounds for the mean and variance by (4) and (5) for $i \in \bar{\mathcal{F}}_{t-1}$.
  Update $\mathcal{F}_t := \{i : U^{\mathrm{v}}_i(t) \leq \bar{\sigma}^2\}$. and $\bar{\mathcal{F}}_t := \{i : L^{\mathrm{v}}_i(t) \leq \bar{\sigma}^2\}$.
  Find $i^\star_t := \mathrm{argmax}\{\hat{\mu}_i(t) : i \in \mathcal{F}_t\}$ if $\mathcal{F}_t \neq \emptyset$.
  Update $\mathcal{P}_t := \{i : L^{\mu}_{i^\star_t}(t) \leq U^{\mu}_i(t), i \neq i^\star_t\}$ if $\{\mathcal{F}_t \neq \emptyset\}$, otherwise $\mathcal{P}_t := [N]$.
  if $\bar{\mathcal{F}}_t \cap \mathcal{P}_t = \emptyset$ **then**
    if $\mathcal{F}_t \neq \emptyset$ **then**
      Set $i_{\mathrm{out}} = i_t = \mathrm{argmax}\left\{\hat{\mu}_i(t) : i \in \bar{\mathcal{F}}_t\right\}$ and $\hat{\mathsf{f}} = 1$.
    else
      Set $\hat{\mathsf{f}} = 0$.
    end if
    break
  end if
  if $|\bar{\mathcal{F}}_t| = 1$ **then**
    Sample arm $i_t = \mathrm{argmax}\left\{\hat{\mu}_i(t) : i \in \bar{\mathcal{F}}_t\right\}$. (in one round).
  else
    Find $i_t = \mathrm{argmax}\left\{\hat{\mu}_i(t) : i \in \bar{\mathcal{F}}_t\right\}$ and competitor arm $c_t = \mathrm{argmax}\{U^{\mu}_i(t) : i \in \bar{\mathcal{F}}_t, i \neq i_t\}$.
    if $U^{\mu}_{c_t}(t) \geq L^{\mu}_{i_t}(t)$ **then**
      Sample arms i_t and c_t (in two rounds)
    else
      Sample arm i_t (in one round).
    end if
  end if
  Find $\mathcal{M}_t = \{i \in \bar{\mathcal{F}}_t : \beta(t+1, T_i(t)) > \sigma^2, i$ has not been sampled at this time step$\}$.
  Sample each arm in $\mathcal{M}_t$ once (in $|\mathcal{M}_t|$ rounds).
end for

Remark 4: When the threshold $\bar{\sigma}^2 \geq \sigma^2$, all the arms are feasible and the problem reduce to vanilla BAI problem. The more interesting case is the case where $\bar{\sigma}^2 < \sigma^2$ and the expectations of the arms are close, e.g., $\Delta_i \leq 2\sigma$ for all $i \in [N]$. In this case, $\Delta^{\mathrm{v}}_i < \sigma^2, \Delta_i \leq 2\sqrt{2}\sigma$ for all $i \in [N]$, leading to $H^{(\sigma)}_{\mathrm{VA},N} = H^{(\sigma)}_{\mathrm{VA}}$. Furthermore,

$$\min\left\{2\sigma^2, 2c\sigma^4\right\} H_{\mathrm{VA}} \leq H^{(\sigma)}_{\mathrm{VA}} \leq \max\left\{2\sigma^2, 2c\sigma^4\right\} H_{\mathrm{VA}},$$

where H_{VA} is defined in (11). These bounds imply that $H^{(\sigma)}_{\mathrm{VA}}$ essentially captures the intrinsic hardness of the instance and is related linearly to the hardness parameter for the bounded rewards case H_{VA}.

A. VA-LUCB for the σ-Sub-Gaussian Case

We extend VA-LUCB to σ-sub-Gaussian distributions. The modified algorithm based on Algorithm 1 is stated in Algorithm 3. The notations from VA-LUCB (Algorithm 1) can be directly adapted to the σ-sub-Gaussian case, except that two notations need to be modified slightly.

- Let $\mathcal{J}_t$ denote the set of arms pulled in time step t. Note that there can be *more than* 2 arms being sampled in one time step.
- The confidence radii for the mean and variance are re-defined to be

$$\alpha(t,T)=\sqrt{\frac{2\sigma^2}{T}\ln\frac{kNt^4}{\delta}},\quad\text{and}\qquad\text{(S.21)}$$
$$\beta(t,T)=\sqrt{\frac{2c\sigma^4}{T}\ln\frac{kNt^4}{\delta}};$$

  respectively, where $k>0$ is an absolute constant to be determined.

Before the analysis of Algorithm 3, we present a convenient concentration bound for the sample variance.

Lemma 9: For an i.i.d. σ-sub-Gaussian random variables $X_1,\dots,X_n$ with expectation μ and variance $\mathrm{Var}(X)$, let $\bar{X}=\frac{1}{n}\sum_{i=1}^n X_i$ denote the sample mean and $S_n^2=\frac{1}{n-1}\sum_{i=1}^n(X_i-\bar{X})^2$ denote the (unbiased) sample variance. For any integer $n\geq 12$ and $\epsilon>0$, we have

$$\begin{aligned}\mathbb{P}[S_n^2-\mathrm{Var}(X)\geq\epsilon] &\leq \exp\left(-\frac{1}{16}\min\left\{\frac{n\epsilon^2}{8\sigma^4},\frac{n\epsilon}{\sigma^2}\right\}\right),\\ \mathbb{P}[S_n^2-\mathrm{Var}(X)\leq-\epsilon] &\leq \exp\left(-\frac{1}{16}\min\left\{\frac{n\epsilon^2}{8\sigma^4},\frac{n\epsilon}{\sigma^2}\right\}\right).\end{aligned}\qquad\text{(S.22)}$$

Proof: We prove the former inequality here; the latter can be derived analogously. According to Honorio and Jaakkola [43, Appendix B], for any σ-sub-Gaussian random variable X,

$$\mathbb{E}\left[\exp\left(t(X^2-\mathbb{E}[X^2])\right)\right]\leq\exp\left(16t^2\sigma^4\right),\quad\forall\,|t|\leq\frac{1}{4\sigma^2},$$

which indicates that $X^2-\mathbb{E}[X^2]$ is sub-Exponential. More precisely, $X^2-\mathbb{E}[X^2]\sim\mathrm{SE}(32\sigma^4,4\sigma^2)$.

The sample variance can be reorganized as

$$\begin{aligned}S_n^2&=\frac{1}{n-1}\sum_{i=1}^n(X_i-\bar{X})^2=\frac{1}{n-1}\sum_{i=1}^n X_i^2-\frac{n}{n-1}\bar{X}^2\\&=\frac{1}{n-1}\sum_{i=1}^n(X_i-\mu)^2-\frac{n}{n-1}(\bar{X}-\mu)^2.\end{aligned}$$

By the properties of sub-Gaussian and sub-Exponential random variables (see Duchi [42, Chapter 3]),

$$\begin{aligned}&X_i-\mu\sim\sigma\text{-sub-Gaussian}\\\Longrightarrow\quad&\frac{X_i-\mu}{\sqrt{n-1}}\sim\frac{\sigma}{\sqrt{n-1}}\text{-sub-Gaussian}\\\Longrightarrow\quad&\frac{(X_i-\mu)^2}{n-1}\sim\mathrm{SE}\left(32\frac{\sigma^4}{(n-1)^2},4\frac{\sigma^2}{n-1}\right)\\\Longrightarrow\quad&S_n^2\sim\mathrm{SE}\left(32\frac{n\sigma^4}{(n-1)^2},4\frac{\sigma^2}{n-1}\right)\end{aligned}$$

where the last implication utilizes the independence of $(X_i-\mu)$ across $i\in[n]$. Likewise,

$$\begin{aligned}&\sum_{i=1}^n X_i-n\mu\sim\sqrt{n}\sigma\text{-sub-Gaussian}\\\Longrightarrow\quad&\bar{X}-\mu\sim\frac{\sigma}{\sqrt{n}}\text{-sub-Gaussian}\\\Longrightarrow\quad&\sqrt{\frac{n}{n-1}}(\bar{X}-\mu)\sim\frac{\sigma}{\sqrt{n-1}}\text{-sub-Gaussian}\\\Longrightarrow\quad&\frac{n}{n-1}(\bar{X}-\mu)^2\sim\mathrm{SE}\left(32\frac{\sigma^4}{(n-1)^2},4\frac{\sigma^2}{n-1}\right)\end{aligned}$$

Therefore, $S_n^2\sim\mathrm{SE}\left(32\sigma^4(\frac{\sqrt{n}+1}{n-1})^2,\frac{4\sigma^2}{n-1}\right)$. According to the concentration property of the sub-Exponential random variables presented in [42, Corollary 3.17], we have

$$\begin{aligned}&\mathbb{P}[S_n^2-\mathrm{Var}(X)\geq\epsilon]\\&\leq\exp\left(-\frac{1}{2}\min\left\{\frac{\epsilon^2}{32\sigma^4(\frac{\sqrt{n}+1}{n-1})^2},\frac{\epsilon}{\frac{4\sigma^2}{n-1}}\right\}\right).\end{aligned}$$

When $n\geq 12$, we have $(\frac{\sqrt{n}+1}{n-1})^2\leq\frac{2}{n}$ and $\frac{4\sigma^2}{n-1}\leq\frac{8\sigma^2}{n}$. Hence.

$$\mathbb{P}[S_n^2-\mathrm{Var}(X)\geq\epsilon]\leq\exp\left(-\frac{1}{2}\min\left\{\frac{n\epsilon^2}{64\sigma^4},\frac{n\epsilon}{8\sigma^2}\right\}\right)$$

as desired. □

For σ-sub-Gaussian distributions, we have the following concentration inequalities for the mean in corresponds to Lemma 6:

$$\begin{aligned}\mathbb{P}[\hat{\mu}_i(t)-\mu_i\geq\epsilon]&\leq\exp\left(-\frac{T_i(t)\epsilon^2}{2\sigma^2}\right),\\\mathbb{P}[\hat{\mu}_i(t)-\mu_i\leq-\epsilon]&\leq\exp\left(-\frac{T_i(t)\epsilon^2}{2\sigma^2}\right).\end{aligned}\qquad\text{(S.23)}$$

In order to get a tightness result, we force the confidence radius for the variance β to be no greater than $c\sigma^2/8$ through out the algorithm (Line 2 and Lines 27 and 28 of Algorithm 3), i.e., $\beta(t,T_i(t))\leq c\sigma^2/8$ which is equivalent to $T_i(t)\geq\frac{128}{c}\ln\frac{kNt^4}{\delta}$. Thus, (S.22) simplifies to

$$\begin{aligned}\mathbb{P}[\hat{\sigma}_i^2(t)-\sigma_i^2\geq\epsilon]&\leq\exp\left(-\frac{T_i(t)\epsilon^2}{2c\sigma^4}\right),\\\mathbb{P}[\hat{\sigma}_i^2(t)-\sigma_i^2\leq-\epsilon]&\leq\exp\left(-\frac{T_i(t)\epsilon^2}{2c\sigma^4}\right).\end{aligned}\qquad\text{(S.24)}$$

We are now ready to present the intuitions for Algorithm 3. In the warm-up procedure (Line 2),

$$T_0:=\min\left\{t\in\mathbb{N}:t\geq\frac{128}{c}\ln\frac{kNt^4}{\delta}\right\}\qquad\text{(S.25)}$$

and all arms are sampled at each of the T_0 time steps. After the warm-up, $t=T_0+1$ and $\beta(t,T_i(t))\leq\frac{c\sigma^2}{8}$ so that (S.24) holds for all arms. The intuitions for Line 3 to Line 26 are the same as Algorithm 1. The only difference here is Lines 27 and 28. The definition of $\mathcal{M}_t$ guarantees each of the arms will be pulled at most once at each time step.

Lemma 10: When Algorithm 3 has not terminated, for any time step $t > T_0$ and arms $i \in \bar{\mathcal{F}}_{t-1}$,

$$\beta(t, T_i(t)) \leq \frac{c\sigma^2}{8}.$$

Proof: We prove this lemma by induction. When $t = T_0 + 1$, by the choice of T_0, the lemma holds.

Assume that for some $t > T_0$, the lemma holds, i.e. for arm $i \in \bar{\mathcal{F}}_{t-1}$, $\beta(t, T_i(t)) \leq \frac{c\sigma^2}{8}$. Conditioned on event E, $\bar{\mathcal{F}}_t \subset \bar{\mathcal{F}}_{t-1}$. If arm $i \in \bar{\mathcal{F}}_t$ is pulled at time step t, we have $T_i(t+1) = T_i(t) + 1$. Thus

$$\begin{aligned}
&\beta(t+1, T_i(t+1)) \leq \frac{c\sigma^2}{8} \\
\iff &T_i(t+1) \geq \frac{128}{c} \ln \frac{kN(t+1)^4}{\delta} \\
\iff &T_i(t) + 1 \geq \frac{128}{c} \ln \frac{kNt^4}{\delta} + 4\frac{128}{c} \ln \frac{t+1}{t}. \qquad \text{(S.26)}
\end{aligned}$$

We now see that if $1 \geq \frac{512}{ct}$ holds, then (S.26) also holds trivially. Consequently, if arm $i \in \bar{\mathcal{F}}_t$ is not pulled at time step t, we must have $T_i(t+1) = T_i(t)$ and $\beta(t+1, T_i(t+1)) \leq \frac{c\sigma^2}{8}$. Therefore, the lemma holds for $t+1$.

By induction, the lemma holds. □

The above lemma guarantees we can always adopt (S.24) for all arms in $\bar{\mathcal{F}}_t$ after the warm-up procedure.

B. Analysis

Define the events

$$\begin{aligned}
E_i^{\mu}(t) &:= \{|\hat{\mu}_i(t) - \mu_i| \leq \alpha(t, T_i(t))\}, \\
E_i^{\mathrm{v}}(t) &:= \{|\hat{\sigma}_i^2(t) - \sigma_i^2| \leq \beta(t, T_i(t))\}, \\
E_i(t) &:= E_i^{\mu}(t) \bigcap E_i^{\mathrm{v}}(t), \quad \forall\, i \in [N].
\end{aligned}$$

For $t > T_0$, define

$$E(t) := \bigcap_{i \in [N]} E_i(t) \quad \text{and} \quad E := \bigcap_{t > T_0} E(t). \qquad \text{(S.27)}$$

Lemma 11 (Analogue of Lemma 1): Define E as in (S.27) and $\alpha(t, T)$ and $\beta(t, T)$ as in (S.21), then E occurs with probability at least $1 - \delta/k$.

Proof: Note that our choice of confidence radii satisfies

$$\begin{aligned}
&\sum_{t=T_0+1}^{\infty} \sum_{T=1}^{t-1} \exp\left(-\frac{T\alpha(t,T)^2}{2\sigma^2}\right) \\
&\quad \leq \sum_{t=2}^{\infty} \sum_{T=1}^{t-1} \exp\left(-\frac{T\alpha(t,T)^2}{2\sigma^2}\right) \leq \frac{\delta}{4kN}, \quad \text{and} \\
&\sum_{t=T_0+1}^{\infty} \sum_{T=1}^{t-1} \exp\left(-\frac{T\beta(t,T)^2}{2c\sigma^4}\right) \\
&\quad \leq \sum_{t=2}^{\infty} \sum_{T=1}^{t-1} \exp\left(-\frac{T\beta(t,T)^2}{2c\sigma^4}\right) \leq \frac{\delta}{4kN}.
\end{aligned}$$

By (S.23) and (S.24), and Lemma 11, we conclude that event E occurs with probability at least $1 - \frac{\delta}{k}$. □

Lemma 2 and 3 still hold for the sub-Gaussian case, since both of them are only established on the confidence bounds. Lemma 4 and 5 need to be modified.

Given a number t large enough and arm i, define $u_i(t)$ and $v_i(t)$ as the smallest numbers of arm pulls such that $\alpha(t, u_i(t)) \leq \Delta_i/2$ and $\beta(t, v_i(t)) \leq \Delta_i^{\mathrm{v}}$, i.e.

$$u_i(t) = \left\lceil \frac{2\sigma^2}{(\frac{\Delta_i}{2})^2} \ln \frac{kNt^4}{\delta} \right\rceil \quad \text{and} \quad v_i(t) = \left\lceil \frac{2c\sigma^4}{(\Delta_i^v)^2} \ln \frac{kNt^4}{\delta} \right\rceil.$$

Lemma 12 (Analogue of Lemma 4): Using Algorithm 3, then 1) for $i^\star$,

$$\mathbb{P}[T_{i^\star}(t) > 4u_{i^\star}(t), i^\star \notin \mathcal{R}_t] \leq \frac{16\sigma^2\delta}{kN\Delta_{i^\star}^2 t^4} =: A_1(i^\star)$$

2) for any suboptimal arm $i \in \mathcal{S}$,

$$\mathbb{P}[T_i(t) > 4u_i(t), i \notin \mathcal{S}_t] \leq \frac{16\sigma^2\delta}{kN\Delta_i^2 t^4} =: A_2(i)$$

3) for any feasible arm $i \in \mathcal{F}$,

$$\mathbb{P}[T_i(t) > 4v_i(t), i \notin \mathcal{F}_t] \leq \frac{4c\sigma^4\delta}{kN(\Delta_i^{\mathrm{v}})^2 t^4} =: A_3(i)$$

4) for any infeasible arm $i \in \bar{\mathcal{F}}^c$,

$$\mathbb{P}[T_i(t) > 4v_i(t), i \notin \bar{\mathcal{F}}_t^c] \leq \frac{4c\sigma^4\delta}{kN(\Delta_i^{\mathrm{v}})^2 t^4} =: A_4(i)$$

Proof: For a suboptimal arm i, note that $\Delta_i/2 \leq \bar{\mu} - \mu_i \leq \Delta_i$.

$$\begin{aligned}
&\mathbb{P}[T_i(t) > 4u_i(t), i \in \mathcal{N}_t] \\
&\leq \mathbb{P}[T_i(t) > 4u_i(t), i \notin \mathcal{S}_t] \\
&= \mathbb{P}[T_i(t) > 4u_i(t), U_i^{\mu}(t) \geq \bar{\mu}] \\
&\leq \sum_{T=4u_i(t)+1}^{\infty} \mathbb{P}[T_i(t) = T, \hat{\mu}_i(t) - \mu_i \geq \bar{\mu} - \alpha(t,T) - \mu_i] \\
&\leq \sum_{T=4u_i(t)+1}^{\infty} \exp\left(-\frac{T}{2\sigma^2}\left(\bar{\mu} - \alpha(t,T) - \mu_i\right)^2\right) \\
&\leq \sum_{T=4u_i(t)+1}^{\infty} \exp\left(-\frac{T}{2\sigma^2}\left(\frac{\Delta_i}{2} - \sqrt{\frac{2\sigma^2}{T}\ln\left(\frac{kNt^4}{\delta}\right)}\right)^2\right) \\
&\leq \sum_{T=4u_i(t)+1}^{\infty} \exp\left(-\frac{1}{2\sigma^2}\left(\frac{\Delta_i}{2}\right)^2\left(\sqrt{T} - \sqrt{u_i(t)}\right)^2\right) \\
&\leq \int_{4u_i(t)}^{\infty} \exp\left(-\frac{1}{2\sigma^2}\left(\frac{\Delta_i}{2}\right)^2\left(\sqrt{x} - \sqrt{u_i(t)}\right)^2\right) \mathrm{d}x \\
&\leq \frac{4\sigma^2\delta}{(\frac{\Delta_i}{2})^2 kNt^4}
\end{aligned}$$

The third inequality results from (S.23) and the last two inequalities can be derived using similar techniques in Lemma 8. Similarly, for arm $i^\star$,

$$\begin{aligned}
\mathbb{P}[T_{i^\star}(t) > 4u_{i^\star}(t), i^\star \in \mathcal{N}_t] &\leq \mathbb{P}[T_{i^\star}(t) > 4u_{i^\star}(t), i^\star \notin \mathcal{R}_t] \\
&\leq \frac{4\sigma^2\delta}{(\frac{\Delta_{i^\star}}{2})^2 kNt^4}
\end{aligned}$$

Note that when $T > 4v_i(t) > v_i(t) > 2\ln(kNt^4/\delta)$, (S.24) can be utilized. For arms $i \in \mathcal{F}$,

$$\mathbb{P}[T_i(t) > 4v_i(t), i \in \partial\mathcal{F}_t]$$

$$
\begin{aligned}
&\leq \mathbb{P}[T_i(t) > 4v_i(t), i \notin \mathcal{F}_t] \\
&= \mathbb{P}[T_i(t) > 4v_i(t), U_i^{\mathrm{v}}(t) \geq \bar{\sigma}^2] \\
&\leq \sum_{T=4v_i(t)+1}^{\infty} \mathbb{P}[T_i(t) = T, \hat{\sigma}_i^2(t) - \sigma_i^2 > \bar{\sigma}^2 - \sigma_i^2 - \beta(t,T)] \\
&\leq \sum_{T=4v_i(t)+1}^{\infty} \exp\left(-\frac{T}{2c\sigma^4}\left(\Delta_i^{\mathrm{v}} - \beta(t, T_i(t))\right)^2\right) \\
&\leq \sum_{T=4v_i(t)+1}^{\infty} \exp\left(-\frac{T}{2c\sigma^4}\left(\Delta_i^{\mathrm{v}} - \sqrt{\frac{\sigma^4}{T}\ln\left(\frac{kNt^4}{\delta}\right)}\right)^2\right) \\
&\leq \sum_{T=4v_i(t)+1}^{\infty} \exp\left(-\frac{(\Delta_i^{\mathrm{v}})^2}{2c\sigma^4}\left(\sqrt{T} - \sqrt{v_i(t)}\right)^2\right) \\
&\leq \int_{4v_i(t)}^{\infty} \exp\left(-2\frac{(\Delta_i^{\mathrm{v}})^2}{4c\sigma^4}\left(\sqrt{x} - \sqrt{v_i(t)}\right)^2\right) \mathrm{d}x \\
&\leq \frac{4c\sigma^4\delta}{(\Delta_i^{\mathrm{v}})^2 kNt^4}
\end{aligned}
$$

The third inequality results from (S.24) and the last two inequalities can again be derived using similar techniques in Lemma 8. Similarly, for arm $i \in \bar{\mathcal{F}}^c$,

$$
\begin{aligned}
\mathbb{P}[T_i(t) > 4v_i(t), i \in \partial\mathcal{F}_t] &\leq \mathbb{P}[T_i(t) > 4v_i(t), i \notin \bar{\mathcal{F}}_t^c] \\
&\leq \frac{4c\sigma^4\delta}{(\Delta_i^{\mathrm{v}})^2 kNt^4}
\end{aligned}
$$

This completes the proof. □

Lemma 13 (Analogue of Lemma 5): Given an instance $(\nu, \bar{\sigma}^2)$, there exists a constant $C_{k,N}$ and

$$
t^\star = \left\lceil C_{k,N} H_{\mathrm{VA},N}^{(\sigma)} \ln \frac{H_{\mathrm{VA},N}^{(\sigma)}}{\delta} \right\rceil,
$$

such that at any time step $t > t^\star$, the probability that Algorithm 3 does not terminate is at most $\frac{22\delta}{kt^2}$.

Proof: Let t be a sufficiently large integer; in particular $t > 2T_0$ (which will be justified after this lemma), and $\mathcal{T} := \{\lceil t/2 \rceil, \ldots, t-1\}$. Define events

- $\mathcal{I}_1$: $\exists\, s \in \mathcal{T}$ such that $i^\star \notin \mathcal{F}_s \cap \mathcal{R}_s$ and $T_{i^\star}(s) > \max\{4u_{i^\star}(s), 4v_{i^\star}(s)\}$.
- $\mathcal{I}_2$: $\exists\, i \in \mathcal{F} \cap \mathcal{S}, s \in \mathcal{T}$ such that $i \notin \mathcal{S}_s$ and $T_i(s) > 4u_i(s)$.
- $\mathcal{I}_3$: $\exists\, i \in \bar{\mathcal{F}}^c \cap \mathcal{R}, s \in \mathcal{T}$ such that $i \notin \bar{\mathcal{F}}_s^c$ and $T_i(s) > 4v_i(s)$
- $\mathcal{I}_4$: $\exists\, i \in \bar{\mathcal{F}}^c \cap \mathcal{S}, s \in \mathcal{T}$ such that $i \notin \bar{\mathcal{F}}_s^c \cup \mathcal{S}_s$ and $T_i(s) > \min\{4u_i(s), 4v_i(s)\}$.
- $\mathcal{I}_5$: $\exists\, s \in \mathcal{T}$ such that $E(s)$ does not occur.

If VA-LUCB terminates before $\lceil t/2 \rceil$, the statement is definitely right. If not, we assume the above five events do not occur. Based on Lemma 3, the additional time steps after $\lceil t/2 \rceil - 1$ for VA-LUCB can be upper bounded as (S.28) in Derivation 3, shown at the bottom of the next page. On the other hand, we observe that if the time step $t > t^\star := \lceil C_{k,N} H_{\mathrm{VA},N}^{(\sigma)} \ln \frac{H_{\mathrm{VA},N}^{(\sigma)}}{\delta} \rceil$ where $C_{k,N}$ is a constant that depends on the constant k and the number of arms N, we have (S.29) in Derivation 4, shown at the bottom of page 2625. So the total number of time steps is bounded by

$$
\begin{aligned}
&\left\lceil \frac{t}{2} \right\rceil - 1 + \max\{4u_{i^\star}(t), 4v_{i^\star}(t)\} + \sum_{i \in \mathcal{F} \cap \mathcal{S}} 4u_i(t) \\
&+ \sum_{i \in \bar{\mathcal{F}}^c \cap \mathcal{R}} 4v_i(t) + \sum_{i \in \bar{\mathcal{F}}^c \cap \mathcal{S}} \min\{4u_i(t), 4v_i(t)\} \leq t
\end{aligned}
$$

We then compute the probability of the event $\cup_{i \in [5]} \mathcal{I}_i$. To simplify notations used in the following, we define the hardness quantity

$$
\begin{aligned}
\widetilde{H}^{(\sigma)} := &\frac{8\sigma^2}{(\Delta_{i^\star})^2} + \frac{2c\sigma^4}{(\Delta_{i^\star}^{\mathrm{v}})^2} + \sum_{i \in \mathcal{F} \cap \mathcal{S}} \frac{8\sigma^2}{(\Delta_i)^2} \\
&+ \sum_{i \in \bar{\mathcal{F}}^c \cap \mathcal{R}} \frac{2c\sigma^4}{(\Delta_i^{\mathrm{v}})^2} + \sum_{i \in \bar{\mathcal{F}}^c \cap \mathcal{S}} g(u_i(t), v_i(t))
\end{aligned}
$$

where $g(u_i(t), v_i(t)) = \frac{8\sigma^2}{(\Delta_i)^2} \cdot \mathbb{1}\{u_i(t) < v_i(t)\} + \frac{2c\sigma^4}{(\Delta_i^{\mathrm{v}})^2} \cdot \mathbb{1}\{u_i(t) \geq v_i(t)\}$ which is exactly $\min\{\frac{2\sigma^2}{(\frac{\Delta_i}{2})^2}, \frac{2c\sigma^4}{(\Delta_i^{\mathrm{v}})^2}\}$ if we ignore the ceiling operators in $u_i(t)$ and $v_i(t)$. By Lemma 6 and Lemma 4,

$$
\begin{aligned}
\mathbb{P}[\mathcal{I}_1] &\leq \sum_{s \in \mathcal{T}} \left(\frac{16\sigma^2\delta}{kN\Delta_{i^\star}^2 s^4} + \frac{4c\sigma^4\delta}{kN(\Delta_{i^\star}^{\mathrm{v}})^2 s^4}\right) \\
\mathbb{P}[\mathcal{I}_2] &\leq \sum_{i \in \mathcal{F} \cap \mathcal{S}} \sum_{s \in \mathcal{T}} \frac{16\sigma^2\delta}{kN\Delta_i^2 s^4} \\
\mathbb{P}[\mathcal{I}_3] &\leq \sum_{i \in \bar{\mathcal{F}}^c \cap \mathcal{R}} \sum_{s \in \mathcal{T}} \frac{4c\sigma^4\delta}{kN(\Delta_i^{\mathrm{v}})^2 s^4} \\
\mathbb{P}[\mathcal{I}_4] &\leq \sum_{i \in \bar{\mathcal{F}}^c \cap \mathcal{S}} \sum_{s \in \mathcal{T}} \frac{2\delta}{kNs^4} g(u_i(s), v_i(s)) \\
\mathbb{P}[\mathcal{I}_5] &\leq \sum_{s \in \mathcal{T}} \frac{4\delta}{ks^3} \leq \frac{6\delta}{kt^2}
\end{aligned}
$$

which implies that

$$
\mathbb{P}\left[\bigcup_{i \in [5]} \mathcal{I}_i\right] \leq 2\widetilde{H}^{(\sigma)} \sum_{s \in \mathcal{T}} \frac{\delta}{kNs^4} + \frac{6\delta}{kt^2} \leq \frac{22\delta}{kt^2}
$$

where the last inequality utilizes the fact that $\widetilde{H}^{(\sigma)} < 2H_{\mathrm{VA}}^{(\sigma)} \leq 2H_{\mathrm{VA},N}^{(\sigma)} < t$ and $N \geq 2$. This yields the upper bound of the probability that the algorithm does not terminate at time step t. □

At this point, we give the constants in the above analysis. We set $k = 2$.

- According to the definition of T_0 in (S.25), $c = 64$, $\delta \leq 0.05$ and the number of arms $N \geq 2$,

$$
\begin{aligned}
T_0 &= \min\left\{t \in \mathbb{N} : t \geq 2\ln\frac{2Nt^4}{\delta}\right\} \\
&\geq \min\left\{t \in \mathbb{N} : t \geq 2\ln(80t^4)\right\} > 12.
\end{aligned}
$$

  In other words, after the warm-up procedure in Line 2 of Algorithm 3, Lemma 9 and the concentration inequality for the variance (S.24) can be applied.
- According to Lemma 11, event E occurs with probability at least $1 - \delta/2$.

- From the definition of T_0 and the computations in Lemma 13, we can easily see that

$$2T_0 \le 4\left\lceil \ln \frac{kNT_0^4}{\delta} \right\rceil < 4N + 4H^{(\sigma)}_{\mathrm{VA},N} \ln\left(\frac{kNt^4}{\delta}\right) < \frac{1}{2} C_{k,N} H^{(\sigma)}_{\mathrm{VA},N} \ln \frac{H^{(\sigma)}_{\mathrm{VA},N}}{\delta}.$$

  This justifies the assumption that $t^\star = C_{k,N} H^{(\sigma)}_{\mathrm{VA},N} \ln(H^{(\sigma)}_{\mathrm{VA},N}/\delta)$ in the analysis of Lemma 12.

The total number of arm pulls can be estimated as follows. Since $T_i(t)$ grows with t, when $T_i(t) \ge \frac{128}{c} \ln \frac{2Nt^4}{\delta} = 2 \ln \frac{2Nt^4}{\delta}$, arm i will not appear in $\mathcal{M}_t$. For any $t > t^\star$, since $4N + 4H^{(\sigma)}_{\mathrm{VA},N} \ln\left(\frac{2Nt^4}{\delta}\right) \le \frac{t}{2}$,

$$\sum_{i\in[N]} 2 \ln \frac{2Nt^4}{\delta} = 2N \ln \frac{2Nt^4}{\delta} \le \frac{t}{4}.$$

So the total number of pulls is upper bounded by $(2 + \frac{1}{4})t$ if the algorithm terminates after t time steps.

Equipped with the above preparatory results, we are now ready to present the proof of Theorem 3.

Proof of Theorem 3: According to Lemma 2 and Lemma 11, on the event E, which occurs with probability at least $1 - \delta/2$, and the termination of Algorithm 3, it succeeds. Lemma 13 indicates that Algorithm 3 terminates at time $t > t^\star > 265 > 5$ with probability at least $11\delta/t^2$. So Algorithm 3 succeeds after $O\left(H^{(\sigma)}_{\mathrm{VA},N} \ln \frac{H^{(\sigma)}_{\mathrm{VA},N}}{\delta}\right)$ time steps with probability at least $1 - (\delta/2 + 11\delta/5^2) \ge 1 - \delta$.

Note that the sample complexity is at most 2.25 times of number of time steps. To get the expected sample complexity, we first compute the expected number of time steps. According to Lemma 13, Algorithm 3 does not terminate at time step $t > t^\star$ with probability at most $\frac{11\delta}{t^2}$. The expected number of time steps is upper bounded by

$$t^\star + \sum_{t=t^\star+1}^{\infty} \frac{11\delta}{t^2} \le t^\star + \frac{11\delta}{t^\star} \le 265\, H^{(\sigma)}_{\mathrm{VA},N} \ln \frac{H^{(\sigma)}_{\mathrm{VA},N}}{\delta} + 1.$$

Thus, the expected sample complexity is upper bounded by $600 H^{(\sigma)}_{\mathrm{VA},N} \ln \frac{H^{(\sigma)}_{\mathrm{VA},N}}{\delta} + 3 = O\left(H^{(\sigma)}_{\mathrm{VA},N} \ln \frac{H^{(\sigma)}_{\mathrm{VA},N}}{\delta}\right)$. This completes the proof of Theorem 3. □

Hence, we have generalized the analysis from the bounded rewards case to the sub-Gaussian rewards case, and the conclusion is that the hardness parameter $H^{(\sigma)}_{\mathrm{VA},N}$ is merely a constant factor off from its bounded rewards counterpart H_{VA}.

Derivation 3:

$$\begin{aligned}
&\sum_{s\in\mathcal{T}} \mathbb{1}\{i_s \in (\mathcal{F}_s \cap \mathcal{N}_s) \cup (\partial\mathcal{F}_s \backslash \mathcal{S}_s) \text{ or } c_s \in (\mathcal{F}_s \cap \mathcal{N}_s) \cup (\partial\mathcal{F}_s \backslash \mathcal{S}_s)\} \\
&\le \sum_{s\in\mathcal{T}} \sum_{i\in[N]} \mathbb{1}\{i = i_s \text{ or } c_s,\, i \in (\mathcal{F}_s \cap \mathcal{N}_s) \cup (\partial\mathcal{F}_s \backslash \mathcal{S}_s)\} \\
&= \sum_{s\in\mathcal{T}} \Bigg(\mathbb{1}\{i^\star = i_s \text{ or } c_s,\, i^\star \in (\mathcal{F}_s \cap \mathcal{N}_s) \cup (\partial\mathcal{F}_s \backslash \mathcal{S}_s)\} + \sum_{i\in\mathcal{F}\cap\mathcal{S}} \mathbb{1}\{i = i_s \text{ or } c_s,\, i \in (\mathcal{F}_s \cap \mathcal{N}_s) \cup (\partial\mathcal{F}_s \backslash \mathcal{S}_s)\} \\
&\quad + \sum_{i\in\bar{\mathcal{F}}^c\cap\mathcal{R}} \mathbb{1}\{i = i_s \text{ or } c_s,\, i \in (\mathcal{F}_s \cap \mathcal{N}_s) \cup (\partial\mathcal{F}_s \backslash \mathcal{S}_s)\} + \sum_{i\in\bar{\mathcal{F}}^c\cap\mathcal{S}} \mathbb{1}\{i = i_s \text{ or } c_s,\, i \in (\mathcal{F}_s \cap \mathcal{N}_s) \cup (\partial\mathcal{F}_s \backslash \mathcal{S}_s)\} \Bigg) \\
&\le \sum_{s\in\mathcal{T}} \Bigg(\mathbb{1}\{i^\star = i_s \text{ or } c_s,\, i^\star \notin \mathcal{F}_s \cap \mathcal{R}_s\} + \sum_{i\in\mathcal{F}\cap\mathcal{S}} \mathbb{1}\{i = i_s \text{ or } c_s,\, i \notin \mathcal{S}_s\} \\
&\quad + \sum_{i\in\bar{\mathcal{F}}^c\cap\mathcal{R}} \mathbb{1}\{i = i_s \text{ or } c_s,\, i \notin \bar{\mathcal{F}}_s^c\} + \sum_{i\in\bar{\mathcal{F}}^c\cap\mathcal{S}} \mathbb{1}\{i = i_s \text{ or } c_s,\, i \notin \bar{\mathcal{F}}_s^c \cup \mathcal{S}_s\} \Bigg) \\
&\le \sum_{s\in\mathcal{T}} \Bigg(\mathbb{1}\{i^\star = i_s \text{ or } c_s,\, T_{i^\star}(s) \le \max\{4u_{i^\star}(s), 4v_{i^\star}(s)\}\} + \sum_{i\in\mathcal{F}\cap\mathcal{S}} \mathbb{1}\{i = i_s \text{ or } c_s,\, T_i(s) \le 4u_i(s)\} \\
&\quad + \sum_{i\in\bar{\mathcal{F}}^c\cap\mathcal{R}} \mathbb{1}\{i = i_s \text{ or } c_s,\, T_i(s) \le 4v_i(s)\} + \sum_{i\in\bar{\mathcal{F}}^c\cap\mathcal{S}} \mathbb{1}\{i = i_s \text{ or } c_s,\, T_i(s) \le \min\{4u_i(s), 4v_i(s)\}\} \Bigg) \\
&\le \sum_{s\in\mathcal{T}} \mathbb{1}\{i^\star = i_s \text{ or } c_s,\, T_{i^\star}(s) \le \max\{4u_{i^\star}(s), 4v_{i^\star}(s)\}\} + \sum_{i\in\mathcal{F}\cap\mathcal{S}} \sum_{s\in\mathcal{T}} \mathbb{1}\{i = i_s \text{ or } c_s,\, T_i(s) \le 4u_i(s)\} \\
&\quad + \sum_{i\in\bar{\mathcal{F}}^c\cap\mathcal{R}} \sum_{s\in\mathcal{T}} \mathbb{1}\{i = i_s \text{ or } c_s,\, T_i(s) \le 4v_i(s)\} + \sum_{i\in\bar{\mathcal{F}}^c\cap\mathcal{S}} \sum_{s\in\mathcal{T}} \mathbb{1}\{i = i_s \text{ or } c_s,\, T_i(s) \le \min\{4u_i(s), 4v_i(s)\}\} \\
&\le \max\{4u_{i^\star}(t), 4v_{i^\star}(t)\} + \sum_{i\in\mathcal{F}\cap\mathcal{S}} 4u_i(t) + \sum_{i\in\bar{\mathcal{F}}^c\cap\mathcal{R}} 4v_i(t) + \sum_{i\in\bar{\mathcal{F}}^c\cap\mathcal{S}} \min\{4u_i(t), 4v_i(t)\}. \qquad \text{(S.28)}
\end{aligned}$$

APPENDIX F
RESULTS USING LIL-BASED CONFIDENCE BOUNDS

While there exist various approaches to apply the LIL techniques to VA-LUCB algorithm [8], [13], [38], [39], we adopt a simple non-asymptotic LIL concentration bound from Jamieson et al. [8] to show that different confidence bounds utilized in VA-LUCB can lead to slightly different upper bounds on the expected stopping time.

Lemma 14 (Lemma 3 in [8]): Let $\{X_i\}_{i=1}^{\infty}$ be a sequence of i.i.d. centered sub-Gaussian random variables with scale parameter σ. Fix any $\epsilon \in (0,1)$ and $\delta \in (0, \ln(1+\epsilon)/e)$. Then one has

$$\mathbb{P}\left[\forall t \in \mathbb{N} : \sum_{s=1}^{t} X_s \leq (1+\sqrt{\epsilon})\sqrt{2\sigma^2\,(1+\epsilon)\,t \ln\left(\frac{\ln((1+\epsilon)t)}{\delta}\right)}\right] \geq 1-\xi(\delta),$$

where $\xi(\delta) := \frac{2+\epsilon}{\epsilon}\left(\frac{\delta}{\log(1+\epsilon)}\right)^{1+\epsilon}$.

Define the "good events"

$$\tilde{E}_i(t) := \left\{|\hat{\mu}_i(t) - \mu_i| \leq \tilde{\alpha}(T_i(t)), |\tilde{\sigma}_i^2(t) - \sigma_i^2| \leq \tilde{\beta}(T_i(t))\right\}$$

for all $t \in \mathbb{N}$ and $i \in [N]$, as well as their intersections

$$\tilde{E}_i := \bigcap_{t\in\mathbb{N}} \tilde{E}_i(t) \quad \text{and} \quad \tilde{E} := \bigcap_{i\in[N]} \tilde{E}_i.$$

Lemma 15: With the choice of the confidence radii in (12), the event $\tilde{E}$ occurs with probability at least $1-\xi(\delta)$. In particular, if we set $\epsilon = 0.9$ and $\delta < 0.1$, then $\xi(\delta) \geq \delta$ which implies that the event $\tilde{E}$ occurs with probability at least $1-\delta$.

Proof: We first record three facts. First, any distribution supported on $[0,1]$ is $1/2$-sub-Gaussian. Second, the rewards of an arm from different time steps are i.i.d. and the realizations from different arms are independent from each other. Third, if arm i is not pulled at time step t, all the statistics for arm i at the time step t (including sample mean, sample variance and concentration bound) remain valid in time step $t+1$. By a direct application of Lemma 14 to the sample mean $\hat{\mu}_i(t)$ and the sample second moment $\hat{M}_2(t) := \frac{1}{T_i(t)}\sum_{s=1}^{t-1} X_{s,i}^2 \mathbb{1}\{i \in \mathcal{J}_s\}$ of arm $i \in [N]$, and setting $\{Y_{s,i}\}_{s=1}^{\infty} \overset{\text{i.i.d.}}{\sim} \nu_i$, we have

$$\mathbb{P}\left[\exists\, t \geq 1 : |\hat{\mu}_i(t) - \mu_i| \geq \tilde{\alpha}(T_i(t))\right]$$
$$= \mathbb{P}\left[\exists\, t \geq 1 : \left|\frac{1}{t}\sum_{s=1}^{t} Y_{s,i} - \mu_i\right| \geq \tilde{\alpha}(t)\right] \leq 2\xi\Big(\frac{\delta}{4N}\Big), \quad \text{and}$$
$$\mathbb{P}\left[\exists\, t \geq 1 : \left|\hat{M}_2(t) - (\mu_i^2+\sigma_i^2)\right| \geq \tilde{\alpha}(T_i(t))\right]$$
$$= \mathbb{P}\left[\exists\, t \geq 1 : \left|\frac{1}{t}\sum_{s=1}^{t} Y_{s,i}^2 - (\mu_i^2+\sigma_i^2)\right| \geq \tilde{\alpha}(t)\right] \leq 2\xi\Big(\frac{\delta}{4N}\Big).$$

Derivation 4:

$$\begin{aligned}
&\max\{4u_{i^\star}(t), 4v_{i^\star}(t)\} + \sum_{i\in\mathcal{F}\cap\mathcal{S}} 4u_i(t) + \sum_{i\in\bar{\mathcal{F}}^c\cap\mathcal{R}} 4v_i(t) + \sum_{i\in\bar{\mathcal{F}}^c\cap\mathcal{S}} \min\{4u_i(t), 4v_i(t)\}\} \\
&= \max\left\{4\left\lceil\frac{2\sigma^2}{(\frac{\Delta_{i^\star}}{2})^2}\ln\left(\frac{kNt^4}{\delta}\right)\right\rceil, 4\left\lceil\frac{2c\sigma^4}{(\Delta_{i^\star}^{\mathrm{v}})^2}\ln\left(\frac{kNt^4}{\delta}\right)\right\rceil\right\} + 4\sum_{i\in\mathcal{F}\cap\mathcal{S}}\left\lceil\frac{2\sigma^2}{(\frac{\Delta_i}{2})^2}\ln\left(\frac{2Nt^4}{\delta}\right)\right\rceil \\
&\quad + \sum_{i\in\bar{\mathcal{F}}^c\cap\mathcal{R}} 4\left\lceil\frac{2c\sigma^4}{(\Delta_i^{\mathrm{v}})^2}\ln\left(\frac{kNt^4}{\delta}\right)\right\rceil + \sum_{i\in\bar{\mathcal{F}}^c\cap\mathcal{S}} \min\left\{4\left\lceil\frac{2\sigma^2}{(\frac{\Delta_i}{2})^2}\ln\left(\frac{kNt^4}{\delta}\right)\right\rceil, 4\left\lceil\frac{2c\sigma^4}{(\Delta_i^{\mathrm{v}})^2}\ln\left(\frac{kNt^4}{\delta}\right)\right\rceil\right\} \\
&\leq 4N + 4\max\left\{\frac{2\sigma^2}{(\frac{\Delta_{i^\star}}{2})^2}\ln\left(\frac{kNt^4}{\delta}\right), \frac{2c\sigma^4}{(\Delta_{i^\star}^{\mathrm{v}})^2}\ln\left(\frac{kNt^4}{\delta}\right)\right\} + \sum_{i\in\mathcal{F}\cap\mathcal{S}}\frac{2\sigma^2}{(\frac{\Delta_i}{2})^2}\ln\left(\frac{2Nt^4}{\delta}\right) \\
&\quad + 4\sum_{i\in\bar{\mathcal{F}}^c\cap\mathcal{R}}\frac{2c\sigma^4}{(\Delta_i^{\mathrm{v}})^2}\ln\left(\frac{kNt^4}{\delta}\right) + 4\sum_{i\in\bar{\mathcal{F}}^c\cap\mathcal{S}}\min\left\{\frac{2\sigma^2}{(\frac{\Delta_i}{2})^2}\ln\left(\frac{kNt^4}{\delta}\right), \frac{2c\sigma^4}{(\Delta_i^{\mathrm{v}})^2}\ln\left(\frac{kNt^4}{\delta}\right)\right\} \\
&= 4N + 4H_{\mathrm{VA}}^{(\sigma)}\ln\left(\frac{kNt^4}{\delta}\right) \\
&\leq 4N + 4H_{\mathrm{VA},N}^{(\sigma)}\ln\left(\frac{kNt^4}{\delta}\right) \\
&= \left(4N + 4H_{\mathrm{VA},N}^{(\sigma)}\ln k\right) + 4H_{\mathrm{VA},N}^{(\sigma)}\ln\frac{N}{\delta} + 16H_{\mathrm{VA},N}^{(\sigma)}\ln\left(C_{k,N}H_{\mathrm{VA},N}^{(\sigma)}\ln\frac{H_{\mathrm{VA},N}^{(\sigma)}}{\delta}\right) \\
&\leq \left(4N + 4H_{\mathrm{VA},N}^{(\sigma)}\ln k + 16H_{\mathrm{VA},N}^{(\sigma)}\ln C_{k,N}\right) + 4H_{\mathrm{VA},N}^{(\sigma)}\ln\frac{N}{\delta} + 32H_{\mathrm{VA},N}^{(\sigma)}\ln\frac{H_{\mathrm{VA},N}^{(\sigma)}}{\delta} \\
&\leq \frac{1}{2}C_{k,N}H_{\mathrm{VA},N}^{(\sigma)}\ln\frac{H_{\mathrm{VA},N}^{(\sigma)}}{\delta} \\
&\leq \frac{t}{2}.
\end{aligned} \tag{S.29}$$

Since the rewards are in $[0,1]$, $|\hat{\mu}_i^2(t)-\mu_i^2| \le |\hat{\mu}_i(t)+\mu_i| \cdot |\hat{\mu}_i(t)-\mu_i| \le 2|\hat{\mu}_i(t)-\mu_i|$. Using this and the triangle inequality, we obtain for every $t \ge 1$,

$$\begin{aligned}\left|\tilde{\sigma}_i^2(t)-\sigma_i^2\right| &\le \left|\hat{M}_2(t)-(\mu_i^2+\sigma_i^2)\right| + \left|\hat{\mu}_i^2(t)-\mu_i^2\right| \\ &\le \tilde{\alpha}(T_i(t)) + 2\tilde{\alpha}(T_i(t)) = \tilde{\beta}(T_i(t)).\end{aligned}$$

Therefore, by a union bound, with probability at least

$$\begin{aligned}1-4N\cdot\xi\Big(\frac{\delta}{4N}\Big) &= 1-4N\cdot\frac{2+\epsilon}{\epsilon}\left(\frac{\delta}{4N\log(1+\epsilon)}\right)^{1+\epsilon} \\ &= 1-(4N)^{-\epsilon}\xi(\delta) \ge 1-\xi(\delta)\end{aligned}$$

event $\tilde{E}$ occurs. □

Theorem 4 (LIL-Based Upper Bound): Given an instance $(\nu,\bar{\sigma}^2)$ and confidence parameter δ, with probability at least $1-\xi(\delta)$, VA-LUCB with the LIL-based confidence bounds succeeds and terminates in

$$O\left(H_{\mathrm{VA}}^{(1)}\ln\frac{N}{\delta}+H_{\mathrm{VA}}^{(3)}\right)$$

time steps,where $H_{\mathrm{VA}}^{(1)}$ and $H_{\mathrm{VA}}^{(3)}$ are defined in (14) and (15) respectively.

Proof: For a suboptimal arm $i \in \mathcal{S}$, when $\tilde{\alpha}(T_i(t)) \le \Delta_i/4$, $\hat{\mu}_i(t)+\tilde{\alpha}(T_i(t)) \le \mu_i + 2\tilde{\alpha}(T_i(t)) \le \mu_i + \frac{\Delta_i}{2} \le \bar{\mu}$, which indicates arm i is not in $\mathcal{N}_t$. The same holds for $i^\star$. For a feasible arm $i \in \mathcal{F}$, when $\tilde{\beta}(T_i(t)) \le \Delta_i^{\mathrm{v}}/2$, $\tilde{\sigma}_i^2(t)+\tilde{\beta}(T_i(t)) \le \mu_i + 2\tilde{\beta}(T_i(t)) \le \bar{\sigma}^2$, which indicates arm i is not in $\partial\mathcal{F}_t$. The same holds for the infeasible arms $i \in \bar{\mathcal{F}}^c$. By a direct computation (see, for example, [12, Eqn. (4)]),

$$\min\left\{t\in\mathbb{N}:\tilde{\alpha}(t)\le\frac{\Delta_i}{4}\right\} \le \frac{2\gamma}{\Delta_i^2}\ln\left(\frac{8N\ln(\gamma\Delta_i^{-2})}{\delta}\right), \text{ and}$$

$$\min\left\{t\in\mathbb{N}:\tilde{\beta}(t)\le\frac{\Delta_i^{\mathrm{v}}}{2}\right\} \le \frac{2\gamma}{(\frac{2}{3}\Delta_i^{\mathrm{v}})^2}\ln\left(\frac{8N\ln(\gamma(\frac{2}{3}\Delta_i^{\mathrm{v}})^{-2})}{\delta}\right),$$

where $\gamma=\gamma_\epsilon=8(1+\sqrt{\epsilon})^2(1+\epsilon)^2$. According to Lemma 3 (which also holds even if the confidence radii have been changed), at least one of pulled arms belongs to the set $(\partial\mathcal{F}_t\backslash\mathcal{S}_t)\cup(\mathcal{F}_t\cap\mathcal{N}_t)$, so after

$$\begin{aligned}\tilde{t} :=& \max\left\{\frac{2\gamma}{\Delta_{i^\star}^2}\ln\left(\frac{8N\ln(\gamma\Delta_{i^\star}^{-2})}{\delta}\right),\right. \\ &\left.\frac{2\gamma}{(\frac{2}{3}\Delta_{i^\star}^{\mathrm{v}})^2}\ln\left(\frac{8N\ln(\gamma(\frac{2}{3}\Delta_{i^\star}^{\mathrm{v}})^{-2})}{\delta}\right)\right\} \\ &+\sum_{i\in\mathcal{F}\cap\mathcal{S}}\frac{2\gamma}{\Delta_i^2}\ln\left(\frac{8N\ln(\gamma\Delta_i^{-2})}{\delta}\right) \\ &+\sum_{i\in\bar{\mathcal{F}}^c\cap\mathcal{R}}\frac{2\gamma}{(\frac{2}{3}\Delta_i^{\mathrm{v}})^2}\ln\left(\frac{8N\ln(\gamma(\frac{2}{3}\Delta_i^{\mathrm{v}})^{-2})}{\delta}\right) \\ &+\sum_{i\in\bar{\mathcal{F}}^c\cap\mathcal{S}}\min\left\{\frac{2\gamma}{\Delta_i^2}\ln\left(\frac{8N\ln(\gamma\Delta_i^{-2})}{\delta}\right),\right. \\ &\left.\frac{2\gamma}{(\frac{2}{3}\Delta_i^{\mathrm{v}})^2}\ln\left(\frac{8N\ln(\gamma(\frac{2}{3}\Delta_i^{\mathrm{v}})^{-2})}{\delta}\right)\right\}\end{aligned}$$

time steps, the algorithm must have terminated. Note that for $x\in(0,1)$, $\frac{1}{x^2}\ln\left(\frac{8N\ln(\gamma x^{-2})}{\delta}\right)$ decreases as x increases and

$$\begin{aligned}&\frac{1}{x^2}\ln\left(\frac{8N\ln(\gamma x^{-2})}{\delta}\right) \\ &= \frac{1}{x^2}\left(\ln\frac{8N}{\delta}+\ln\big(\ln(\gamma(1+\epsilon)+\ln x^{-2})\big)\right) \\ &\le \frac{1}{x^2}\ln\frac{8N}{\delta}+\frac{1}{x^2}\left(\ln\ln_+(x^{-2})+\gamma\right) \\ &\le \frac{c_1}{x^2}\ln\frac{N}{\delta}+\frac{\ln\ln_+(x^{-2})}{x^2}\end{aligned}$$

where $c_1=\ln 8+\gamma$ is a known constant that only depends on ϵ. Thus, $\tilde{t}$ can be upper bounded by

$$2\gamma c_1 H_{\mathrm{VA}}^{(1)}\ln\frac{N}{\delta}+2\gamma H_{\mathrm{VA}}^{(3)} = O\left(H_{\mathrm{VA}}^{(1)}\ln\frac{N}{\delta}+H_{\mathrm{VA}}^{(3)}\right)$$

where $H_{\mathrm{VA}}^{(1)}$ and $H_{\mathrm{VA}}^{(3)}$ are defined in (14) and (15) respectively. □

APPENDIX G
PROOF OF LOWER BOUND

Let $\mathrm{KL}(\nu,\nu')$ denote the KL divergence between distributions ν and ν', and

$$d(x,y) := x\ln\left(\frac{x}{y}\right)+(1-x)\ln\left(\frac{1-x}{1-y}\right)$$

denote the Kullback–Leibler (KL) divergence between the Bernoulli distributions $\mathrm{Bern}(x)$ and $\mathrm{Bern}(y)$.

Lemma 16 (Pinsker's and reverse Pinsker's inequality [44]): Let P and Q be two distributions that are defined in the same finite space $\mathcal{A}$ and have the same support. We have

$$\delta(P,Q)^2 \le \frac{1}{2}\mathrm{KL}(P,Q) \le \frac{1}{\alpha_Q}\delta(P,Q)^2$$

where $\delta(P,Q) := \sup\{|P(A)-Q(A)| : A\subset\mathcal{A}\} = \frac{1}{2}\sum_{x\in\mathcal{A}}|P(x)-Q(x)|$ is the total variational distance, and $\alpha_Q := \min_{x\in\mathcal{A}:Q(x)>0}Q(x)$.

Lemma 17 (Lemma 1 in [10]): For any $1\le j\le N$,

$$\sum_{i=1}^N \mathbb{E}_{\mathcal{G}_0}[T_i(\tau^{(0)})]\cdot d\big(\mu_i^{(0)},\mu_i^{(j)}\big) \ge \sup_{\mathcal{E}\in\mathcal{G}_0} d\big(\mathbb{P}_{\mathcal{G}_0}(\mathcal{E}),\ \mathbb{P}_{\mathcal{G}_j}(\mathcal{E})\big).$$

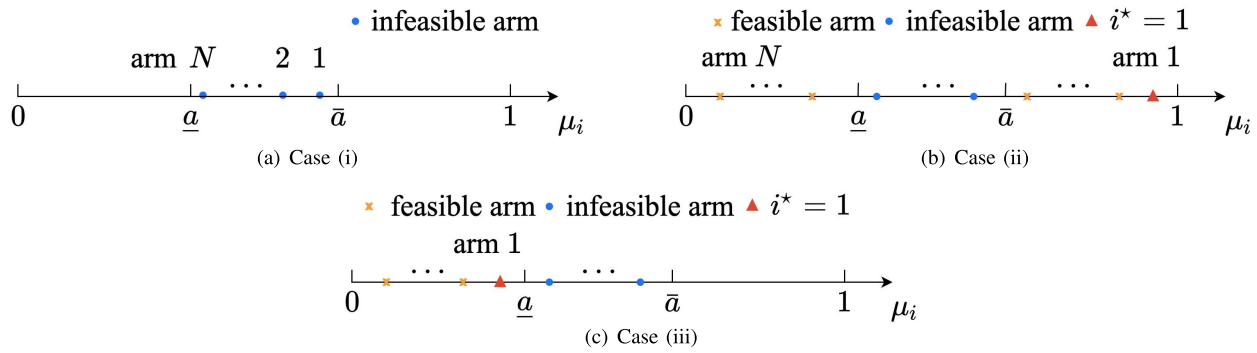
Proof of Theorem 2: Fix a δ-PAC algorithm π. We consider the instances containing arms with Bernoulli reward distributions. By simple algebra, an arm $i\in[N]$ with reward distribution $\nu_i=\mathrm{Bern}(\mu_i)$ and hence variance $\mu_i(1-\mu_i)$ is infeasible if and only if $\mu_i\in(\underline{a},\bar{a})$ and this arm is feasible otherwise.

Step 1 (Classification of Instances): Based on the values of $\{\mu_i\}_{i=1}^N$, we have one of the following cases:

(i) $\underline{a}<\mu_N\le\ldots\le\mu_2\le\mu_1<\bar{a}$,
(ii) $0<\mu_N\le\ldots\le\mu_2<\mu_1<1$, and $\bar{a}<\mu_1$,
(iii) $\mu_i<\bar{a}$ for all $i\in[N]$, $\{i\in[N]:\mu_i=\underline{a}\}=\emptyset$, $\{i\in[N]:\mu_i<\underline{a}\}\ne\emptyset$, and arm $1=\mathrm{argmax}\{\mu_i:\mu_i<\underline{a}\}$.

which are shown in Figure 7. There exists no feasible arm in Case (i), while there exists at least one feasible arm in Cases (ii) and (iii). Then we construct instances by making small modifications to the reward distributions of the arms. These constructed instances are hard to distinguish from each of the cases above, in the sense that the constructed instances will lead to different conclusions towards the feasibility of the instance or the best feasible arm.

Step 2 (Analysis of Each Case): Subsequently, we analyze each case individually. In each case, we construct $N+1$



(a) Case (i) (b) Case (ii)

(c) Case (iii)

Fig. 7. Illustrations of the three cases for the proof of the lower bound.

instances such that under instance j ($0 \le j \le N$), the stochastic reward of arm i ($1 \le i \le N$) is drawn from distribution

$$\nu_i^{(j)} := \nu_i \cdot \mathbb{1}\{i \neq j\} + \nu_i' \cdot \mathbb{1}\{i = j\},$$

where ν_i' will be specified in each case. Let $\mu_i^{(j)}$ denote the expectation of arm i and $(\sigma_i^{(j)})^2$ denote the variance of arm i under instance j. Under instance $0 \le j \le N$, we define several other notations as follows:

- Let $X_{i,r}^{(j)}$ be the random reward of arm i at round r. Then $X_{i,r}^{(j)} \in \{0, 1\}$.
- Let $i_r^{(j)}$ be the pulled arm at round r, and $\mathcal{G}_r^{(j)} = \{(i_s^{(j)}, X_{i_s^{(j)},s})\}_{s=1}^r$ be the sequence of pulled arms and observed rewards up to and including round r.
- Let $\tau^{(j)}$ denote the stopping time.

For simplicity, we abbreviate $\mathcal{G}_{\tau^{(j)}}^{(j)}$ as $\mathcal{G}_j$.

Case (i): $\underline{a} < \mu_N \le \ldots \le \mu_2 \le \mu_1 < \bar{a}$

Construction of Instances: Fix any $0 < b_1 < \underline{a} < \bar{a} < b_2 < 1$. We define $\nu_i' = \mathrm{Bern}(\mu_i')$ for arm $i \in [N]$ with

$$\mu_i' = \begin{cases} b_1 & \mu_i < 1/2, \\ b_2 & \mu_i \ge 1/2. \end{cases}$$

Therefore,

- under instance 0, since $(\sigma_i^{(0)})^2 = \mu_i(1-\mu_i) > \bar{\sigma}^2$ for all arm $i \in [N]$, there is no feasible arm;
- under instance j ($1 \le j \le L$), we see that

$$(\sigma_i^{(j)})^2 = \mu_i(1-\mu_i) > \bar{\sigma}^2, \quad \forall\, i \neq j, \text{ and}$$
$$(\sigma_j^{(j)})^2 = \mu_i'(1-\mu_i') < \bar{\sigma}^2,$$

  implying that j is the unique optimal feasible arm.

Since algorithm π is δ-PAC, we have $\mathbb{P}_{\mathcal{G}_0}[i_{\text{out}} = j] < \delta$ and $\mathbb{P}_{\mathcal{G}_j}[i_{\text{out}} \neq j] < \delta$ for all $1 \le j \le N$.

Change Measure: Next, we lower bound $\mathbb{E}_{\mathcal{G}_0}[T_i(\tau^{(0)})]$ with the KL divergence by applying Lemma 17.

Note that $d(a, b)$ equals to the KL divergence between $\mathrm{Bern}(a)$ and $\mathrm{Bern}(b)$ and

$$d(x, 1-x) \ge \ln\left(\frac{1}{2.4x}\right) \quad \forall\, x \in (0, 1].$$

Let $\mathcal{E}_j := \{i_{\text{out}} = j\}$, then

$$\delta \ge \mathbb{P}_{\mathcal{G}_0}[\{i_{\text{out}}\} \neq \emptyset] \ge \mathbb{P}_{\mathcal{G}_0}[\mathcal{E}_j],$$
$$1 - \delta \le \mathbb{P}_{\mathcal{G}_j}[i_{\text{out}} = j] = \mathbb{P}_{\mathcal{G}_j}[\mathcal{E}_j].$$

Since $\mu_i^{(0)} = \mu_i^{(j)}$ for all $i \neq j$ under instance $1 \le j \le N$, we have

$$\mathbb{E}_{\mathcal{G}_0}[T_j(\tau^{(0)})] \ge \frac{d(\mathbb{P}_{\mathcal{G}_0}[\mathcal{E}_j],\ \mathbb{P}_{\mathcal{G}_j}[\mathcal{E}_j])}{d(\mu_j^{(0)}, \mu_j^{(j)})} \ge \frac{d(\delta, 1-\delta)}{d(\mu_j, \mu_j')} \ge \frac{-\ln(2.4\delta)}{d(\mu_j, \mu_j')}.$$

Therefore,

$$\mathbb{E}_{\mathcal{G}_0}[\tau^{(0)}] \ge \sum_{j=1}^N \mathbb{E}_{\mathcal{G}_0}[T_j(\tau^{(0)})] \ge \ln\left(\frac{1}{2.4\delta}\right) \cdot \sum_{i=1}^N \frac{1}{d(\mu_i, \mu_i')}$$
$$= \ln\left(\frac{1}{2.4\delta}\right) \cdot \left(\sum_{i:\mu_i < 1/2} \frac{1}{d(\mu_i, b_1)} + \sum_{i:\mu_i \ge 1/2} \frac{1}{d(\mu_i, b_2)}\right).$$

Since (b_1, b_2) can be chosen arbitrarily from $B = \{(b_1, b_2) : 0 < b_1 < \underline{a} < \bar{a} < b_2 < 1\}$, we have

$$\mathbb{E}_{\mathcal{G}_0}[\tau^{(0)}]$$
$$\ge \sup_{(b_1,b_2)\in B} \ln\left(\frac{1}{2.4\delta}\right) \cdot \left(\sum_{i:\mu_i < 1/2} \frac{1}{d(\mu_i, b_1)} + \sum_{i:\mu_i \ge 1/2} \frac{1}{d(\mu_i, b_2)}\right)$$
$$= \ln\left(\frac{1}{2.4\delta}\right) \cdot \left(\sum_{i:\mu_i < 1/2} \frac{1}{d(\mu_i, \underline{a})} + \sum_{i:\mu_i \ge 1/2} \frac{1}{d(\mu_i, \bar{a})}\right).$$

Case (ii): $0 < \mu_N \le \ldots \le \mu_2 < \mu_1 < 1$, and $\bar{a} < \mu_1$.

Construction of Instances: Fix any $\underline{a} < b_1 < \bar{a} < \mu_1 < b_2 < 1$. We define $\nu_1' = \mathrm{Bern}(b_1)$ and $\nu_i' = \mathrm{Bern}(b_2)$ for all arms $i \neq 1$. Therefore,

- under instance 0, we see that

$$\mu_1^{(0)} = \mu_1 > \mu_i = \mu_i^{(0)} \quad \forall\, i \neq 1, \text{ and}$$
$$(\sigma_1^{(0)})^2 = \mu_1(1-\mu_1) < \bar{a}(1-\bar{a}) = \bar{\sigma}^2;$$

- under instance 1, we see that

$$(\sigma_1^{(1)})^2 = b_1(1-b_1) > \bar{a}(1-\bar{a}) = \bar{\sigma}^2;$$

- under instance j ($2 \le j \le L$), we see that

$$\mu_j^{(j)} = b_2 > \mu_i = \mu_i^{(j)} \quad \forall\, i \neq j, \text{ and}$$

$$(\sigma_j^{(j)})^2 = b_2(1-b_2) < \bar{a}(1-\bar{a}) = \bar{\sigma}^2.$$

Since arm 1 is the unique best feasible arm under instance 0, arm 1 is not feasible under instance 1, and arm j is the unique best feasible arm under instance j ($2 \le j \le N$), we have $\mathbb{P}_{\mathcal{G}_0}[i_{\text{out}} \neq 1] < \delta$ and $\mathbb{P}_{\mathcal{G}_1}[i_{\text{out}} = 1] < \delta$, and $\mathbb{P}_{\mathcal{G}_j}[i_{\text{out}} \neq j] < \delta$ for all $2 \le j \le N$.

Change of Measure: We again lower bound $\mathbb{E}_{\mathcal{G}_0}[T_i(\tau^{(0)})]$ with the KL divergence by applying Lemma 17. Let $\mathcal{E} := \{i_{\text{out}} \neq 1\}$, then

$$\delta \ge \mathbb{P}_{\mathcal{G}_0}[i_{\text{out}} \neq 1] = \mathbb{P}_{\mathcal{G}_0}[\mathcal{E}], 1-\delta \le \mathbb{P}_{\mathcal{G}_1}[i_{\text{out}} \neq 1] = \mathbb{P}_{\mathcal{G}_1}[\mathcal{E}],$$
$$1-\delta \le \mathbb{P}_{\mathcal{G}_j}[i_{\text{out}} = j] \le \mathbb{P}_{\mathcal{G}_j}[i_{\text{out}} \neq 1] = \mathbb{P}_{\mathcal{G}_j}[\mathcal{E}]$$

for $\forall\, 2 \le j \le L$. Since $\mu_i^{(0)} = \mu_i^{(j)}$ for all $i \neq j$ under instance $1 \le j \le N$, we have

$$\mathbb{E}_{\mathcal{G}_0}[T_j(\tau^{(0)})] \ge \frac{d\big(\mathbb{P}_{\mathcal{G}_0}(\mathcal{E}),\ \mathbb{P}_{\mathcal{G}_j}(\mathcal{E})\big)}{d\big(\mu_j^{(0)}, \mu_j^{(j)}\big)} \ge \frac{d(\delta, 1-\delta)}{d\big(\mu_j^{(0)}, \mu_j^{(j)}\big)}$$
$$\ge \frac{-\ln(2.4\delta)}{d\big(\mu_j^{(0)}, \mu_j^{(j)}\big)} \ge \begin{cases} \dfrac{-\ln(2.4\delta)}{d(\mu_1, b_1)}, & j = 1 \\ \dfrac{-\ln(2.4\delta)}{d(\mu_j, b_2)}, & 2 \le j \le L \end{cases}.$$

Therefore,

$$\mathbb{E}_{\mathcal{G}_0}[\tau^{(0)}] \ge \ln\left(\frac{1}{2.4\delta}\right) \cdot \left(\frac{1}{d(\mu_1, b_1)} + \sum_{i=2}^{N} \frac{1}{d(\mu_i, b_2)}\right).$$

Since (b_1, b_2) can be chosen arbitrarily from $B = \{(b_1, b_2) : \underline{a} < b_1 < \bar{a} < \mu_1 < b_2 < 1\}$, we have

$$\mathbb{E}_{\mathcal{G}_0}[\tau^{(0)}]$$
$$\ge \sup_{(b_1,b_2)\in B} \ln\left(\frac{1}{2.4\delta}\right) \cdot \left(\frac{1}{d(\mu_1, b_1)} + \sum_{i=2}^{N} \frac{1}{d(\mu_i, b_2)}\right)$$
$$= \ln\left(\frac{1}{2.4\delta}\right) \cdot \left(\frac{1}{d(\mu_1, \bar{a})} + \sum_{i=2}^{N} \frac{1}{d(\mu_i, \mu_1)}\right).$$

Case (iii): $\mu_i < \bar{a}$ for all $i \in [N]$, $\{i \in [N] : \mu_i = \underline{a}\} = \emptyset$, $\{i \in [N] : \mu_i < \underline{a}\} \neq \emptyset$, and arm $1 = \operatorname{argmax}\{\mu_i : \mu_i < \underline{a}\}$.

Construction of Instances: Fix any $\mu_1 < b_1 < \underline{a} < b_2 < \bar{a} < b_3 < 1$. We define $\nu_i' = \text{Bern}(\mu_i')$ for arm $i \in [N]$ with

$$\mu_i' = \begin{cases} b_1, & \mu_i < 1/2 \text{ and } \mu_i \neq \mu_1, \\ b_2, & \mu_i = \mu_1, \\ b_3, & \mu_i \ge 1/2. \end{cases}$$

Therefore, arm 1 is the unique best feasible arm under instance 0, arm 1 is not feasible under instance 1, and arm j is the unique best feasible arm under instance j ($2 \le j \le N$), we have $\mathbb{P}_{\mathcal{G}_0}[i_{\text{out}} \neq 1] < \delta$, $\mathbb{P}_{\mathcal{G}_1}[i_{\text{out}} = 1] < \delta$, and $\mathbb{P}_{\mathcal{G}_j}[i_{\text{out}} \neq j] < \delta$ for all $2 \le j \le N$.

Change of Measure: We again lower bound $\mathbb{E}_{\mathcal{G}_0}[T_i(\tau^{(0)})]$ with the KL divergence by applying Lemma 17. Let $\mathcal{E} := \{i_{\text{out}} \neq 1\}$, then

$$\delta \ge \mathbb{P}_{\mathcal{G}_0}[i_{\text{out}} \neq 1] = \mathbb{P}_{\mathcal{G}_0}[\mathcal{E}], 1-\delta \le \mathbb{P}_{\mathcal{G}_1}[i_{\text{out}} \neq 1] = \mathbb{P}_{\mathcal{G}_1}[\mathcal{E}].$$
$$1-\delta \le \mathbb{P}_{\mathcal{G}_j}[i_{\text{out}} = j] \le \mathbb{P}_{\mathcal{G}_j}[i_{\text{out}} \neq 1] = \mathbb{P}_{\mathcal{G}_j}[\mathcal{E}]$$

for $\forall\, 2 \le j \le L$. Since $\mu_i^{(0)} = \mu_i^{(j)}$ for all $i \neq j$ under instance $1 \le j \le N$, we have

$$\mathbb{E}_{\mathcal{G}_0}[T_j(\tau^{(0)})] \ge \frac{d\big(\mathbb{P}_{\mathcal{G}_0}[\mathcal{E}], \mathbb{P}_{\mathcal{G}_j}[\mathcal{E}]\big)}{d\big(\mu_j^{(0)}, \mu_j^{(j)}\big)} \ge \frac{d(\delta, 1-\delta)}{d\big(\mu_j, \mu_j'\big)}$$
$$\ge \frac{-\ln(2.4\delta)}{d\big(\mu_j, \mu_j'\big)} \ge \begin{cases} \dfrac{-\ln(2.4\delta)}{d(\mu_1, b_1)}, & \mu_j < 1/2 \text{ and } \mu_j \neq \mu_1 \\ \dfrac{-\ln(2.4\delta)}{d(\mu_j, b_2)}, & \mu_j = \mu_1 \\ \dfrac{-\ln(2.4\delta)}{d(\mu_j, b_3)}, & \mu_j \ge 1/2 \end{cases}.$$

Therefore,

$$\mathbb{E}_{\mathcal{G}_0}[\tau^{(0)}] \ge \sum_{j=1}^{N} \mathbb{E}_{\mathcal{G}_0}[T_j(\tau^{(0)})]$$
$$\ge \ln\left(\frac{1}{2.4\delta}\right) \cdot \left(\frac{1}{d(\mu_1, b_2)} + \sum_{\substack{i:\mu_i<1/2,\\ \mu_i\neq\mu_1}} \frac{1}{d(\mu_i, b_1)} + \sum_{i:\mu_i\ge 1/2} \frac{1}{d(\mu_i, b_3)}\right).$$

Since (b_1, b_2, b_3) can be chosen arbitrarily from $B = \{(b_1, b_2, b_3) : \mu_1 < b_1 < \underline{a} < b_2 < \bar{a} < b_3 < 1\}$, we have

$$\mathbb{E}_{\mathcal{G}_0}[\tau^{(0)}] \ge \sup_{(b_1,b_2,b_3)\in B} \ln\left(\frac{1}{2.4\delta}\right) \cdot \left(\frac{1}{d(\mu_1, b_2)} + \sum_{\substack{i:\mu_i<1/2,\\ \mu_i\neq\mu_1}} \frac{1}{d(\mu_i, b_1)} + \sum_{i:\mu_i\ge 1/2} \frac{1}{d(\mu_i, b_3)}\right)$$
$$= \ln\left(\frac{1}{2.4\delta}\right) \cdot \left(\frac{1}{d(\mu_1, \underline{a})} + \sum_{\substack{i:\mu_i<1/2,\\ \mu_i\neq\mu_1}} \frac{1}{d(\mu_i, \mu_1)} + \sum_{i:\mu_i\ge 1/2} \frac{1}{d(\mu_i, \bar{a})}\right).$$

Step 3 (Simplification of the Bounds With H_{VA}): We further lower bound the sample complexity in each case.

Case (i): When $\underline{a} < \mu_N \le \ldots \le \mu_2 \le \mu_1 < \bar{a}$, we have

$$\mathbb{E}_{\mathcal{G}_0}[\tau^{(0)}] \ge \ln\left(\frac{1}{2.4\delta}\right) \cdot \left(\sum_{i:\mu_i<1/2} \frac{1}{d(\mu_i, \underline{a})} + \sum_{i:\mu_i\ge 1/2} \frac{1}{d(\mu_i, \bar{a})}\right).$$

By Theorem 16, we have

$$d(\mu_i, \bar{a}) \le \frac{1}{\underline{a}} \cdot (\mu_i - \bar{a})^2 \quad \text{and} \quad d(\mu_i, \underline{a}) \le \frac{1}{\underline{a}} \cdot (\mu_i - \underline{a})^2,$$
$$\text{where} \quad \underline{a} < 1/2 < \bar{a} = 1 - \underline{a}.$$

Let σ_a^2 be the variance of $\text{Bern}(a)$ and σ_b^2 be the variance of $\text{Bern}(b)$ for any $a, b \in (0,1)$. Then

$$\sigma_a^2 - \sigma_b^2 = a(1-a) - b(1-b) = a - b - a^2 + b^2$$
$$= (a-b)(1-a-b),$$
$$(\sigma_a^2 - \sigma_b^2)^2 = (a-b)^2(1-a-b)^2.$$

Note that $\bar{a}(1-\bar{a}) = \bar{\sigma}^2$. For $\mu_i \geq 1/2$,

$$d(\mu_i, \bar{a}) \leq \frac{(\mu_i - \bar{a})^2 \cdot (1-\mu_i-\bar{a})^2}{\underline{a}(1-\mu_i-\bar{a})^2} = \frac{(\sigma_i^2 - \bar{\sigma}^2)^2}{\underline{a}(1-\mu_i-\bar{a})^2} \leq \frac{(\sigma_i^2 - \bar{\sigma}^2)^2}{\underline{a}(\bar{a}-1/2)^2} = \frac{(\sigma_i^2 - \bar{\sigma}^2)^2}{\underline{a}(1/2-\underline{a})^2};$$

for $\mu_i < 1/2$,

$$d(\mu_i, \underline{a}) \leq \frac{(\mu_i - \underline{a})^2 \cdot (1-\mu_i-\underline{a})^2}{\underline{a}(1-\mu_i-\underline{a})^2} = \frac{(\sigma_i^2 - \bar{\sigma}^2)^2}{\underline{a}(1-\mu_i-\underline{a})^2} \leq \frac{(\sigma_i^2 - \bar{\sigma}^2)^2}{\underline{a}(1/2-\underline{a})^2}.$$

Since there is no feasible arm, $\bar{\mathcal{F}}^c \cap \mathcal{R} = [N]$ and $\bar{\mathcal{F}}^c \cap \mathcal{S} = \mathcal{F} = \emptyset$. Notice an obvious fact

$$(1/2-\underline{a})^2 = \underline{a}^2 - \underline{a} + 1/4 = 1/4 - \bar{\sigma}^2, \tag{S.30}$$

thus

$$\mathbb{E}[\tau] \geq \underline{a}(1/4-\bar{\sigma}^2)\ln\left(\frac{1}{2.4\delta}\right) \cdot \sum_{i\in[N]} \frac{1}{(\sigma_i^2-\bar{\sigma}^2)^2} = H_{\mathrm{VA}} \ln\left(\frac{1}{2.4\delta}\right) \cdot \underline{a}(1/4-\bar{\sigma}^2).$$

Case (ii): When $0 < \mu_N \leq \ldots \leq \mu_2 < \mu_1 < 1$, and $\bar{a} < \mu_1$, we have

$$\mathbb{E}_{\mathcal{G}_0}[\tau^{(0)}] \geq \ln\left(\frac{1}{2.4\delta}\right) \cdot \left(\frac{1}{d(\mu_1, \bar{a})} + \sum_{j=2}^{N} \frac{1}{d(\mu_j, \mu_1)}\right).$$

We first apply Theorem 16 to see that

$$d(\mu_i, \mu_1) \leq \frac{(\mu_i - \mu_1)^2}{1-\mu_1}, \quad \forall\, i \neq 1,$$

$$d(\mu_1, \bar{a}) \leq \frac{(\mu_1 - \bar{a})^2 \cdot (1-\mu_1-\bar{a})^2}{\underline{a}(1-\mu_1-\bar{a})^2} = \frac{(\sigma_1^2 - \bar{\sigma}^2)^2}{\underline{a}(1-\mu_1-\bar{a})^2} \leq \frac{(\sigma_1^2 - \bar{\sigma}^2)^2}{\underline{a}(2\bar{a}-1)^2} = \frac{(\sigma_1^2 - \bar{\sigma}^2)^2}{4\underline{a}(1/2-\underline{a})^2}. \tag{S.31}$$

Since $\bar{\mathcal{F}}^c \cap \mathcal{R}$ is empty, $\mathcal{S} = [N] \setminus \{i^\star\}$ and hence

$$\Delta_1 = \Delta_{i^\star} = \min_{i\in\mathcal{S}} \Delta_i \cdot \mathbb{1}\{\mathcal{S} \neq \emptyset\} + \infty \cdot \mathbb{1}\{\mathcal{S} = \emptyset\} = \min_{i\in[N]\setminus\{i^\star\}} \Delta_i. \tag{S.32}$$

Lastly,

$$\begin{aligned}
\mathbb{E}[\tau] &\overset{(a)}{\geq} \ln\left(\frac{1}{2.4\delta}\right) \cdot \frac{4\underline{a}(1/2-\underline{a})^2}{(\sigma_1^2-\bar{\sigma}^2)^2} + (1-\mu_1)\ln\left(\frac{1}{2.4\delta}\right) \cdot \sum_{i=2}^{N} \frac{1}{(\mu_i-\mu_1)^2} \\
&\overset{(b)}{\geq} \ln\left(\frac{1}{2.4\delta}\right) \cdot \frac{4\underline{a}(1/4-\bar{\sigma}^2)}{(\Delta_1^{\mathrm{v}})^2} + \frac{1-\mu_1}{2}\ln\left(\frac{1}{2.4\delta}\right) \cdot \sum_{i=1}^{N} \frac{1}{\Delta_i^2} \\
&\geq \ln\left(\frac{1}{2.4\delta}\right) \cdot \frac{\min\{4\underline{a}(1/4-\bar{\sigma}^2), (1-\mu_1)/8\}}{\max\{(\Delta_1^{\mathrm{v}})^2, (\Delta_1/2)^2\}} + \frac{1-\mu_1}{8}\ln\left(\frac{1}{2.4\delta}\right) \cdot \left(\sum_{i\in\mathcal{F}\cap\mathcal{S}} \frac{1}{(\Delta_i/2)^2} + \sum_{i\in\bar{\mathcal{F}}^c\cap\mathcal{S}} \frac{1}{\max\{(\Delta_i/2)^2, (\Delta_i^{\mathrm{v}})^2\}}\right) \\
&= H_{\mathrm{VA}} \ln\left(\frac{1}{2.4\delta}\right) \cdot \min\left\{4\underline{a}\left(\frac{1}{4}-\bar{\sigma}^2\right), \frac{1-\mu_1}{8}\right\}.
\end{aligned}$$

We derive (a) by applying the lower bounds on the KL divergences in (S.31), and (b) follows from (S.30) and (S.32).

Case (iii): When $\mu_i < \bar{a}$ for all $i \in [N]$, $\{i \in [N] : \mu_i = \underline{a}\} = \emptyset$, $\{i \in [N] : \mu_i < \underline{a}\} \neq \emptyset$, and arm $1 = \mathrm{argmax}\{\mu_i : \mu_i < \underline{a}\}$, we have

$$\mathbb{E}_{\mathcal{G}_0}[\tau^{(0)}] \geq \ln\left(\frac{1}{2.4\delta}\right) \cdot \left(\frac{1}{d(\mu_1, \underline{a})} + \sum_{\substack{i:\mu_i<1/2,\\ \mu_i\neq\mu_1}} \frac{1}{d(\mu_i, \mu_1)} + \sum_{i:\mu_i\geq 1/2} \frac{1}{d(\mu_i, \bar{a})}\right).$$

Similar to the analysis of Cases (i) and (ii), we have

$$\begin{aligned}
d(\mu_i, \bar{a}) &\leq \frac{(\sigma_i^2-\bar{\sigma}^2)^2}{\underline{a}(1/4-\bar{\sigma}^2)} \quad \forall\, \mu_i \geq 1/2, \\
d(\mu_i, \underline{a}) &\leq \frac{(\sigma_i^2-\bar{\sigma}^2)^2}{\underline{a}(1/4-\bar{\sigma}^2)} \quad \forall\, \mu_i < 1/2, \\
d(\mu_1, \underline{a}) &\leq \frac{(\mu_1-\underline{a})^2 \cdot (1-\mu_1-\underline{a})^2}{\underline{a}(1-\mu_1-\underline{a})^2} = \frac{(\sigma_1^2-\bar{\sigma}^2)^2}{\underline{a}(1-\mu_1-\underline{a})^2} \leq \frac{(\sigma_1^2-\bar{\sigma}^2)^2}{\underline{a}(1-2\underline{a})^2} = \frac{(\sigma_1^2-\bar{\sigma}^2)^2}{4\underline{a}(1/2-\underline{a})^2}.
\end{aligned}$$

Note that $i^\star = 1$ and $\mu_{i^\star} = \mu_1 < \underline{a}$. For $\mu_i < \mu_1$, arm i is feasible and we also have

$$d(\mu_i, \underline{a}) \leq \frac{(\mu_i-\mu_1)^2}{\underline{a}} = \frac{(\mu_i-\mu_{i^\star})^2}{\underline{a}}.$$

Therefore,

$$\mathbb{E}[\tau] \geq \ln\left(\frac{1}{2.4\delta}\right) \cdot \frac{4\underline{a}(1/2-\underline{a})^2}{(\sigma_1^2-\bar{\sigma}^2)^2} + \underline{a}\ln\left(\frac{1}{2.4\delta}\right) \cdot \left(\sum_{i\in\mathcal{F}\cap\mathcal{S}} \frac{1}{(\mu_i-\mu_{i^\star})^2} + \sum_{i\notin\mathcal{F}} \frac{1/4-\bar{\sigma}^2}{(\sigma_i^2-\bar{\sigma}^2)^2}\right).$$

By definition, $\bar{\mathcal{F}}^c \cap \mathcal{S}$ is empty and hence $\Delta_1 = \min_{i\in\mathcal{S}} \Delta_i = \min_{i\in\mathcal{F}\cap\mathcal{S}} \Delta_i$. Combined with (S.30), the above analysis yields

$$\begin{aligned}
\mathbb{E}[\tau] &\geq \underline{a}\ln\left(\frac{1}{2.4\delta}\right) \cdot \left(\frac{4(1/4-\bar{\sigma}^2)}{(\Delta_1^{\mathrm{v}})^2} + \sum_{i\in\mathcal{F}\cap\mathcal{S}} \frac{1}{\Delta_i^2} + \sum_{i\in\bar{\mathcal{F}}^c\cap\mathcal{R}} \frac{1/4-\bar{\sigma}^2}{(\Delta_i^{\mathrm{v}})^2}\right) \\
&\geq \underline{a}\ln\left(\frac{1}{2.4\delta}\right) \cdot \left(\frac{4(1/4-\bar{\sigma}^2)}{(\Delta_1^{\mathrm{v}})^2} + \sum_{i\in\mathcal{F}} \frac{1/2}{\Delta_i^2} + \sum_{i\in\bar{\mathcal{F}}^c\cap\mathcal{R}} \frac{1/4-\bar{\sigma}^2}{(\Delta_i^{\mathrm{v}})^2}\right) \\
&\geq \underline{a}\ln\left(\frac{1}{2.4\delta}\right) \cdot \left(\frac{\min\{4(1/4-\bar{\sigma}^2), 1/8\}}{\max\{(\Delta_1^{\mathrm{v}})^2, (\Delta_1/2)^2\}}\right.
\end{aligned}$$

$$+\sum_{i\in\mathcal{F}\cap\mathcal{S}}\frac{1/8}{(\Delta_i/2)^2}+\sum_{i\in\bar{\mathcal{F}}^c\cap\mathcal{R}}\frac{1/4-\bar{\sigma}^2}{(\Delta_i^{\mathrm{v}})^2}\Bigg)$$
$$\geq H_{\mathrm{VA}}\ln\left(\frac{1}{2.4\delta}\right)\cdot\underline{a}\cdot\min\left\{\frac{1}{8},\frac{1}{4}-\bar{\sigma}^2\right\}.$$

Step 4 (Conclusion):

Case (i): When $\underline{a}<\mu_N\leq\ldots\leq\mu_2\leq\mu_1<\bar{a}$, we have

$$\mathbb{E}[\tau]\geq H_{\mathrm{VA}}\ln\left(\frac{1}{2.4\delta}\right)\cdot\underline{a}(1/4-\bar{\sigma}^2).$$

Case (ii): When $0<\mu_N\leq\ldots\leq\mu_2<\mu_1<1$, and $\bar{a}<\mu_1$, we have

$$\mathbb{E}[\tau]\geq H_{\mathrm{VA}}\ln\left(\frac{1}{2.4\delta}\right)\cdot\min\left\{4\underline{a}\left(\frac{1}{4}-\bar{\sigma}^2\right),\frac{1-\mu_1}{8}\right\}.$$

Case (iii): When $\mu_i<\bar{a}$ for all $i\in[N]$, $\{i\in[N]:\mu_i=\underline{a}\}=\emptyset$, $\{i\in[N]:\mu_i<\underline{a}\}\neq\emptyset$, and arm $1=\mathrm{argmax}\{\mu_i:\mu_i<\underline{a}\}$, we have

$$\mathbb{E}[\tau]\geq H_{\mathrm{VA}}\ln\left(\frac{1}{2.4\delta}\right)\cdot\underline{a}\cdot\min\left\{\frac{1}{8},\frac{1}{4}-\bar{\sigma}^2\right\}.$$

In either case, we have

$$\mathbb{E}[\tau]\geq H_{\mathrm{VA}}\ln\left(\frac{1}{2.4\delta}\right)\cdot\min\left\{\underline{a}\left(\frac{1}{4}-\bar{\sigma}^2\right),\ \frac{\underline{a}}{8},\ \frac{1-\mu_{i^\star}}{8}\right\},$$

which completes the proof of the lower bound. □

Proof of Corollary 1: The only statement that requires proof is the fact that the average sample complexity of VA-LUCB $\mathbb{E}[\tau^{\text{VA-LUCB}}]$ is $O\big(H_{\mathrm{VA}}\ln(H_{\mathrm{VA}}/\delta)\big)$. From Lemma 5, the expected time steps can be upper bounded by[5]

$$t^\star+\sum_{t=t^\star+1}^{\infty}\frac{5\delta}{t^2}\leq t^\star+\frac{5\delta}{t^\star}\leq 152\,H_{\mathrm{VA}}\ln\frac{H_{\mathrm{VA}}}{\delta}+1. \tag{S.33}$$

Note the average sample complexity is at most twice the number of time steps, thus

$$\tau_\delta^\star\leq\mathbb{E}[\tau^{\text{VA-LUCB}}]\leq 304\,H_{\mathrm{VA}}\ln\frac{H_{\mathrm{VA}}}{\delta}+2$$
$$=O\left(H_{\mathrm{VA}}\ln\frac{H_{\mathrm{VA}}}{\delta}\right),$$

which completes the proof. □

APPENDIX H
EXPERIMENTAL DETAILS

A. Experiment Design for the Fourth Term of H_{VA}

We complete the description of the experiment design in Section V, i.e., for the fourth term $\max\{\Delta_i/2,\Delta_i^{\mathrm{v}}\}^{-2}$,

(a). Under the condition that $\Delta_i/2\leq\Delta_i^{\mathrm{v}}$, when $\Delta_{i^\star}(\geq 2\Delta_{i^\star}^{\mathrm{v}})$ and Δ_i for all $i\in\bar{\mathcal{F}}^c\cap\mathcal{S}$ increase, H_{VA} and the sample complexity will stay the same.

(b). Under the condition that $\Delta_i/2\leq\Delta_i^{\mathrm{v}}$,when Δ_i^{v} increases, H_{VA} and the sample complexity will decrease.

(c). Under the condition that $\Delta_i/2\geq\Delta_i^{\mathrm{v}}$, as $\Delta_{i^\star}(\leq 2\Delta_{i^\star}^{\mathrm{v}})$ and Δ_i for all $i\in\bar{\mathcal{F}}^c\cap\mathcal{S}$ increase, H_{VA} and the sample complexity will decrease.

(d). Under the condition that $\Delta_i/2\geq\Delta_i^{\mathrm{v}}$, as Δ_i^{v} increase, H_{VA} and the sample complexity will stay unchanged.

B. Specific Parameters for Each Instance

There are 4 cases for the first and fourth term in H_{VA} respectively, as well as one case for the second and third term respectively. In each case, there are 11 instances, indexed by $j\in\{0,1,2\ldots,10\}$. Each instance consists of $N=20$ arms, including $i^\star$ (if it exists), $i^{\star\star}$ (if it exists), and the other 18 arms with exactly the same parameters. Beta distribution are adopted as the reward distributions for the arms because they are supported on $[0,1]$ and due to their flexibility in assigning the expectations and the variances for the arms. To be more specific, given a Beta distribution $B(\alpha,\beta)$ with expectation a and variance b, where $\alpha,\beta>0, a(1-a)>b$, the four parameters are related according to the following equations:

$$a=\frac{\alpha}{\alpha+\beta},\qquad b=\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$
$$\alpha=\frac{a^2(1-a)-ab}{b},\quad \beta=\frac{1-a}{a}\alpha=\frac{a(1-a)^2-(1-a)b}{b}.$$

Thus, when the expectation and the variance are given, the two parameters of the Beta distribution α and β can be readily computed. To demonstrate the effects on the sample complexity of the four terms more clearly, each instance is designed to consist of the arms which are associated with the term to be examined. The parameters for each instance indexed by $j\in\{0,1\ldots,10\}$ in each case are described below. Recall the definition of H_{VA} in (11). Since the first term $\min\{\Delta_{i^\star}/2,\Delta_{i^\star}^{\mathrm{v}}\}^{-2}$ involves arm $i^\star$, arm $i^{\star\star}$, Case 1 is comprised of the best feasible arm $i^\star$ and feasible and subptimal arms (including $i^{\star\star}$, i.e., $H_{\mathrm{VA}}=\min\{\Delta_{i^\star}/2,\Delta_{i^\star}^{\mathrm{v}}\}^{-2}+\sum_{i\in\mathcal{F}\cap\mathcal{S}}(\Delta_i/2)^{-2}$. We control the mean gap of $i^\star$ by changing the mean of $i^{\star\star}$ and the variance gap by changing the variance of $i^\star$.

Case 1(a): $\Delta_{i^\star}/2\leq\Delta_{i^\star}^{\mathrm{v}}$ and

$$H_{\mathrm{VA}}=\frac{1}{\min\{\frac{\Delta_{i^\star}}{2},\Delta_{i^\star}^{\mathrm{v}}\}^2}+\left(\frac{\Delta_{i^{\star\star}}}{2}\right)^{-2}+\sum_{i\in\mathcal{F}\cap\mathcal{S}\setminus\{i^{\star\star}\}}\left(\frac{\Delta_i}{2}\right)^{-2}$$
$$=\left(\frac{\Delta_{i^\star}}{2}\right)^{-2}+\left(\frac{\Delta_{i^{\star\star}}}{2}\right)^{-2}+\sum_{i\in\mathcal{F}\cap\mathcal{S}\setminus\{i^{\star\star}\}}\left(\frac{\Delta_i}{2}\right)^{-2}$$

The parameters that are varied are $\Delta_{i^\star}$ and $\Delta_{i^{\star\star}}$. See Table II for details.

Case 1(b): $\Delta_{i^\star}/2\leq\Delta_{i^\star}^{\mathrm{v}}$ and

$$H_{\mathrm{VA}}=\frac{1}{\min\{\frac{\Delta_{i^\star}}{2},\Delta_{i^\star}^{\mathrm{v}}\}^2}+\left(\frac{\Delta_{i^{\star\star}}}{2}\right)^{-2}+\sum_{i\in\mathcal{F}\cap\mathcal{S}\setminus\{i^{\star\star}\}}\left(\frac{\Delta_i}{2}\right)^{-2}$$
$$=\left(\frac{\Delta_{i^\star}}{2}\right)^{-2}+\left(\frac{\Delta_{i^{\star\star}}}{2}\right)^{-2}+\sum_{i\in\mathcal{F}\cap\mathcal{S}\setminus\{i^{\star\star}\}}\left(\frac{\Delta_i}{2}\right)^{-2}$$

[5]The power of t in the summation is 2. The number 2 can be traced back to choice/design of the power of t (which is 4) in the confidence radii α and β in (3). In fact, the power 4 in (3) can be replaced by any number slightly greater than 3 so that the infinite summation in (S.33) still converges. By doing so, the sample complexity of the upper bound remains unchanged, but the empirical performance will be improved. One can replace the 4 with 2 and the empirical performance will be improved significantly. However, this comes at the expense of the loss of tightness in the expected sample complexity result (cf. Corollary 1).

TABLE II
CASE 1(a)

$\bar{\sigma}^2 = 0.25$					
$\mu_{i^\star}$	0.7	$\mu_{i^{\star\star}}$	$\mu_{i^\star} - \Delta_{i^\star}$	μ_i	0.2
$\Delta_{i^\star}$	0.01×1.2^j	$\Delta_{i^{\star\star}}$	$\Delta_{i^\star}$	Δ_i	0.5
$\sigma^2_{i^\star}$	0.09	$\sigma^2_{i^{\star\star}}$	0.09	σ^2_i	0.09
$\Delta^{\mathrm{v}}_{i^\star}$	0.16	$\Delta^{\mathrm{v}}_{i^{\star\star}}$	0.16	Δ^{v}_i	0.16

TABLE III
CASE 1(b)

$\bar{\sigma}^2 = 0.25, N = 20$					
$\mu_{i^\star}$	0.55	$\mu_{i^{\star\star}}$	0.53	μ_i	0.15
$\Delta_{i^\star}$	0.02	$\Delta_{i^{\star\star}}$	0.02	Δ_i	0.4
$\sigma^2_{i^\star}$	$\bar{\sigma}^2 - \Delta^{\mathrm{v}}_{i^\star}$	$\sigma^2_{i^{\star\star}}$	0.09	σ^2_i	0.09
$\Delta^{\mathrm{v}}_{i^\star}$	0.01×1.2^j	$\Delta^{\mathrm{v}}_{i^{\star\star}}$	0.16	Δ^{v}_i	0.16

TABLE IV
CASE 1(c)

$\bar{\sigma}^2 = 0.25, N = 20$					
$\mu_{i^\star}$	0.55	$\mu_{i^{\star\star}}$	0.15	μ_i	0.15
$\Delta_{i^\star}$	0.4	$\Delta_{i^{\star\star}}$	0.4	Δ_i	0.4
$\sigma^2_{i^\star}$	$\bar{\sigma}^2 - \Delta^{\mathrm{v}}_{i^\star}$	$\sigma^2_{i^{\star\star}}$	0.09	σ^2_i	0.09
$\Delta^{\mathrm{v}}_{i^\star}$	0.01×1.2^j	$\Delta^{\mathrm{v}}_{i^{\star\star}}$	0.16	Δ^{v}_i	0.16

TABLE V
CASE 1(d)

$\bar{\sigma}^2 = 0.04$					
$\mu_{i^\star}$	0.7	$\mu_{i^{\star\star}}$	$\mu_{i^\star} - \Delta_{i^\star}$	μ_i	0.3
$\Delta_{i^\star}$	0.02×1.1^j	$\Delta_{i^{\star\star}}$	$\Delta_{i^\star}$	Δ_i	0.4
$\sigma^2_{i^\star}$	0.03	$\sigma^2_{i^{\star\star}}$	0.03	σ^2_i	0.03
$\Delta^{\mathrm{v}}_{i^\star}$	0.01	$\Delta^{\mathrm{v}}_{i^{\star\star}}$	0.01	Δ^{v}_i	0.01

The parameter that is varied is $\Delta^{\mathrm{v}}_{i^\star}$. See Table III for details.

Case 1(c): $\Delta_{i^\star}/2 \geq \Delta^{\mathrm{v}}_{i^\star}$ and

$$H_{\mathrm{VA}} = \frac{1}{\min\{\frac{\Delta_{i^\star}}{2}, \Delta^{\mathrm{v}}_{i^\star}\}^2} + \left(\frac{\Delta_{i^{\star\star}}}{2}\right)^{-2} + \sum_{i \in \mathcal{F}\cap\mathcal{S}\setminus\{i^{\star\star}\}} \left(\frac{\Delta_i}{2}\right)^{-2}$$
$$= (\Delta^{\mathrm{v}}_{i^\star})^{-2} + \left(\frac{\Delta_{i^{\star\star}}}{2}\right)^{-2} + \sum_{i \in \mathcal{F}\cap\mathcal{S}\setminus\{i^{\star\star}\}} \left(\frac{\Delta_i}{2}\right)^{-2}$$

See Table V for details.

Case 1(d): $\Delta_{i^\star}/2 \geq \Delta^{\mathrm{v}}_{i^\star}$ and

$$H_{\mathrm{VA}} = \frac{1}{\min\{\frac{\Delta_{i^\star}}{2}, \Delta^{\mathrm{v}}_{i^\star}\}^2} + \left(\frac{\Delta_{i^{\star\star}}}{2}\right)^{-2} + \sum_{i \in \mathcal{F}\cap\mathcal{S}\setminus\{i^{\star\star}\}} \left(\frac{\Delta_i}{2}\right)^{-2}$$
$$= (\Delta^{\mathrm{v}}_{i^\star})^{-2} + \left(\frac{\Delta_{i^{\star\star}}}{2}\right)^{-2} + \sum_{i \in \mathcal{F}\cap\mathcal{S}\setminus\{i^{\star\star}\}} \left(\frac{\Delta_i}{2}\right)^{-2}$$

The parameters that are varied are $\Delta_{i^\star}$ and $\Delta_{i^{\star\star}}$. See Table V for details.

Case 2: To see the effect of the **second term** $\sum_{i\in\mathcal{F}\cap\mathcal{S}}(\Delta_i/2)^{-2}$, Case 2 is compriesed of the best feasible arm and arms in $\mathcal{F}\cap\mathcal{S}$, including $i^{\star\star}$. We set $\Delta_{i^\star}/2 \leq \Delta^{\mathrm{v}}_{i^\star}$ and $\bar{\mathcal{F}}^c = \emptyset$. Therefore,

$$H_{\mathrm{VA}} = \frac{1}{\min\{\frac{\Delta_{i^\star}}{2}, \Delta^{\mathrm{v}}_{i^\star}\}^2} + \left(\frac{\Delta_{i^{\star\star}}}{2}\right)^{-2} + \sum_{i \in \mathcal{F}\cap\mathcal{S}\setminus\{i^{\star\star}\}} \left(\frac{\Delta_i}{2}\right)^{-2}$$
$$= \left(\frac{\Delta_{i^\star}}{2}\right)^{-2} + \sum_{i \in \mathcal{F}\cap\mathcal{S}} \left(\frac{\Delta_i}{2}\right)^{-2}$$

See Table VI for details.

Case 3: As for the **third term** $\sum_{i\in\bar{\mathcal{F}}^c\cap\mathcal{R}}(\Delta^{\mathrm{v}}_i)^{-2}$, $\bar{\mathcal{F}}^c\cap\mathcal{R}$ is nonempty. In Case 3, we set $\mathcal{F} = \emptyset$, i.e. it is an infeasible instance and there are 20 infeasible arms with the same parameters. Hence, $H_{\mathrm{VA}} = \sum_{i\in\bar{\mathcal{F}}^c\cap\mathcal{R}}(\Delta^{\mathrm{v}}_i)^{-2}$. See Table VII for details.

For the **fourth term** $\sum_{i\in\bar{\mathcal{F}}^c\cap\mathcal{S}} \max\{\Delta_i/2, \Delta^{\mathrm{v}}_i\}^{-2}$, the arms that are involved are $i^\star$ and the both infeasible and suboptimal arms. In Case 4, $i^\star$ is designed to be the unique feasible arm and the rest of the arms are set to be infeasible and suboptimal arms with the same parameters, i.e., $H_{\mathrm{VA}} = \min\{\Delta_{i^\star}/2, \Delta^{\mathrm{v}}_{i^\star}\}^{-2} + \sum_{i\in\bar{\mathcal{F}}^c\cap\mathcal{S}} \max\{\Delta_i/2, \Delta^{\mathrm{v}}_i\}^{-2}$. In Case 4(c) we set $\Delta_{i^\star}/2 \leq \Delta^{\mathrm{v}}_{i^\star}$ while in other cases $\Delta_{i^\star}/2 \geq \Delta^{\mathrm{v}}_{i^\star}$.

Case 4(a): $\Delta_i/2 \leq \Delta^{\mathrm{v}}_i$ and

$$H_{\mathrm{VA}} = \frac{1}{\min\{\frac{\Delta_{i^\star}}{2}, \Delta^{\mathrm{v}}_{i^\star}\}^2} + \sum_{i\in\bar{\mathcal{F}}^c\cap\mathcal{S}} \frac{1}{\max\{\frac{\Delta_i}{2}, \Delta^{\mathrm{v}}_i\}^2}$$
$$= (\Delta^{\mathrm{v}}_{i^\star})^{-2} + \sum_{i\in\bar{\mathcal{F}}^c\cap\mathcal{S}} (\Delta^{\mathrm{v}}_i)^{-2}.$$

The parameters that are varied are Δ_i for all $i \in [N]$. See Table VIII for details.

Case 4(b): $\Delta_i/2 \leq \Delta^{\mathrm{v}}_i$ and

$$H_{\mathrm{VA}} = \frac{1}{\min\{\frac{\Delta_{i^\star}}{2}, \Delta^{\mathrm{v}}_{i^\star}\}^2} + \sum_{i\in\bar{\mathcal{F}}^c\cap\mathcal{S}} \frac{1}{\max\{\frac{\Delta_i}{2}, \Delta^{\mathrm{v}}_i\}^2}$$
$$= (\Delta^{\mathrm{v}}_{i^\star})^{-2} + \sum_{i\in\bar{\mathcal{F}}^c\cap\mathcal{S}} (\Delta^{\mathrm{v}}_i)^{-2}.$$

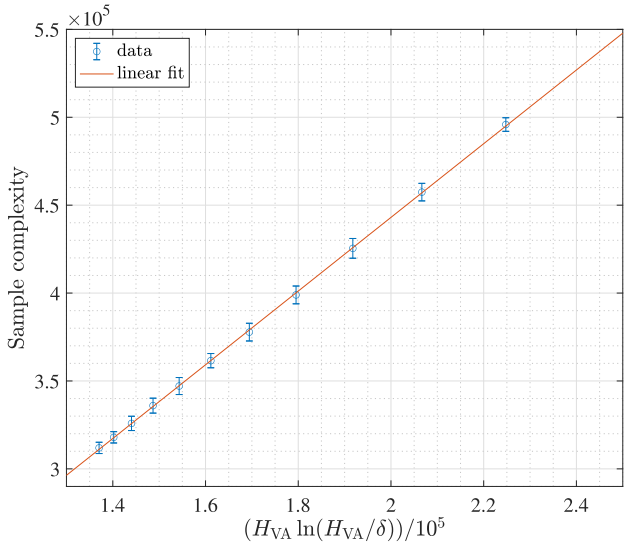
The parameters that are varied are Δ^{v}_i for all $i \in \bar{\mathcal{F}}^c\cap\mathcal{S}$. See Table IX for details.

TABLE VI
CASE 2

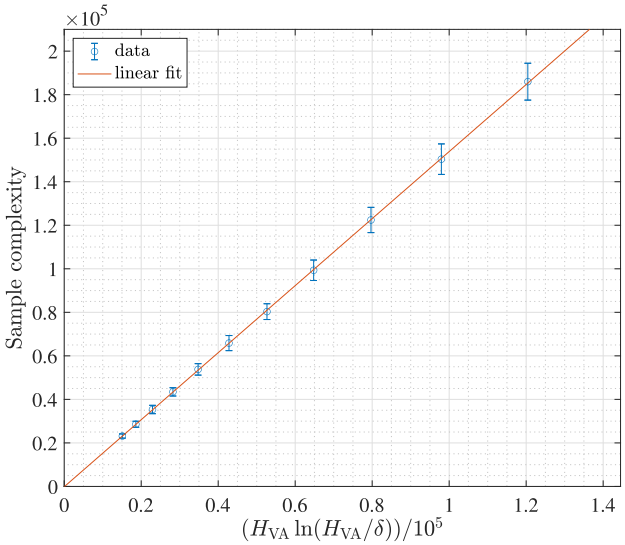
$\bar{\sigma}^2 = 0.25, N = 20$					
$\mu_{i^\star}$	0.7	$\mu_{i^{\star\star}}$	$\mu_{i^\star} - \Delta_{i^\star}$	μ_i	$\mu_{i^\star} - \Delta_i$
$\Delta_{i^\star}$	Δ_i	$\Delta_{i^{\star\star}}$	$\Delta_{i^\star}$	Δ_i	0.02×1.2^j
$\sigma^2_{i^\star}$	0.09	$\sigma^2_{i^{\star\star}}$	0.09	σ^2_i	0.09
$\Delta^{\mathrm{v}}_{i^\star}$	0.16	$\Delta^{\mathrm{v}}_{i^{\star\star}}$	0.16	Δ^{v}_i	0.16

TABLE VII
CASE 3

$\bar{\sigma}^2 = 0.04, N = 20$					
$\mu_{i^\star}$	NA	$\mu_{i^{\star\star}}$	NA	μ_i	0.55
$\Delta_{i^\star}$	$+\infty$	$\Delta_{i^{\star\star}}$	0	Δ_i	0
$\sigma^2_{i^\star}$	NA	$\sigma^2_{i^{\star\star}}$	NA	σ^2_i	$\bar{\sigma}^2 + \Delta^{\mathrm{v}}_i$
$\Delta^{\mathrm{v}}_{i^\star}$	NA	$\Delta^{\mathrm{v}}_{i^{\star\star}}$	NA	Δ^{v}_i	0.01×1.2^j



(a) Case 4(b)



(b) Case 4(c)

Fig. 8. The time complexities of cases 4(b) and 4(c) with respect to $H_{\mathrm{VA}} \ln(H_{\mathrm{VA}}/\delta)$.

TABLE VIII
CASE 4(a)

$\bar{\sigma}^2 = 0.04, N = 20$					
$\mu_{i^\star}$	0.7	$\mu_{i^{\star\star}}$	$\mu_{i^\star} - \Delta_{i^\star}$	μ_i	$\mu_{i^\star} - \Delta_i$
$\Delta_{i^\star}$	Δ_i	$\Delta_{i^{\star\star}}$	$\Delta_{i^\star}$	Δ_i	0.02×1.2^j
$\sigma^2_{i^\star}$	0.03	$\sigma^2_{i^{\star\star}}$	0.2	σ^2_i	0.2
$\Delta^{\mathrm{v}}_{i^\star}$	0.01	$\Delta^{\mathrm{v}}_{i^{\star\star}}$	0.16	Δ^{v}_i	0.16

TABLE IX
CASE 4(b)

$\bar{\sigma}^2 = 0.04, N = 20$					
$\mu_{i^\star}$	0.55	$\mu_{i^{\star\star}}$	0.53	μ_i	0.53
$\Delta_{i^\star}$	0.02	$\Delta_{i^{\star\star}}$	0.02	Δ_i	0.02
$\sigma^2_{i^\star}$	0.03	$\sigma^2_{i^{\star\star}}$	σ^2_i	σ^2_i	$\bar{\sigma}^2 + \Delta^{\mathrm{v}}_i$
$\Delta^{\mathrm{v}}_{i^\star}$	0.01	$\Delta^{\mathrm{v}}_{i^{\star\star}}$	Δ^{v}_i	Δ^{v}_i	0.05×1.1^j

Case 4(c): $\Delta_i/2 \geq \Delta^{\mathrm{v}}_i$ and

$$H_{\mathrm{VA}} = \frac{1}{\min\{\frac{\Delta_{i^\star}}{2}, \Delta^{\mathrm{v}}_{i^\star}\}^2} + \sum_{i \in \bar{\mathcal{F}}^c \cap \mathcal{S}} \frac{1}{\max\{\frac{\Delta_i}{2}, \Delta^{\mathrm{v}}_i\}^2} = \left(\frac{\Delta_{i^\star}}{2}\right)^{-2} + \sum_{i \in \bar{\mathcal{F}}^c \cap \mathcal{S}} \left(\frac{\Delta_i}{2}\right)^{-2}.$$

The parameters that are varied are Δ_i for all $i \in [N]$. See Table X for details.

Case 4(d): $\Delta_i/2 \geq \Delta^{\mathrm{v}}_i$ and

$$H_{\mathrm{VA}} = \frac{1}{\min\{\frac{\Delta_{i^\star}}{2}, \Delta^{\mathrm{v}}_{i^\star}\}^2} + \sum_{i \in \bar{\mathcal{F}}^c \cap \mathcal{S}} \frac{1}{\max\{\frac{\Delta_i}{2}, \Delta^{\mathrm{v}}_i\}^2} = (\Delta^{\mathrm{v}}_{i^\star})^{-2} + \sum_{i \in \bar{\mathcal{F}}^c \cap \mathcal{S}} \left(\frac{\Delta_i}{2}\right)^{-2}.$$

TABLE X
CASE 4(c)

$\bar{\sigma}^2 = 0.2, N = 20$					
$\mu_{i^\star}$	0.7	$\mu_{i^{\star\star}}$	$\mu_{i^\star} - \Delta_{i^\star}$	μ_i	$\mu_{i^\star} - \Delta_i$
$\Delta_{i^\star}$	Δ_i	$\Delta_{i^{\star\star}}$	$\Delta_{i^\star}$	Δ_i	0.09×1.1^j
$\sigma^2_{i^\star}$	0.04	$\sigma^2_{i^{\star\star}}$	0.21	σ^2_i	0.21
$\Delta^{\mathrm{v}}_{i^\star}$	0.16	$\Delta^{\mathrm{v}}_{i^{\star\star}}$	0.01	Δ^{v}_i	0.01

TABLE XI
CASE 4(d)

$\bar{\sigma}^2 = 0.04, N = 20$					
$\mu_{i^\star}$	0.7	$\mu_{i^{\star\star}}$	0.3	μ_i	0.3
$\Delta_{i^\star}$	0.4	$\Delta_{i^{\star\star}}$	0.4	Δ_i	0.4
$\sigma^2_{i^\star}$	0.03	$\sigma^2_{i^{\star\star}}$	σ^2_i	σ^2_i	$\bar{\sigma}^2 + \Delta^{\mathrm{v}}_i$
$\Delta^{\mathrm{v}}_{i^\star}$	0.01	$\Delta^{\mathrm{v}}_{i^{\star\star}}$	Δ^{v}_i	Δ^{v}_i	0.01×1.2^j

The parameters that are varied are Δ^{v}_i for all $i \in \bar{\mathcal{F}}^c \cap \mathcal{S}$. See Table XI for details.

C. Additional Experimental Results for VA-LUCB

The plots of the time complexities of Case 4(b) and Case 4(c) with respect to the corresponding $H_{\mathrm{VA}} \ln(H_{\mathrm{VA}}/\delta)$ are shown in Figure 8. For Cases 1(b), 4(a) and 4(d), the time complexities in each of these cases are expected to remain the same as the instances (and hence, hardness) vary. This is corroborated by the experimental results which are displayed in Table XII.

We remark that some of the terms are correlated, e.g., $\Delta_{i^\star}$ and $\Delta_{i^{\star\star}}$. Hence, we sometimes have to make a compromise by changing the parameters of arms in other terms, like in Case 1(d) in which when $\Delta_{i^\star}$ increases, $\Delta_{i^{\star\star}}$ also increases. Thus, the decrease in sample complexity in this case results from $\Delta^{-2}_{i^{\star\star}}$ and not from $\min\{\Delta_{i^\star}/2, \Delta^{\mathrm{v}}_{i^\star}\}^{-2}$.

TABLE XII

THE TIME COMPLEXITIES OF CASES 1(B), 4(A) AND 4(D). "TC" AND "STD" ARE SHORT FOR "TIME COMPLEXITY" AND "STANDARD DEVIATION" RESPECTIVELY. THE TIME COMPLEXITIES ARE ALMOST CONSTANT ACROSS INSTANCES

	Case 1(b)		Case 4(a)		Case 4(d)	
instance	TC	STD	TC	STD	TC	STD
0	5.648×10^5	4.185×10^4	3.026×10^5	4.770×10^3	2.895×10^5	4.303×10^3
1	5.585×10^5	6.713×10^4	3.028×10^5	4.774×10^3	2.896×10^5	4.336×10^3
2	5.531×10^5	6.741×10^4	3.031×10^5	4.861×10^3	2.896×10^5	4.173×10^3
3	5.666×10^5	8.515×10^4	3.029×10^5	4.426×10^3	2.892×10^5	4.209×10^3
4	5.787×10^5	7.152×10^4	3.029×10^5	4.932×10^3	2.896×10^5	4.019×10^3
5	5.733×10^5	7.558×10^4	3.030×10^5	5.125×10^3	2.894×10^5	4.224×10^3
6	5.578×10^5	7.112×10^4	3.031×10^5	4.887×10^3	2.894×10^5	4.471×10^3
7	5.596×10^5	4.398×10^4	3.031×10^5	5.059×10^3	2.897×10^5	4.134×10^3
8	5.630×10^5	5.354×10^4	3.031×10^5	5.074×10^3	2.896×10^5	4.022×10^3
9	5.726×10^5	7.683×10^4	3.031×10^5	4.891×10^3	2.895×10^5	4.157×10^3
10	5.767×10^5	4.815×10^4	3.029×10^5	4.839×10^3	2.895×10^5	4.150×10^3

ACKNOWLEDGMENT

The authors would like to sincerely thank the two anonymous reviewers for their detailed and constructive reviews that have helped to improve the quality of the present paper.

REFERENCES

[1] T. Lattimore and C. Szepesvári, *Bandit Algorithms*. Cambridge, U.K.: Cambridge Univ. Press, 2020.

[2] A. Cassel, S. Mannor, and A. Zeevi, "A general approach to multi-armed bandits under risk criteria," in *Proc. 31st Conf. Learn. Theory*, vol. 75, 2018, pp. 1295–1306.

[3] J. Lee, S. Park, and J. Shin, "Learning bounds for risk-sensitive learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 13867–13879.

[4] J. Q. L. Chang and V. Y. F. Tan, "A unifying theory of Thompson sampling for continuous risk-averse bandits," in *Proc. 36th AAAI Conf. Artif. Intell. (AAAI)*, 2022, pp. 1–8.

[5] Y. David, B. Szörényi, M. Ghavamzadeh, S. Mannor, and N. Shimkin, "PAC bandits with risk constraints," in *Proc. Int. Symp. Artif. Intell. Math. (ISAIM)*, 2018, pp. 1–9.

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

[7] J.-Y. Audibert, S. Bubeck, and R. Munos, "Best arm identification in multi-armed bandits," in *Proc. 23th Conf. Learn. Theory*, 2010, pp. 41–53.

[8] K. Jamieson, M. Malloy, R. Nowak, and S. Bubeck, "Lil'UCB: An optimal exploration algorithm for multi-armed bandits," in *Proc. 27th Conf. Learn. Theory*, vol. 35, Barcelona, Spain, 2014, pp. 423–439.

[9] S. Kalyanakrishnan, A. Tewari, P. Auer, and P. Stone, "PAC subset selection in stochastic multi-armed bandits," in *Proc. 29th Int. Conf. Mach. Learn.*, 2012, pp. 227–234.

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

[11] D. Russo, "Simple Bayesian algorithms for best arm identification," in *Proc. 29th Annu. Conf. Learn. Theory*, vol. 49, Jun. 2016, pp. 1417–1418.

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

[13] S. R. Howard, A. Ramdas, J. McAuliffe, and J. Sekhon, "Time-uniform, nonparametric, nonasymptotic confidence sequences," *Ann. Statist.*, vol. 49, no. 2, pp. 1055–1080, Apr. 2021.

[14] A. Sani, A. Lazaric, and R. Munos, "Risk-aversion in multi-armed bandits," in *Proc. 25th Int. Conf. Neural Inf. Process. Syst.* Red Hook, NY, USA: Curran Associates, 2012, pp. 3275–3283.

[15] S. Vakili and Q. Zhao, "Risk-averse multi-armed bandit problems under mean-variance measure," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 6, pp. 1093–1111, Sep. 2016.

[16] Q. Zhu and V. Y. F. Tan, "Thompson sampling algorithms for mean-variance bandits," in *Proc. 37th Int. Conf. Mach. Learn.*, 2020, pp. 11599–11608.

[17] J. Q. L. Chang, Q. Zhu, and V. Y. F. Tan, "Risk-constrained Thompson sampling for CVaR bandits," 2020, *arXiv:2011.08046*.

[18] A. Zimin, R. Ibsen-Jensen, and K. Chatterjee, "Generalized risk-aversion in stochastic multi-armed bandits," 2014, *arXiv:1405.0833*.

[19] L. A. Prashanth, K. Jagannathan, and R. Kolla, "Concentration bounds for CVaR estimation: The cases of light-tailed and heavy-tailed distributions," in *Proc. 37th Int. Conf. Mach. Learn.*, vol. 119, 2020, pp. 5577–5586.

[20] A. Kagrecha, J. Nair, and K. Jagannathan, "Distribution oblivious, risk-aware algorithms for multi-armed bandits with unbounded rewards," in *Proc. 33rd Int. Conf. Neural Inf. Process. Syst.*, vol. 32. Red Hook, NY, USA: Curran Associates, 2019, pp. 11272–11281.

[21] A. Kagrecha, J. Nair, and K. Jagannathan, "Statistically robust, risk-averse best arm identification in multi-armed bandits," *IEEE Trans. Inf. Theory*, vol. 68, no. 8, pp. 5248–5267, Aug. 2022.

[22] Y. David and N. Shimkin, "Pure exploration for max-quantile bandits," in *Machine Learning and Knowledge Discovery in Databases* (Lecture Notes in Computer Science). Cham, Switzerland: Springer, 2016, pp. 556–571.

[23] A. Kagrecha, J. Nair, and K. Jagannathan, "Constrained regret minimization for multi-criterion multi-armed bandits," 2020, *arXiv:2006.09649*.

[24] D. Baudry, R. Gautron, E. Kaufmann, and O. Maillard, "Optimal Thompson sampling strategies for support-aware CVaR bandits," in *Proc. 38th Int. Conf. Mach. Learn.*, vol. 139, Jul. 2021, pp. 716–726.

[25] E. Even-Dar, M. Kearns, and J. Wortman, "Risk-sensitive online learning," in *Proc. Int. Conf. Algorithmic Learn. Theory*. Cham, Switzerland: Springer, 2006, pp. 199–213.

[26] O.-A. Maillard, "Robust risk-averse stochastic multi-armed bandits," in *Proc. Int. Conf. Algorithmic Learn. Theory*. Cham, Switzerland: Springer, 2013, pp. 218–233.

[27] H. S. Chang, "An asymptotically optimal strategy for constrained multi-armed bandit problems," *Math. Methods Oper. Res.*, vol. 91, no. 3, pp. 545–557, Jun. 2020.

[28] Y. Wu, R. Shariff, T. Lattimore, and C. Szepesvári, "Conservative bandits," in *Proc. 33rd Int. Conf. Mach. Learn.*, vol. 48, 2016, pp. 1254–1262.

[29] S. Amani, M. Alizadeh, and C. Thrampoulidis, "Linear stochastic bandits under safety constraints," in *Proc. 33rd Int. Conf. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 9256–9266.

[30] V. Y. F. Tan, L. A. Prashanth, and K. Jagannathan, "A survey of risk-aware multi-armed bandits," in *Proc. 31st Int. Joint Conf. Artif. Intell.*, Vienna, Austria, Jul. 2022, pp. 1–11.

[31] P. Auer, C.-K. Chiang, R. Ortner, and M. Drugan, "Pareto front identification from stochastic bandit feedback," in *Proc. 19th Int. Conf. Artif. Intell. Statistics*, vol. 51, Cadiz, Spain, May 2016, pp. 939–947.

[32] E. Turgay, D. Oner, and C. Tekin, "Multi-objective contextual bandit problem with similarity information," in *Proc. 21st Int. Conf. Artif. Intell. Statist.*, vol. 84, Apr. 2018, pp. 1673–1681.
[33] M. Zuluaga, A. Krause, and M. Püschel, "ϵ-PAL: An active learning approach to the multi-objective optimization problem," *J. Mach. Learn. Res.*, vol. 17, no. 104, pp. 1–32, 2016.
[34] J. Katz-Samuels and C. Scott, "Top feasible arm identification," in *Proc. 22nd Int. Conf. Artif. Intell. Statist.*, vol. 89, Apr. 2019, pp. 1593–1601.
[35] P. Lu, C. Tao, and X. Zhang, "Variance-dependent best arm identification," in *Proc. 37th Conf. Uncertainty Artif. Intell.*, vol. 161, Jul. 2021, pp. 1120–1129.
[36] M. Faella, A. Finzi, and L. Sauro, "Rapidly finding the best arm using variance," in *Proc. 24th Eur. Conf. Artif. Intell.*, 2020, pp. 2585–2591.
[37] S. P. Bhat and L. A. Prashanth, "Concentration of risk measures: A Wasserstein distance approach," in *Proc. 33rd Int. Conf. Neural Inf. Process. Syst.*, vol. 32. Red Hook, NY, USA: Curran Associates, 2019, pp. 11762–11771.
[38] M. Simchowitz, K. Jamieson, and B. Recht, "The simulator: Understanding adaptive sampling in the moderate-confidence regime," in *Proc. Conf. Learn. Theory*, vol. 65, 2017, pp. 1794–1834.
[39] E. Tánczos, R. Nowak, and B. Mankoff, "A KL-LUCB algorithm for large-scale crowdsourcing," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30. Red Hook, NY, USA: Curran Associates, 2017, pp. 1–10.
[40] A. Garivier and E. Kaufmann, "Optimal best arm identification with fixed confidence," in *Proc. 29th Conf. Learn. Theory*, vol. 49, 2016, pp. 998–1027.
[41] C. McDiarmid, "On the method of bounded differences," *Surv. Combinatorics*, vol. 141, no. 1, pp. 148–188, 1989.
[42] J. Duchi, *Lecture Notes for Statistics 311/Electrical Engineering 377*. Stanford, CA, USA: Stanford Univ., 2016.
[43] J. Honorio and T. Jaakkola, "Tight bounds for the expected risk of linear classifiers and PAC-Bayes finite-sample guarantees," in *Proc. 17th Int. Conf. Artif. Intell. Statist. (AISTATS)*, 2014, pp. 384–392.
[44] F. Götze, H. Sambale, and A. Sinulis, "Higher order concentration for functions of weakly dependent random variables," *Electron. J. Probab.*, vol. 24, pp. 1–19, Jan. 2019.

Yunlong Hou (Graduate Student Member, IEEE) received the B.S. degree from Beijing Normal University in 2020. He is currently pursuing the Ph.D. degree with the Department of Mathematics, National University of Singapore (NUS). His research interests include machine learning, e.g., online learning.

Vincent Y. F. Tan (Senior Member, IEEE) was born in Singapore in 1981. He received the B.A. and M.Eng. degrees in electrical and information sciences from Cambridge University in 2005 and the Ph.D. degree in electrical engineering and computer science (EECS) from the Massachusetts Institute of Technology (MIT) in 2011.

He is currently an Associate Professor with the Department of Mathematics and the Department of Electrical and Computer Engineering, National University of Singapore (NUS). His research interests include network information theory, machine learning, and statistical signal processing. He is a member of the IEEE Information Theory Society Board of Governors. He received the MIT EECS Jin-Au Kong Outstanding Doctoral Thesis Prize in 2011, the NUS Young Investigator Award in 2014, the Singapore National Research Foundation (NRF) Fellowship (Class of 2018), and the NUS Young Researcher Award in 2019. He is also serving as a Senior Area Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING and an Associate Editor of Machine Learning for the IEEE TRANSACTIONS ON INFORMATION THEORY. He was also an IEEE Information Theory Society Distinguished Lecturer for 2018/9.

Zixin Zhong was born in China in 1995. She received the Ph.D. degree from the Department of Mathematics, National University of Singapore (NUS), in October 2021.

She was privileged to be supervised by Prof. Vincent Y. F. Tan and Prof. Wang Chi Cheung during her Ph.D. study, and she worked with them as a Research Fellow from June 2021 to July 2022. She is currently a Post-Doctoral Fellow at the Department of Computing Science, University of Alberta (UofA). She is supervised by Prof. Csaba Szepesvári. Her work has been presented at top machine learning (ML) conferences, including ICML and AISTATS, and also in top journals, such as the *Journal of Machine Learning Research* (JMLR) and the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS. Her research interests include reinforcement learning, online machine learning and, in particular, multi-armed bandits. She also serves as a Reviewer for several conferences and journals, including AISTATS, ICLR, ICML, NeurIPS, IEEE TRANSACTIONS ON INFORMATION THEORY, IEEE TRANSACTIONS ON SIGNAL PROCESSING, and *Transactions on Machine Learning Research* (TMLR). She was selected as a Top Reviewer of NeurIPS 2022.