Multi-Metric Adaptive Experimental Design Under a Fixed Budget with Validation

Qining Zhang
University of Michigan

Tanner Fiez
Amazon

Yi Liu
Amazon

Wenyang Liu
Amazon

Abstract

A/B tests in online experiments face statistical power challenges when testing multiple candidates simultaneously, while adaptive experimental designs (`AED`) alone fall short in inferring experiment statistics such as the average treatment effect, especially with many metrics (e.g., revenue, safety) and heterogeneous variances. This paper proposes a fixed-budget multi-metric `AED` framework with a two-phase structure: an adaptive exploration phase to identify the best treatment, and a validation phase with an A/B test to verify the treatment's quality and infer statistics. We propose `SHRVar`, which generalizes sequential halving (`SH`) with a novel relative-variance-based sampling and an elimination strategy built on reward z values. It achieves a provable error probability that decreases exponentially, where the exponent H_3 generalizes the complexity measure for `SH` and `SHVar` with homogeneous and heterogeneous variances, respectively. Numerical experiments demonstrate its performance and robustness.

1 INTRODUCTION

Randomized online experimental design, which aims to evaluate and compare the performance of different system versions, is a standard procedure to support statistically valid decisions in industrial developments, such as new webpage layouts, new services, and new software features. A/B test (or A/B/N test), where the designer assigns an experiment subject to either the control (current version) or one of the multiple treatments (new versions) with a fixed probability and then measures their response, has demonstrated empirical success (Kohavi et al., 2020) due to its easy-to-implement nature and the accuracy in inferring experimental statistics such as the average treatment effect (`ATE`) (Imbens, 2004). However, A/B tests demonstrate poor statistical power when the number of treatments scales, which becomes insufficient in modern experiments where hundreds of new treatments may be developed simultaneously from machine learning methods such as generative AIs. On the other hand, adaptive experimental design (`AED`) methods, such as best arm identification (`BAI`), where the probability of experiment subject assignment can be adaptively chosen, have received increasing popularity as an alternative to reduce the cost of experimentation. However, the adaptivity also prohibits designers from the merits of classic A/B tests, such as the accurate inference of `ATE` (Cook et al., 2023; Deep et al., 2023).

Experimentation Framework. To combine the advantages of both `AED` and A/B tests, specifically fast and efficient treatment selection in experimentation with accurate statistical inference, we propose and study an experimentation framework shown in Fig. 1. This framework consists of two phases, an exploration phase and a validation phase. In exploration, an `AED` method, such as a multi-armed bandit (`MAB`) algorithm, is employed to identify the "best" treatment, for example, a new layout of the shopping website preferred by customers. Then, in validation, the recommended treatment (new layout) and the control (old layout) are placed in an A/B test to verify their performance in multiple metrics and obtain experiment statistics such as the `ATE`. When the superior-than-control performance of the recommended treatment is validated under all metrics, the new layout will be deployed into production to replace the old website layout, and the cycle continues. Therefore, the exploration phase aims to identify a treatment most likely to pass the validation.

Challenges. Both `AED` (Jamieson and Nowak, 2014; Kaufmann et al., 2016) and A/B tests (Siroker and Koomen, 2015; Kaufmann et al., 2014) have been extensively studied, but the new two-phase experimentation framework poses new challenges. First, most real-world

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300. Copyright 2026 by the author(s).

industrial experiments consist of multiple metrics such as revenue, safety, and customer engagement. Second, the control plays a different role from all other treatments in exploration, and thus, its experimental subject assignment should also be different. Third, the validation and exploration phases are correlated, and we aim to maximize the probability of successful validation. Both the AED and A/B test literature fail to address these challenges simultaneously.

Related Work. Our paper is closely related to best arm identification under a fixed budget. Classic AED algorithms in this setting (Karnin et al., 2013; Bubeck et al., 2011; Jamieson and Nowak, 2014; Garivier and Kaufmann, 2016), including SH and SR, often focus on a single metric as a reward. While follow-up works such as the feasible arm identification (Katz-Samuels and Bagherjeiran, 2019; Katz-Samuels and Scott, 2018, 2019) or Pareto set identification (Auer et al., 2016) extend AED to multiple metrics, these works still do not distinguish between control and treatments, where both are modeled as a mere "arm". Some work (Russac et al., 2021) considers an explicit control arm with the multi-population problem, but as with most other existing works, the exploration and validation phases are studied separately. The intertwined influence across phases has not been well understood despite its practical importance. Our framework follows the general variance equalization principle commonly used both in the fixed-budget setting (Lalitha et al., 2023) and the fixed-confidence setting (Weltz et al., 2023), and generalizes SH into multi-metric problems.

Contributions. To address these limitations, we formulate a multi-metric MAB problem with both exploration and validation as the framework shown in Fig. 1, where the goal is to design a fixed-budget BAI algorithm to identify the treatment with the best chance of passing the A/B test validation of all metrics:

- We formulate the multi-metric AED problem with exploration and validation, and study both Bayesian and non-Bayesian validation to characterize the best treatment with z values.
- We propose a sequential halving algorithm called SHRVar with a novel relative-variance-based sampling and estimates of reward z values for treatment elimination. SHRVar enjoys strong theoretical guarantees, where the error probability decreases exponentially as the exploration budget increases. The exponent depends on H_3 complexity, which generalizes results of SH (Karnin et al., 2013) for classic MAB and SHVar (Lalitha et al., 2023) for MAB with heterogeneous variance.
- We conduct numerical experiments to demonstrate the superior performance of SHRVar.

2 PRELIMINARIES

In this section, we formulate the *Multi-Metric Multi-Armed Bandit* (M3AB) model shown in Fig. 1, define the best treatment, and then introduce the objective. Throughout the paper, we use $[\cdot]_+$ to denote $\max\{0, \cdot\}$, use $\propto$ to refer to "proportional to", and use $\lesssim$ to hide absolute constants in $\leq$. For a random variable $\mathbf{x}$, we use $\mathrm{Var}(\mathbf{x})$ to denote its variance.

M3AB with Heterogeneous Variance. We consider a stochastic MAB model with $A+1$ arms denoted as $\mathcal{A} = \{0, 1, \cdots, A\}$. The arm 0 is called *control*, and all other arms are called *treatments*, as shown in Fig. 1. The rewards of arms are vectors of dimension M. We call each dimension a reward *metric* and use $\boldsymbol{\nu}_a = \nu_{a,1,} \times \cdots \times \nu_{a,M}$ to denote the joint distribution of all reward metrics for arm a, where $\nu_{a,i}$ is a normal distribution for reward metric i and arm a. The developed algorithm and theoretical guarantees could be generalized to sub-Gaussian and even heavier-tailed distributions following the development in this paper. For each metric i and each arm a, the reward distribution has mean $\mu_{a,i}$ and variance $\sigma^2_{a,i}$, where $\sigma_{a,i}$ can be heterogeneous, both across metrics for a fixed arm a or across different arms. Let $\boldsymbol{\mu}_a = [\mu_{a,1}, \cdots, \mu_{a,M}]$ and $\boldsymbol{\sigma}^2_a = [\sigma^2_{a,1}, \cdots, \sigma^2_{a,M}]$. We assume the expectations $\{\boldsymbol{\mu}_a\}_{a=1}^A$ are unknown to the agent but $\{\boldsymbol{\sigma}^2_a\}_{a=1}^A$ are known. The M3AB model is a generalization of classic MAB. If $M = 1$, we recover the MAB problem with heterogeneous variance studied in (Lalitha et al., 2023), and if the variances of different arms are the same, we recover the classic MAB model with homogeneous variance (Karnin et al., 2013).

Exploration (BAI). The agent explores the treatments and control in a best-arm-identification framework with a fixed budget T. At each time $t \in [T]$, the agent pulls an arm a_t and a random reward vector $\mathbf{x}_t = [\mathrm{x}_{t,1}, \cdots, \mathrm{x}_{t,M}]$ is sampled from $\boldsymbol{\nu}_{\mathrm{a}_t}$, the joint reward distribution of the pulled arm. After observing the reward vector, the agent moves to the next time step $t+1$, and decides which arm to pull to repeat the process. At time T, the agent is expected to recommend (select) a *treatment* $\hat{a} \in \mathcal{A} \setminus \{0\}$ and proceed to validation. Ideally, $\hat{a}$ should be the treatment that has a reward better than the control in all metrics, and at the same time maximizes the probability that a significant positive result occurs if it is placed in an A/B test against control.

Validation (A/B). In fixed-budget exploration, the error probability is usually unobserved and can be large. Therefore, a validation A/B test with budget T_{v} is conducted to verify the quality of the selected treatment $\hat{a}$ and obtain detailed experiment statistics such as the average treatment effect (ATE). We consider two types of validation A/B tests, the *non-Bayesian validation* and

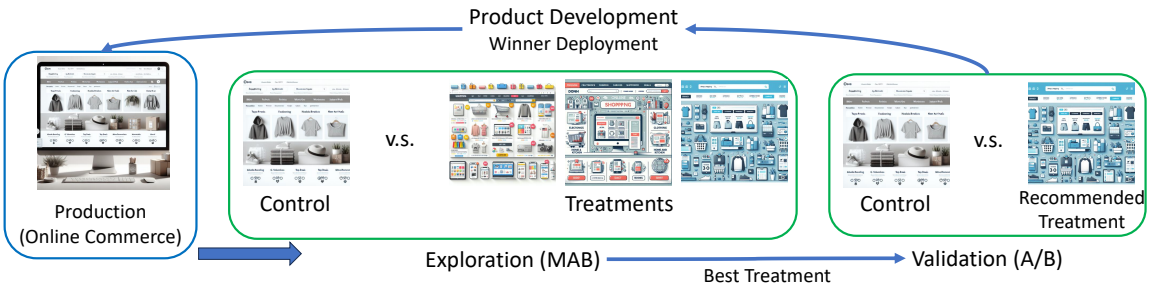


Figure 1: E-Commerce Webpage Development: the framework consists of two phases: exploration and validation. In exploration, a multi-armed bandit algorithm is used to select a treatment. In validation, the selected treatment undergoes an A/B test against the control to validate its quality.

the *Bayesian validation*, both of which have been widely used in industrial experiments (Kirk, 2009; Seltman, 2012; Box and Tiao, 2011). In both tests, the selected treatment $\hat{a}$ and the control are both pulled $T_{\mathrm{v}}/2$ times. Then, the empirical means of metric i of both arms are denoted as $\hat{\mu}_{\mathrm{v}}(\hat{a}, i)$ and $\hat{\mu}_{\mathrm{v}}(0, i)$ respectively. For each metric i, we define the null hypothesis and new hypothesis as $\mathbb{H}_{0,i} = \{\mu_{\hat{a},i} < \mu_{0,i}\}$ and $\mathbb{H}_{1,i} = \{\mu_{\hat{a},i} > \mu_{0,i}\}$. If the null hypothesis can be rejected, we say the treatment passes the validation of metric i, and denote this event as $\mathcal{E}_{\mathrm{v},i} = \{\text{reject } \mathbb{H}_{0,i} \text{ in validation}\}$. For both tests, we measure the probability $\mathbb{P}(\mathcal{E}_{\mathrm{v},i})$ that the recommended treatment passes validation.

Non-Bayesain Validation. The non-Bayesian validation is usually a z-test or a t-test (Seltman, 2012; Student, 1908). Since the variances $\sigma_{a,i}$ are known, a z-test is conducted between the control arm 0 and recommended treatment $\hat{a}$ for each metric. Let $\Phi(\cdot)$ be the cumulative distribution function of the standard normal distribution. Given a confidence level δ_i for each metric i, we reject the null hypothesis $\mathbb{H}_{0,i}$ if:

$$\hat{\mu}_{\mathrm{v}}(\hat{a}, i) - \hat{\mu}_{\mathrm{v}}(0, i) \geq \Phi^{-1}(1-\delta_i)\sqrt{2\left(\sigma^2_{\hat{a},i} + \sigma^2_{0,i}\right) T_{\mathrm{v}}^{-1}}.$$

Bayesian Validation (Kamalbasha and Eugster, 2021). We assume the `ATE` $\mu_{\hat{a},i} - \mu_{0,i}$ for metric i follows a normal prior distribution $\mathcal{N}(0, \tau_i^2)$ independent of other metrics, where τ_i^2 is the prior variance. The posterior distribution of `ATE` for the metric i is normal and calculated as $\nu_{\mathrm{v},i} = \mathcal{N}(\hat{\Delta}_{\mathrm{v},i}, \hat{\sigma}^2_{\mathrm{v},i})$ (Atsidakou et al., 2022), where $\hat{\Delta}_{\mathrm{v},i}$ and $\hat{\sigma}^2_{\mathrm{v},i}$ are the posterior mean and variances with explicit forms in the appendix section E. We obtain the posterior p_i of $\{\mu_{\hat{a},i} > \mu_{0,i}\}$ as $p_i = \int_{x=0}^{\infty} f_{\nu_{\mathrm{v},i}}(x)dx = \Phi(\hat{\Delta}_{\mathrm{v},i}/\hat{\sigma}^2_{\mathrm{v},i})$, where $f_{\nu_{\mathrm{v},i}}$ is the posterior density. We reject the null hypothesis $\mathbb{H}_{0,i}$ if p_i is larger than a threshold q_i.

Robust Treatment Identification. The goal is to find a treatment $\hat{a}$ that is most "robust" and "likely" to pass the validation for all reward metrics. Specifically, a risk-averse view is considered, and the best treatment a^* is defined as maximizing the minimum probability of rejecting the null hypothesis for all metrics:

$$a^* = \arg\max_{a \in \mathcal{A}/\{0\}} \min_i \mathbb{P}\left(\mathcal{E}_{\mathrm{v},i}\right). \tag{1}$$

The goal of this paper is to design a BAI algorithm to recommend a^*, and the performance is measured by the probability of error, i.e., $\mathbb{P}(\hat{a} \neq a^*)$.

2.1 Risk-Averse and Robust Validation

In the real world, a new version should generally not have a poorer quality than its previous version in any metric. Otherwise, it may lead to catastrophic results, e.g., loss of revenue and customer trust. Therefore, the objective of the validation phase shown in Fig. 1 is to ensure that the recommended treatment $\hat{a}$ is better in all metrics and avoid any mistakes during exploration.

Compared to a more natural angle that maximizes $\mathbb{P}(\cap_i \mathcal{E}_{\mathrm{v},i})$, the probability of rejecting all null hypotheses for all metrics simultaneously in validation, our max-min problem formulation in equation 1 takes a robust perspective that evaluates a treatment not only by its superior quality in some metrics but also the risk of misidentifying its inferiority to control in other metrics. We illustrate with a toy example in Tab. 1. This example uses Bayesian validation, and we evaluate the probability of rejecting the null hypothesis for all treatments and metrics based on the expected reward and variance. In this example, only treatment T1 has better quality than the control C in both metrics, while T2 does not satisfy the aforementioned requirement since it has a worse quality in metric 1. However, T2 has a higher overall validation passing probability $\mathbb{P}(\mathcal{E}_{\mathrm{v},1} \cap \mathcal{E}_{\mathrm{v},2})$ compared to T1, i.e., 0.30 versus 0.19. This is because the inferiority of T2 in metric 1 is difficult to identify due to the large variance and

arms	metric 1 $(\mu_{a,i}, \sigma_{a,i})$	$\mathbb{P}(\mathcal{E}_{\mathrm{v},1})$	metric 2 $(\mu_{a,i}, \sigma_{a,i})$	$\mathbb{P}(\mathcal{E}_{\mathrm{v},2})$	$\mathbb{P}(\mathcal{E}_{\mathrm{v},1} \cap \mathcal{E}_{\mathrm{v},2})$	$\min\{\mathbb{P}(\mathcal{E}_{\mathrm{v},1}), \mathbb{P}(\mathcal{E}_{\mathrm{v},2})\}$
C	$(0, 10)$		$(0, 10)$			
T1	$(0.6, 10)$	0.44	$(0.6, 10)$	0.44	0.19	**0.44**
T2	$(-0.2, 30)$	0.30	$(6, 10)$	0.99	**0.30**	0.30

Table 1: `M3AB` example: 2 treatments (T1, T2) and 1 control (C) and a typical Bayesian validation setup in industry with $q_1 = q_2 = 0.67$, prior variance $\tau_1 = \tau_2 = 10$, and $T_{\mathrm{v}} = 100$. The mean and standard deviations for both metrics are shown, as well as the probability of passing validation for each metric based on the normal posterior. In general, T1 is better than control for all metrics, and T2 is worse than control in metric 1. Therefore, T1 should be chosen for validation instead of T2.

small effect gap, and the other metric has a large probability of passing $\mathbb{P}(\mathcal{E}_{\mathrm{v},2})$ as well, which makes this treatment easy to pass validation in general. Therefore, the exploration phase should avoid recommending such treatments since they violate the development principle and may not be detected through validation. On the other hand, both metrics of T1 have a moderate probability of passing so $\mathbb{P}(\mathcal{E}_{\mathrm{v},1} \cap \mathcal{E}_{\mathrm{v},2})$ is smaller compared to T2. This shows the joint probability formulation would discourage better-than-control treatments with moderate treatment effects over metrics and mis-detect an inferior treatment if the validations for other metrics are easy to pass. On the other hand, under the max-min problem formulation which aims for the treatment with the largest $\min\{\mathbb{P}(\mathcal{E}_{\mathrm{v},1}), \mathbb{P}(\mathcal{E}_{\mathrm{v},2})\}$, we will favor T1 over T2, which aligns with the real-world interests. The equivalence of the two formulations in the large T_{v} regime is also discussed in Sec. C.2 of the appendix.

3 CHARACTERIZATION OF BEST TREATMENT

Since a^* is defined on the intangible $\mathbb{P}(\mathcal{E}_{\mathrm{v},i})$, we first analyze the validation A/B test, which gives a unified characterization for a^* in both non-Bayesian and Bayesian validations as follows:

Proposition 3.1. *In both non-Bayesian and Bayesian validation tests, the best treatment a^* defined in equation 1 is equivalently defined as follows:*

$$a^* = \arg\max_{a \in \mathcal{A}/\{0\}} \min_i \underbrace{\frac{\mu_{a,i} - \mu_{0,i}}{\sqrt{\sigma_{a,i}^2 + \sigma_{0,i}^2}} + \xi_{a,i}}_{z_{a,i}}, \tag{2}$$

*where $z_{a,i}$ is the **z value** and $\xi_{a,i}$ is the **validation constant** which we obtain from known values as:*

$$\xi_{a,i} = \frac{\Phi^{-1}(\delta_i)}{\sqrt{T_{\mathrm{v}}/2}} \quad \text{(non-Bayesian)},$$

$$\xi_{a,i} = \frac{\Phi^{-1}(1 - q_i)}{\sqrt{T_{\mathrm{v}}/2}} \sqrt{1 + 2\frac{\sigma_{a,i}^2 + \sigma_{0,i}^2}{\tau_i^2 T_{\mathrm{v}}}} \quad \text{(Bayesian)}.$$

Best Treatment a^*. The quality of each treatment is characterized by the z value, which consists of two components: the first is the signal-to-noise ratio, where the numerator represents the `ATE`, and the variance in the denominator represents the hardness of validation. The second term $\xi_{a,i}$ reflects validation requirements such as the required confidence level and the validation time horizon. The smaller $z_{a,i}$ is, the harder it will be for treatment a to pass validation of metric i. Therefore, the best treatment a^* optimizes the minimum z-value to ensure a high passing probability for all metrics. For simplicity, we define the bottleneck metric i_a^*, where the subscript is omitted when no ambiguity.

Definition 3.2 (Bottleneck Metric)**.** For each treatment a, define the bottleneck metric i_a^* as the metric with the smallest z value, i.e., $i_a^* = \arg\min_i z_{a,i}$.

Variance Heterogeneity. a^* is influenced by the heterogeneity of variances among both metrics and treatments. Let's consider $M = 1$ and omit the subscript i. Suppose we use non-Bayesian validation with confidence level δ and horizon T_{v}, then all treatments have the same validation constants $\xi_a = \Phi^{-1}(\delta)/T_{\mathrm{v}}$. If furthermore, all treatments have the same variance σ_a^2, then the best treatment a^* would be the arm with the largest reward expectation μ_a. However, if treatments have different variances σ_a^2, the arm with the largest reward expectation may no longer be a^* as it may have a large variance, resulting in a small z value. In multi-metric models, the heterogeneity of the validation constant $\xi_{a,i}$ will also influence the definition of a^*. For example, in non-Bayesian validations, some metrics are primary while others are considered as guardrails, and therefore, smaller confidence levels δ_i may be required for primary metrics but larger δ_i for guardrails. Similarly, in Bayesian validations, the heterogeneous prior variance of different metrics may also lead to heterogeneous validation constants ξ that further influence the choice of the best treatment a^*.

Algorithm 1 Sequential Halving with Relative Variance (`SHRVar`)

1: initialize $\mathcal{A}_1 \leftarrow \mathcal{A}/\{0\}$;
2: **for** $s = 1, 2, \cdots, \lceil \log_2 A \rceil$ **do**
3: for control, $N_s(0) = \left\lfloor \frac{\lambda_{\mathcal{A}_s,\Sigma}}{\rho_{\mathcal{A}_s,\Sigma}+\lambda_{\mathcal{A}_s,\Sigma}} \frac{T}{\lceil \log_2 A \rceil} \right\rfloor$.
4: for active treatments $a \in \mathcal{A}_s$,

$$N_s(a) = \left\lfloor \frac{\max_i \rho_{a,i}^2}{\rho_{\mathcal{A}_s,\Sigma}\left(\rho_{\mathcal{A}_s,\Sigma} + \lambda_{\mathcal{A}_s,\Sigma}\right)} \frac{T}{\lceil \log_2 A \rceil} \right\rfloor;$$

5: **for** $a \in \mathcal{A}_s \cup \{0\}$ **do**
6: sample a for $N_s(a)$ times and obtain reward vectors $\mathbf{x}_{a,1}, \mathbf{x}_{a,2}, \cdots, \mathbf{x}_{a,N_s(a)}$;
7: compute $\hat{\mu}_s(a,i)$ and $\hat{z}_s(a,i)$ as equation 4.
8: let $\mathcal{A}_{s+1} \subset \mathcal{A}_s$ be the set of $\lceil |\mathcal{A}_s|/2 \rceil$ arms with larger $\min_i \hat{z}_s(a,i)$;
9: recommend the arm in $\mathcal{A}_{\lceil \log_2 A \rceil + 1}$

4 SEQUENTIAL HALVING ALGORITHM

We propose an algorithm called *Sequential Halving with Relative Variance* (`SHRVar`) to tackle the proposed `M3AB` problem. The proposed algorithm is based on the sequential halving framework, and details are summarized in algorithm 1. In vanilla `SH` (Karnin et al., 2013), the total budget T is divided into $\log_2 A$ stages, and all active arms are sampled uniformly. Then, `SH` obtains the empirical reward of each arm and eliminates half of them with a lower empirical reward. Similarly, in `SHRVar`, we also divide the time horizon into $\log_2 A$ stages and sample each arm according to a designed fraction (Lines 3-4). Here, we define the *relative variance* $\rho_{a,i}^2$ and $\lambda_{a,i}^2$ which shows the proportion of reward randomness of treatment a compared to control as:

$$\rho_{a,i}^2 = \frac{\sigma_{a,i}^2}{\sigma_{a,i}^2 + \sigma_{0,i}^2}, \quad \lambda_{a,i}^2 = \frac{\sigma_{0,i}^2}{\sigma_{a,i}^2 + \sigma_{0,i}^2}. \tag{3}$$

Then, let $\rho_{\mathcal{A}_s,\Sigma}^2 = \sum_{a\in\mathcal{A}_s} \max_i \rho_{a,i}^2$ and $\lambda_{\mathcal{A}_s,\Sigma}^2 = \max_{a\in\mathcal{A}_s} \max_i \lambda_{a,i}^2$ be the overall treatment relative variance in set $\mathcal{A}_s$ and the overall control relative variance. Both can be defined for any set $\mathcal{S}$ of treatments by replacing the subscript. The intuition of the proposed relative-variance-based sampling rule will be discussed shortly. After observing the samples, `SHRVar` computes the empirical reward $\hat{\mu}_s(a,i)$ and the empirical z values with a plug-in estimator for each treatment and metric of the current stage, i.e., $\forall a \in \mathcal{A}_s, i \in [M]$,

$$\hat{\mu}_s(a,i) = \frac{1}{N_s(a)} \sum_{k=1}^{N_s(a)} \mathbf{x}_{a,i,k}, \tag{4a}$$

$$\hat{z}_s(a,i) = \frac{\hat{\mu}_s(a,i) - \hat{\mu}_s(0,i)}{\sqrt{\sigma_{a,i}^2 + \sigma_{0,i}^2}} + \xi_{a,i}. \tag{4b}$$

where $\mathbf{x}_{a,i,k}$ is the random reward of metric i and treatment a for the k-th pull. Take a minimum over metrics to estimate the z value for the bottleneck metric $\min_i z_{a,i}$, and half of the treatments with smaller $\min_i \hat{z}_{a,i}$ will be eliminated (Line 9).

Variance Equalization. The principle of sampling rules in `BAI` is to equalize the variance of the estimate used in elimination. Other rules, such as Neyman allocation (Neyman, 1992), have also been studied, and we discuss and compare them with our approach in Sec. C.1 of the appendix. If two eliminated arms have the same estimate, the probability that either is the best arm would be the same. In classic `MAB` with a single metric and homogeneous unit variance, the vanilla `SH` samples each active arm the same number of times N_s. Then, the empirical mean estimate $\hat{\mu}_s(a)$, which is used in elimination, will have the same variance for all arms, i.e., $\mathrm{Var}(\hat{\mu}_s(a)) = 1/N_s$. If heterogeneous variances are considered (Lalitha et al., 2023), uniform exploration no longer equalizes the variance of the empirical mean. Variance-based sampling is used to equalize the variance of the empirical mean estimator $\mathrm{Var}(\hat{\mu}_s(a))$, which samples each arm $N_s(a)$ times in proportion to their reward variance σ_a^2, i.e., $N_s(a) \propto \sigma_a^2$. Our problem is both multi-metric and uses the z value instead of the reward means for elimination, so we follow the principle to design the relative-variance-based sampling to handle both challenges.

Sampling Rule in Single-Metric Model. We first consider $M = 1$ and omit the index i for metrics. Since the estimate we use for treatment elimination is the z values, we analyze the variance of $\hat{z}_s(a)$:

$$\mathrm{Var}(\hat{z}_s(a)) = \frac{\rho_a^2}{N_s(a)} + \frac{\lambda_a^2}{N_s(0)} \le \frac{\rho_a^2}{N_s(a)} + \frac{\max_a \lambda_a^2}{N_s(0)}.$$

Then, we perform a min-max optimization over the variance upper bound to equalize the variance as:

$$\min_{\{N_s(a)\}} \max_{a\in\mathcal{A}_s} \left(\frac{\rho_a^2}{N_s(a)} + \frac{\max_a \lambda_a^2}{N_s(0)} \right).$$

Notice that only active treatments in $\mathcal{A}_s$ will be pulled. Solving the problem results in an allocation as:

$$N_s(0) \propto \max_{a\in\mathcal{A}_s} \lambda_a \sqrt{\sum_{a\in\mathcal{A}_s} \rho_a^2}, \; N_s(a) \propto \rho_a^2, \; \forall a \in \mathcal{A}_s,$$

which depends on relative variance and resembles the rule used in Lines 3-4 in algorithm 1 if $M = 1$.

Sampling Rule in Multi-Metric Model. To generalize the single-metric sampling rule to the multi-metric

model, we first suppose the "bottleneck" metric i_a^* is known a priori, and then we only need to consider the metric i_a^* for each treatment a and equalize its variance $\mathrm{Var}\left(\hat{z}_s(a, i_a^*)\right)$. This suffices to replace ρ_a^2 and λ_a^2 with ρ_{a,i^*}^2 and λ_{a,i^*}^2 in the sampling rule. However, i_a^* is in reality not known before the sampling rule is designed, so we pursue a pessimistic surrogate to use the maximum relative variance $\max_i \rho_{a,i}$ and $\max_i \lambda_{a,i}$ to replace ρ_{a,i^*} and λ_{a,i^*}, which results in our proposed sampling rule in algorithm 1 as:

$$N_s(0) \propto \ \rho_{\mathcal{A}_s,\Sigma} \cdot \lambda_{\mathcal{A}_s,\Sigma},\ N_s(a) \propto \ \max_i \rho_{a,i}^2,\ \forall a \in \mathcal{A}_s.$$

This rule equalizes $\max_i \mathrm{Var}\left(\hat{z}_s(a,i)\right)$ for treatments to safeguard against the worst case, where i_a^* is the metric with the largest relative variance.

Discussion on Relative Variance. The relative variance $\rho_{a,i}$ and $\lambda_{a,i}$ point out whether the difficulty of passing the validation is from the variance of the treatment or the variance of the control. If $\rho_{a,i}^2$ is large, the control will look like a "deterministic" arm since its variance is much smaller. Moreover, for any treatment a, if there exists a metric i with a large $\rho_{a,i}$, it should be allocated more samples so that its estimate is as accurate as other treatments, since passing the validation of metric i requires estimating its mean reward more accurately. To understand terms $\rho_{\mathcal{A}_s,\Sigma}$ and $\lambda_{\mathcal{A}_s,\Sigma}$, we first consider the single metric model. Then $\rho_{\mathcal{A}_s,\Sigma}^2$ becomes the sum of relative variances ρ_a^2 over all active treatments in $\mathcal{A}_s$, representing the total randomness in z value estimates from treatments, and $\lambda_{\mathcal{A}_s,\Sigma}^2$ will be the maximum of λ_a^2, indicating the variance of z value estimates from the randomness of control. Multi-metric models simply replace the relative variances with the maximum over metrics.

5 THEORETICAL RESULTS

In this section, we present the probability of error guarantee for `SHRVar`. For simplicity, we assume A is a power of 2. We first introduce some useful notations. For an arbitrary set $\mathcal{S}$ of treatments, we define the heterogeneity of relative variance $\kappa_{\mathcal{S},a,i}$ for treatment a and metric i as:

$$\kappa_{\mathcal{S},a,i} = \frac{\rho_{a,i}^2}{\max_i \rho_{a,i}^2} \frac{\rho_{\mathcal{S},\Sigma}}{\rho_{\mathcal{S},\Sigma} + \lambda_{\mathcal{S},\Sigma}} + \frac{\lambda_{a,i}^2}{\max_{a\in\mathcal{S},i} \lambda_{a,i}^2} \frac{\lambda_{\mathcal{S},\Sigma}}{\rho_{\mathcal{S},\Sigma} + \lambda_{\mathcal{S},\Sigma}}.$$

We define the effective gap $D_{\mathcal{S},a} > 0$ compared to the best treatment in terms of its z values as:

$$D_{\mathcal{S},a}^2 = \min_{i\in[M]} \max_{j\in[M]} \frac{[z_{a^*,i} - z_{a,j}]_+^2}{(\kappa_{\mathcal{S},a,j} + \kappa_{\mathcal{S},a^*,i})^2}.$$

Finally, for any set of treatments $\mathcal{S}$, let $\mathcal{S}_c'$ be its subset excluding 1/4 of the treatments with smallest $D_{\mathcal{S},a}$. The H_3 complexity is defined as follows:

$$H_3 = \left(\min_{\mathcal{S}: a^* \in \mathcal{S}} \frac{\min_{a\in\mathcal{S}_c'} D_{\mathcal{S},a}^2}{\left(\rho_{\mathcal{S},\Sigma} + \lambda_{\mathcal{S},\Sigma}\right)^2} \right)^{-1}.$$

Theorem 5.1. *If we use `SHRVar` in algorithm 1 to identify the best treatment a^* in our exploration phase, the probability of mistake can be upper bounded as:*

$$\mathbb{P}\left(\hat{a} \neq a^*\right) \leq 6M \log_2 A \cdot \exp\left(-\frac{T}{2H_3 \log_2 A}\right).$$

Proof Roadmap. The proof of the theorem is in the appendix Sec. F and builds on the sequential halving framework where we first fix a stage s and the active treatment set $\mathcal{A}_s$ which has not been eliminated, and then analyze the probability that at least half of the active treatments in $\mathcal{A}_s$ will "outperform" a^* in the elimination metric $\min_i \hat{z}_s(a,i)$. This requires analyzing the probability that each sub-optimal treatment a outperforms a^*. For each treatment a, we first characterize the probability that the minimum z value estimate is larger than a^*, i.e., $\min_i \hat{z}_s(a,i) > \min_i \hat{z}_s(a^*,i)$, which requires decoupling the dependence of different treatments when estimating the z values since they all involve the empirical reward of control. This argument leads to the instance-dependent exponent $D_{\mathcal{A}_s,a}$ analogous to the reward gap in classic bandit models. Then, we use Markov's inequality to characterize the probability of eliminating a^* at each stage and use a union bound over the stages to reach the final result. Next, we discuss the instance-dependent parameters.

Effective Gap. For any set $\mathcal{S}$, the effective gap $D_{\mathcal{S},a}$ measures the difficulty of eliminating a sub-optimal arm a from it, which is analogous to the expected reward gap in `MAB`. Since $\kappa_{\mathcal{S},a,i} \leq 1$ by definition, we have:

$$\begin{aligned} D_{\mathcal{S},a} &= \min_{i\in[M]} \max_{j\in[M]} \frac{[z_{a^*,i} - z_{a,j}]_+}{(\kappa_{\mathcal{S},a,j} + \kappa_{\mathcal{S},a^*,i})} \\ &\geq \min_{i\in[M]} \max_{j\in[M]} [z_{a^*,i} - z_{a,j}]_+ \geq z_{a^*,i^*} - z_{a,i^*}. \end{aligned}$$

Thus, $D_{\mathcal{S},a}$ can be viewed as the gap between the z value of the bottleneck metric i_a^* of two treatments. If the z values have a larger gap between a^* and a, $D_{\mathcal{S},a}$ will be larger, and the quality of the treatment will be easier to distinguish.

Relative Variance Heterogeneity. The heterogeneity $\kappa_{\mathcal{S},a,i}$ measures the error if we use the maximum relative variance $\max_i \rho_{a,i}$ to approximate each $\rho_{a,i}$ as in our sampling rule. If $\kappa_{\mathcal{S},a,i}$ is large and close to 1, the heterogeneity of $\rho_{a,i}$ is small, and approximating each with $\max_i \rho_{a,i}$ would be accurate. If $\kappa_{\mathcal{S},a,i}$ becomes

smaller, the relative variance $\rho_{a,i}$ of this metric would be much smaller than $\max_i \rho_{a,i}$, making this metric easy to estimate. Potentially, the hardness of identification would become smaller, resulting in a larger effective gap $D_{\mathcal{S},a}$.

Complexity Measure H_3. The role of the two minimum arguments in the exponent H_3^{-1} is to trade off the difficulty of distinguishing between more arms with a larger gap $D_{\mathcal{S},a}$ and fewer arms with a smaller gap $D_{\mathcal{S},a}$. If we choose a large treatment set $\mathcal{S}$, more arms with a smaller gap will be excluded from $\mathcal{S}'_{\mathrm{c}}$, which results in a larger numerator in the definition of H_3^{-1}. At the same time, more treatments are included in $\mathcal{S}$, so the denominator will also be larger since the total relative variance $\rho_{\mathcal{S},\Sigma}$ and $\lambda_{\mathcal{S},\Sigma}$ increase. Therefore, the exponent is determined by the worst set $\mathcal{S}$ with a moderate size, small gaps, and large relative variances. It is also possible that with a less refined analysis, we can obtain a looser but more explainable complexity measure H'_3 as follows:

$$H_3 \leq \left(\min_{\mathcal{S}: a^* \in \mathcal{S}} \frac{\min_{a \neq a^*} (z_{a^*,i^*} - z_{a,i^*})^2}{\left(\rho_{\mathcal{S},\Sigma} + \lambda_{\mathcal{S},\Sigma}\right)^2} \right)^{-1} \lesssim \left(\frac{\min_a (z_{a^*,i^*} - z_{a,i^*})^2}{\sum_{a \in \mathcal{A}} \max_i \rho_{a,i}^2 + \max_{a \in \mathcal{A}, i} \lambda_{a,i}^2} \right)^{-1} \equiv H'_3,$$

where $\lesssim$ hides absolute constants. The form of H'_3 resembles the complexity measure used in (Lalitha et al., 2023, Theorem 3) for the theoretical analysis of `SHVar` where the minimum reward gap is replaced by the minimum z value gap of the bottleneck metrics between the best treatment and sub-optimal treatments, and the sum of variances are replaced by the sum of maximum relative variances in the denominator. With our proof methodology, we can obtain a tighter error probability bound for `SHVar` with a complexity measure similar to H_3. We further discuss this complexity measure and achievable lower bounds in the appendix Sec C.4.

Comparation to `SH` and `SHVar`. We simplify our model to a single metric with homogeneous variance among treatments. Then, the sum of relative variances in the denominator will be approximately $|\mathcal{S}|$, and the effective gap will be the difference of z values, i.e., $D_{\mathcal{S},a} = z_{a^*} - z_a$. If we change the index of treatments to order them from large to small z-values, the complexity measure H_3 becomes:

$$H_3 = \left(\min_{\mathcal{S}: a^* \in \mathcal{S}} \frac{\min_{a \in \mathcal{S}'_{\mathrm{c}}} (z_{a^*} - z_a)^2}{|\mathcal{S}|} \right)^{-1} \lesssim \max_{a \in \mathcal{A}} \frac{a}{(z_{a^*} - z_a)^2},$$

where $\lesssim$ hides absolute constants and it almost recovers H_2 in (Karnin et al., 2013, Theorem 4.1) for `SH`, except for replacing the reward gap with the z value gap. We can also simplify our model to a single metric with heterogeneous variance and compare with (Lalitha et al., 2023). The looser exponent H'_3 is simplified as:

$$H_3 \lesssim H'_3 = \left(\frac{\min_a (z_{a^*} - z_a)^2}{\sum_{a \in \mathcal{A}} \rho_a^2 + \max_a \lambda_a^2} \right)^{-1},$$

which almost recovers (Lalitha et al., 2023, Theorem 3) for `SHVar` except for replacing the reward gap with the z-value gap and variance with relative variance. The two reductions show that `SHRVar` is a generalization of both `SH` and `SHVar` to the multi-metric with validation, and does not suffer performance loss in simpler models, which also shows H_3 provides a tighter characterization.

Algorithm 2 Sequential Halving with Relative Variance and Confidence-Based Elimination (`SHRVar-c`)

initialize $\mathcal{A}_1 \leftarrow \mathcal{A}/\{0\}$;
for $s = 1, 2, \cdots, \lceil \log_2 A \rceil$ **do**
  for control, $N_s(0) = \left\lfloor \frac{\lambda_{\mathcal{A}_s,\Sigma}}{\rho_{\mathcal{A}_s,\Sigma} + \lambda_{\mathcal{A}_s,\Sigma}} \frac{T}{\lceil \log_2 A \rceil} \right\rfloor$.
  for active treatments $a \in \mathcal{A}_s$,

$$N_s(a) = \left\lfloor \frac{\max_i \rho_{a,i}^2}{\rho_{\mathcal{A}_s,\Sigma} \left(\rho_{\mathcal{A}_s,\Sigma} + \lambda_{\mathcal{A}_s,\Sigma}\right)} \frac{T}{\lceil \log_2 A \rceil} \right\rfloor;$$

  for $a \in \mathcal{A}_s \cup \{0\}$ **do**
    sample a for $N_s(a)$ times and obtain reward vectors $\mathbf{x}_{a,1}, \mathbf{x}_{a,2}, \cdots, \mathbf{x}_{a,N_s(a)}$;
  compute $\hat{\mu}_s(a,i)$ and $\hat{z}_s(a,i)$ as equation 4.
  for $a \in \mathcal{A}_s$ **do**
    compute $\delta_s(a)$ as follows:

$$\delta_s(a) = \inf \left\{ \delta \,\middle|\, \mathrm{UCB}_s^{\delta}(a) \leq \max_{a' \in \mathcal{A}_s} \mathrm{LCB}_s^{\delta}(a') \right\};$$

  Let $\mathcal{A}_{s+1}$ be the set of $\lceil |\mathcal{A}_s|/2 \rceil$ arms with smaller $\delta_s(a)$;
recommend the arm in $\mathcal{A}_{\lceil \log_2 A \rceil + 1}$.

5.1 Discussions on the Elimination Rule

As for the elimination rule, we note that the minimum z value estimator over metrics $\min_i \hat{z}_s(a,i)$ used in `SHRVar` may be subject to underestimation since $\mathbb{E}[\min_i \hat{z}_s(a,i)] \leq \min_i z_{a,i}$, and create bias in treatment elimination. In our design of `SHRVar`, the variances of different treatments and metrics are approximately equalized through our sampling rule, which mitigates the underestimation effect. However, to study this effect, especially under the circumstances where the sampling strategy is not controlled, we consider a confidence-based elimination rule as follows. Let $b_s^{\delta}(a,i)$ be the confidence bonus when the out-of-concentration

probability is δ as follows:

$$b_s^\delta(a,i) = 2\sqrt{\left(\frac{\rho_{a,i}^2}{N_s(a)} + \frac{\lambda_{a,i}^2}{N_s(0)}\right)\log\left(\frac{|\mathcal{A}_s|M}{\delta}\right)}.$$

Let the following be the upper and lower confidence bounds where $\min_i z_{a,i}$ falls into with probability at least $1-\delta$:

$$\mathrm{LCB}_s^\delta(a) = \min_i \left\{\hat{z}_s(a,i) - b_s^\delta(a,i)\right\},$$
$$\mathrm{UCB}_s^\delta(a) = \min_i \left\{\hat{z}_s(a,i) + b_s^\delta(a,i)\right\}.$$

For each treatment a, we calculate the confidence level $\delta_s(a) = \delta$ such that $\mathrm{UCB}_s^\delta(a)$ equals the largest lower confidence bound over treatments, i.e., $\max_a \mathrm{LCB}_s^\delta(a)$. Then, the confidence-based elimination rule eliminates half of the arms with smaller $\delta_s(a)$ at each stage. The algorithm, denoted as `SHRVar-c`, is in algorithm 2 and more detailed discussions of confidence-based elimination are discussed in Sec. B of the appendix. We also provide a theoretical guarantee similar to Theorem 5.1:

Theorem 5.2. *If we use `SHRVar-c` in algorithm 2 to identify the best treatment a^* in our exploration phase, the probability of mistake can be upper bounded as:*

$$\mathbb{P}\left(\hat{a} \neq a^*\right) \leq 6M\log_2 A \cdot \exp\left(-\frac{T}{2\tilde{H}_3 \log_2 A}\right).$$

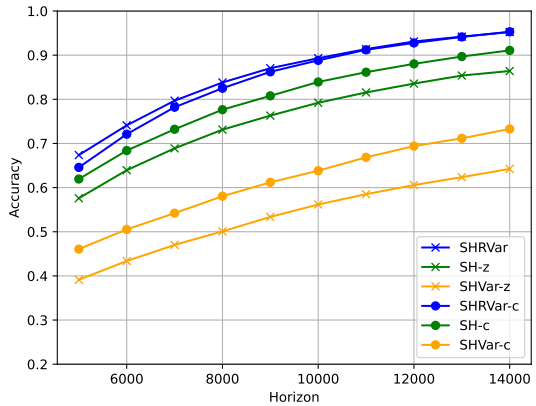
Here, $\tilde{H}_3$ is a complexity notion similar to H_3, and its explicit form and discussions are in Appendix B.

6 NUMERICAL RESULTS

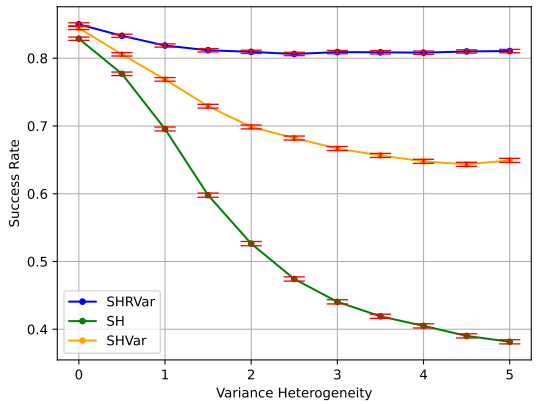
We demonstrate the empirical performance of `SHRVar`. We first study the exploration phase and compare the accuracy of identifying a^*. Then, we experiment with both the exploration and validation phases to show that `SHRVar` significantly increases the probability of validating a better treatment, especially in heterogeneous variance instances. This demonstrates the superiority of z value estimates over expected reward in `AED` with validation. The results are averaged over 10^5 repetitions. We report the error probability $\mathbb{P}(\hat{a} = a^*)$ that the algorithm successfully recommends a^* in Fig. 2(a), and the probability $\mathbb{P}(\mathcal{E}_\mathrm{v})$ that the recommended treatment $\hat{a}$ passes validation in Fig. 2(b). A larger-scale experiment is conducted in Sec. A of the appendix, where the findings are exactly the same.

6.1 Error Probability in Exploration

We consider a bandit instance with $A = 16$ treatments, and $M = 3$ metrics. Suppose we use non-Bayesian validation, and the confidence level δ_i is the same for each



(a) Exploration Accuracy



(b) Validation Success Probability

Figure 2: (a) Exploration Accuracy: rate of identifying the best treatment in exploration as T increases. (b) Validation Success Probability: rate of passing validation, i.e., successfully validating a better-than-control treatment. Error bars represent confidence intervals.

metric. So the validation constant $\xi_{a,i}$ would be the same for all (a,i). The expected reward and variance of control are set to be $[0,0,0]$ and $[1,1,1]$, respectively, for all metrics. For the treatments, the expected reward of the best treatment a^* is randomly chosen as $[2.4697, 1.5556, 1.1180]$, and all sub-optimal treatments have the same expected reward $[2.0125, 1.6971, 1.3416]$. All treatments have the same variances $[4, 1, 0.25]$.

Baselines. Since fixed-budget BAI algorithms in the literature, such as `SH` and `SHVar`, do not deal with multi-metric problems, we adapt them to `M3AB` by revising their elimination strategies to either the z-value-based or confidence-based eliminations proposed in algorithm 1 and algorithm 2. Together with `SHRVar-c`, this constructs 5 baselines where the rest are called

SH-z, SH-c, SHVar-z, and SHVar-c. Specifically, to study the influence of relative-variance-based sampling, SH-z replaces Lines 3-4 of algorithm 1 with uniform sampling and SHVar-z replaces it with a variance-based sampling rule similar to SHVar, which samples each arm proportional to $\max_i \sigma^2_{a,i}$. The comparison between the SH, SHVar, and SHRVar under the same z-value-based elimination rule reflects the efficiency of the relative-variance-based sampling rule. Additionally, SH-c and SHVar-c are created by replacing the elimination rule in SH-z and SHVar-z with the confidence-based elimination. Under each sampling rule, i.e., uniform sampling in SH, variance-based sampling in SHVar, and relative-variance-based sampling in SHRVar, we could compare the performance of the elimination rules and measure how the potential under-estimation effect of the $\min_i \hat{z}_{a,i}$ estimator would take place in different levels of variance equalization.

Performance. For all time horizons T, the relative-variance-based sampling strategy significantly improves over uniform sampling and variance-based sampling used in SHVar, with SHRVar and SHRVar-c both having a better error probability than other baselines. This is because the relative-variance-based sampling rule balances the estimated variance across treatments and avoids the error from a single uncertain estimate. Therefore, relative-variance-based sampling should be used to achieve the best efficiency. We also observe that with a small horizon, SHVar slightly performs better than SHVar-c. The difference disappears as the total horizon increases, which is consistent with our theory as H_3 and $\tilde{H}_3$ become the same in large T.

Robustness. On the other hand, if we fix the elimination rule and evaluate the performance drop from SHRVar to SHVar and SH, we can observe that under the confidence-based elimination strategy, the performance degradation is much milder, i.e., the performance difference between SHRVar-c and SH-c is much smaller than the difference between SHRVar and SH. This shows that the vanilla z-value estimate indeed suffers from the under-estimation effect. Even though it is not an issue when the variances of treatments are approximately equalized with the relative-variance-based sampling, it incurs a non-negligible performance loss when the sampling rule deviates from optimal. This demonstrates the robustness of the confidence-based elimination rule, which is useful in practice. For example, some system constraints may prevent the use of unbalanced sampling fractions for each active treatment, i.e., uniform sampling is required due to fairness. In other cases, the variances $\sigma^2_{a,i}$ also come from estimates, which may be inaccurate or subject to non-stationarity. So, a perfect z-estimate variance equalization is very difficult. More discussions are in Appendix B.

6.2 Validation Success Probability

We simplify our model to $M = 1$ and experiment with both exploration and validation phases to show that SHRVar significantly improves the chance of passing validation compared to the vanilla SH proposed in (Karnin et al., 2013) and SHVar proposed in (Lalitha et al., 2023), especially under heterogeneous variance instances. Both SH and SHVar use the estimate of expected reward for treatment elimination instead of the z value used in SHRVar. We consider $A = 27$ treatments where the control has an expected reward $\mu_0 = 0$ and variance $\sigma_0^2 = 1$. To make the results more comparable, we fix the z value of each treatment with $z_a = 0.3 - 0.1\sqrt{a}$ so the first treatment is the best. Then, we gradually increase the level of variance heterogeneity l, which is a constant in $[0, 5]$. The standard deviation σ_a of each treatment a is set as $1 + a^l$, so more sub-optimal arms have larger variance. We set the horizons for the exploration and validation phases to be 500 and 100, respectively. It can be observed that when the variance heterogeneity is small, e.g., $l \leq 0.5$, all algorithms have relatively good performance. However, as the variance heterogeneity l over the treatments increases, our proposed algorithm SHRVar maintains a good validation success rate over 80%, but the performances of both other baselines quickly decrease. This is because treatments with a larger variance will be hard to validate, which is not taken into account by the reward estimate used in SH and SHVar. The results demonstrate the superiority of z value estimation in AED with validation. Notice that SH is also worse than SHVar since it is completely variance-agnostic and thus not robust to variance heterogeneity.

7 CONCLUSION

In this paper, we consider M3AB with both exploration and validation to find the treatment with the best chance to pass an A/B test against control. We characterize the best treatment through z values and propose SHRVar with a novel relative-variance-based sampling rule. Theoretical analysis shows that SHRVar has an exponentially decreasing error probability, generalizing SH or SHVar, which also works for bounded or sub-Gaussian rewards. Numerical results also demonstrate its superiority. Our future directions include incorporating unknown variance estimations and generalizing our algorithms and results for linear bandits.

Acknowledgment

We thank Cathy Jiao for her support of this project.

References

Abbasi-Yadkori, Y., Bartlett, P., Gabillon, V., Malek, A., and Valko, M. (2018). Best of both worlds: Stochastic & adversarial best-arm identification. In *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 918–949. PMLR.

Antos, A., Grover, V., and Szepesvári, C. (2010). Active learning in heteroscedastic noise. *Theoretical Computer Science*, 411(29):2712–2728. Algorithmic Learning Theory (ALT 2008).

Ararat, C. and Tekin, C. (2023). Vector optimization with stochastic bandit feedback. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 2165–2190. PMLR.

Atsidakou, A., Katariya, S., Sanghavi, S., and Kveton, B. (2022). Bayesian fixed-budget best-arm identification. *arXiv preprint arXiv:2211.08572*.

Audibert, J.-Y. and Bubeck, S. (2010). Best Arm Identification in Multi-Armed Bandits. In *COLT - 23th Conference on Learning Theory - 2010*, page 13 p., Haifa, Israel.

Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47:235–256.

Auer, P., Chiang, C.-K., Ortner, R., and Drugan, M. (2016). Pareto front identification from stochastic bandit feedback. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 939–947, Cadiz, Spain. PMLR.

Auer, P. and Ortner, R. (2010). Ucb revisited: Improved regret bounds for the stochastic multi-armed bandit problem. *Periodica Mathematica Hungarica*, 61(1-2):55 – 65.

Bachoc, F., Cesari, T., and Gerchinovitz, S. (2021). The sample complexity of level set approximation. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 424–432. PMLR.

Badanidiyuru, A., Kleinberg, R., and Slivkins, A. (2018). Bandits with knapsacks. *J. ACM*, 65(3).

Berkenkamp, F., Krause, A., and Schoellig, A. P. (2023). Bayesian optimization with safety constraints: safe and automatic parameter tuning in robotics. *Machine learning*, 112(10):3713—3747.

Berkenkamp, F., Schoellig, A. P., and Krause, A. (2016). Safe controller optimization for quadrotors with gaussian processes. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 491–496.

Bertsimas, D., Nohadani, O., and Teo, K. M. (2010). Nonconvex robust optimization for problems with constraints. *INFORMS Journal on Computing*, 22(1):44–58.

Bogunovic, I., Scarlett, J., Krause, A., and Cevher, V. (2016). Truncated variance reduction: A unified approach to bayesian optimization and level-set estimation. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Bottero, A., Luis, C., Vinogradska, J., Berkenkamp, F., and Peters, J. R. (2022). Information-theoretic safe exploration with gaussian processes. In *Advances in Neural Information Processing Systems*, volume 35, pages 30707–30719. Curran Associates, Inc.

Box, G. E. and Tiao, G. C. (2011). *Bayesian inference in statistical analysis*. John Wiley & Sons.

Bryan, B., Nichol, R. C., Genovese, C. R., Schneider, J., Miller, C. J., and Wasserman, L. (2005). Active learning for identifying function threshold boundaries. In *Advances in Neural Information Processing Systems*, volume 18. MIT Press.

Bubeck, S., Munos, R., and Stoltz, G. (2011). Pure exploration in finitely-armed and continuous-armed bandits. *Theoretical Computer Science*, 412(19):1832–1852. Algorithmic Learning Theory (ALT 2009).

Cai, Y. and Daskalakis, C. (2011). *On Minmax Theorems for Multiplayer Games*, pages 217–234. SIAM.

Camilleri, R., Wagenmaker, A., Morgenstern, J. H., Jain, L., and Jamieson, K. G. (2022). Active learning with safety constraints. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A., editors, *Advances in Neural Information Processing Systems*, volume 35, pages 33201–33214. Curran Associates, Inc.

Carlsson, E., Basu, D., Johansson, F., and Dubhashi, D. (2024). Pure exploration in bandits with linear constraints. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 334–342. PMLR.

Chang, H. S. (2020). An asymptotically optimal strategy for constrained multi-armed bandit problems. *Mathematical Methods of Operations Research*, 91(3):545–557.

Chaudhuri, K., Jain, P., and Natarajan, N. (2017). Active heteroscedastic regression. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 694–702. PMLR.

Chen, L., Gupta, A., and Li, J. (2016). Pure exploration of multi-armed bandit under matroid constraints. In *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 647–669, Columbia University, New York, New York, USA. PMLR.

Chen, L., Li, J., and Qiao, M. (2017a). Nearly Instance Optimal Sample Complexity Bounds for Top-k Arm Selection. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 101–110. PMLR.

Chen, R. S., Lucier, B., Singer, Y., and Syrgkanis, V. (2017b). Robust optimization for non-convex objectives. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Cook, T., Mishler, A., and Ramdas, A. (2023). Semiparametric efficient inference in adaptive experiments. In *NeurIPS 2023 Workshop on Adaptive Experimental Design and Active Learning in the Real World*.

Crepon, E., Garivier, A., and Koolen, W. (2024). Sequential learning of the Pareto front for multi-objective bandits. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 3583–3591. PMLR.

David, Y., Szörényi, B., Ghavamzadeh, M., Mannor, S., and Shimkin, N. (2018). Pac bandits with risk constraints. In *ISAIM*.

Deep, V., Bassamboo, A., and Juneja, S. (2023). Asymptotically optimal adaptive a/b tests for average treatment effect. *SSRN*.

Degenne, R. and Koolen, W. (2019). Pure exploration with multiple correct answers. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Ding, W., Qin, T., Zhang, X., and Liu, T. (2013). Multi-armed bandit with budget constraint and variable costs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 27(1):232–238.

Drugan, M. and Manderick, B. (2015). Exploration versus exploitation trade-off in infinite horizon pareto multi-armed bandits algorithms. In *Proceedings of the International Conference on Agents and Artificial Intelligence - Volume 2: ICAART,*, pages 66–77. SciTePress.

Drugan, M. M. and Nowe, A. (2013). Designing multi-objective multi-armed bandits algorithms: A study. In *The 2013 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Drugan, M. M. and Nowe, A. (2014). Scalarization based pareto optimal set of arms identification algorithms. In *2014 International Joint Conference on Neural Networks (IJCNN)*, pages 2690–2697.

Duivenvoorden, R. R., Berkenkamp, F., Carion, N., Krause, A., and Schoellig, A. P. (2017). Constrained bayesian optimization with particle swarms for safe adaptive controller tuning. *IFAC-PapersOnLine*, 50(1):11800–11807. 20th IFAC World Congress.

Faizal, F. Z. and Nair, J. (2022). Constrained pure exploration multi-armed bandits with a fixed budget. *arXiv preprint arXiv:2211.14768*.

Gardner, J. R., Kusner, M. J., Xu, Z., Weinberger, K. Q., and Cunningham, J. P. (2014). Bayesian optimization with inequality constraints. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, page II–937–II–945. JMLR.org.

Garivier, A. and Kaufmann, E. (2016). Optimal best arm identification with fixed confidence. In *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 998–1027, Columbia University, New York, New York, USA. PMLR.

Garivier, A., Kaufmann, E., and Koolen, W. M. (2016). Maximin action identification: A new bandit framework for games. In *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 1028–1050, Columbia University, New York, New York, USA. PMLR.

Gelbart, M. A., Snoek, J., and Adams, R. P. (2014). Bayesian optimization with unknown constraints. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, UAI'14, page 250–259, Arlington, Virginia, USA. AUAI Press.

Glynn, P. and Juneja, S. (2004). A large deviations perspective on ordinal optimization. In *Proceedings of the 2004 Winter Simulation Conference, 2004.*, volume 1. IEEE.

Gotovos, A., Casati, N., Hitz, G., and Krause, A. (2013). Active learning for level set estimation. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, IJCAI '13, page 1344–1350. AAAI Press.

Ha, H., Gupta, S., Rana, S., and Venkatesh, S. (2021). High dimensional level set estimation with bayesian neural network. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(13):12095–12103.

Hasselt, H. (2010). Double q-learning. In *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc.

Hernández-Lobato, J. M., Gelbart, M. A., Adams, R. P., Hoffman, M. W., and Ghahramani, Z. (2016). A general framework for constrained bayesian optimization using information-based search. *Journal of Machine Learning Research*, 17(160):1–53.

Hou, Y., Tan, V. Y. F., and Zhong, Z. (2023). Almost optimal variance-constrained best arm identification. *IEEE Transactions on Information Theory*, 69(4):2603–2634.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *The Review of Economics and Statistics*, 86(1):4–29.

Inatsu, Y., Karasuyama, M., Inoue, K., and Takeuchi, I. (2019). Active learning for level set estimation under cost-dependent input uncertainty. *arXiv preprint arXiv:1909.06064*.

Iwazaki, S., Inatsu, Y., and Takeuchi, I. (2020). Bayesian experimental design for finding reliable level set under input uncertainty. *IEEE Access*, 8:203982–203993.

Jain, L. and Jamieson, K. G. (2019). A new perspective on pool-based active classification and false-discovery control. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

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

Jeunen, O., Mandav, J., Potapov, I., Agarwal, N., Vaid, S., Shi, W., and Ustimenko, A. (2024). Multi-objective recommendation via multivariate policy learning. *arXiv preprint arXiv:2405.02141*.

Jourdan, M., Degenne, R., Baudry, D., de Heide, R., and Kaufmann, E. (2022). Top two algorithms revisited. In *Advances in Neural Information Processing Systems*, volume 35, pages 26791–26803. Curran Associates, Inc.

Jourdan, M. and Réda, C. (2023). An anytime algorithm for good arm identification. *arXiv preprint arXiv:2310.10359*.

Kalyanakrishnan, S. and Stone, P. (2010). Efficient selection of multiple bandit arms: theory and practice. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, page 511–518, Madison, WI, USA. Omnipress.

Kalyanakrishnan, S., Tewari, A., Auer, P., and Stone, P. (2012). Pac subset selection in stochastic multi-armed bandits. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, page 227–234, Madison, WI, USA. Omnipress.

Kamalbasha, S. and Eugster, M. (2021). Bayesian a/b testing for business decisions. In *Data Science – Analytics and Applications*, pages 50–57, Wiesbaden. Springer Fachmedien Wiesbaden.

Kanarios, K., Zhang, Q., and Ying, L. (2024). Cost aware best arm identification. *Reinforcement Learning Journal*, 4:1533–1545.

Kano, H., Honda, J., Sakamaki, K., Matsuura, K., Nakamura, A., and Sugiyama, M. (2019). Good arm identification via bandit feedback. *Mach. Learn.*, 108(5):721–745.

Karnin, Z., Koren, T., and Somekh, O. (2013). Almost optimal exploration in multi-armed bandits. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1238–1246, Atlanta, Georgia, USA. PMLR.

Katz-Samuels, J. and Bagherjeiran, A. (2019). Feasible bidding strategies through pure-exploration bandits.

Katz-Samuels, J. and Jamieson, K. (2020). The true sample complexity of identifying good arms. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 1781–1791. PMLR.

Katz-Samuels, J. and Scott, C. (2018). Feasible arm identification. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2535–2543. PMLR.

Katz-Samuels, J. and Scott, C. (2019). Top feasible arm identification. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 1593–1601. PMLR.

Kaufmann, E., Cappé, O., and Garivier, A. (2016). On the Complexity of Best Arm Identification in Multi-Armed Bandit Models. *Journal of Machine Learning Research*, 17:1–42.

Kaufmann, E., Cappé, O., and Garivier, A. (2014). On the complexity of a/b testing. In *Proceedings of The 27th Conference on Learning Theory*, volume 35 of *Proceedings of Machine Learning Research*, pages 461–481, Barcelona, Spain. PMLR.

Kaufmann, E., Koolen, W. M., and Garivier, A. (2018). Sequential test for the lowest mean: From thompson to murphy sampling. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Kirk, R. E. (2009). Experimental design. *Sage handbook of quantitative methods in psychology*, pages 23–45.

Kirschner, J., Mutny, M., Hiller, N., Ischebeck, R., and Krause, A. (2019). Adaptive and safe Bayesian optimization in high dimensions via one-dimensional subspaces. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3429–3438. PMLR.

Kohavi, R., Tang, D., and Xu, Y. (2020). *Trustworthy online controlled experiments: A practical guide to a/b testing*. Cambridge University Press.

Lalitha, A. L., Kalantari, K., Ma, Y., Deoras, A., and Kveton, B. (2023). Fixed-budget best-arm identification with heterogeneous reward variances. In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pages 1164–1173. PMLR.

Li, T., Shi, C., Wang, J., Zhou, F., et al. (2023). Optimal treatment allocation for efficient policy evaluation in sequential decision making. *Advances in Neural Information Processing Systems*, 36:48890–48905.

Li, Z. and Chi Cheung, W. (2024). Best arm identification with resource constraints. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 253–261. PMLR.

Lindner, D., Tschiatschek, S., Hofmann, K., and Krause, A. (2022). Interactively learning preference constraints in linear bandits. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 13505–13527. PMLR.

Liu, X., Li, B., Shi, P., and Ying, L. (2021). An efficient pessimistic-optimistic algorithm for stochastic linear bandits with general constraints. In *Advances in Neural Information Processing Systems*, volume 34, pages 24075–24086. Curran Associates, Inc.

Locatelli, A., Gutzeit, M., and Carpentier, A. (2016). An optimal algorithm for the thresholding bandit problem. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1690–1698, New York, New York, USA. PMLR.

Marchesi, A., Trovò, F., Gatti, N., et al. (2019). Learning maximin strategies in simulation-based games with infinite strategy spaces. In *Proceedings of the Smooth Games Optimization and Machine Learning Workshop*, pages 1–5.

Mason, B., Jain, L., Mukherjee, S., Camilleri, R., Jamieson, K., and Nowak, R. (2022). Nearly optimal algorithms for level set estimation. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 7625–7658. PMLR.

Mason, B., Jain, L., Tripathy, A., and Nowak, R. (2020). Finding all epsilon-good arms in stochastic bandits. In *Advances in Neural Information Processing Systems*, volume 33, pages 20707–20718. Curran Associates, Inc.

Mukherjee, S., Purushothama, N. K., Sudarsanam, N., and Ravindran, B. (2017). Thresholding bandits with augmented ucb. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, page 2515–2521. AAAI Press.

Neyman, J. (1992). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. In *Breakthroughs in statistics: Methodology and distribution*, pages 123–150. Springer.

Ngo, G., Nguyen, D., Phan-Trong, D., and Gupta, S. (2024). Active level set estimation for continuous search space with theoretical guarantee. In *Proceedings of the 15th Asian Conference on Machine Learning*, volume 222 of *Proceedings of Machine Learning Research*, pages 943–958. PMLR.

Nouiehed, M., Sanjabi, M., Huang, T., Lee, J. D., and Razaviyayn, M. (2019). Solving a class of non-convex min-max games using iterative first order methods. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Ouhamma, R., Maillard, O.-A., and Perchet, V. (2021). Online sign identification: Minimization of the number of errors in thresholding bandits. In *Advances in Neural Information Processing Systems*, volume 34, pages 18577–18589. Curran Associates, Inc.

Pacchiano, A., Ghavamzadeh, M., Bartlett, P., and Jiang, H. (2021). Stochastic bandits with linear constraints. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 2827–2835. PMLR.

Poiani, R., Degenne, R., Kaufmann, E., Metelli, A. M., and Restelli, M. (2024). Optimal multi-fidelity best-arm identification. *arXiv preprint arXiv:2406.03033*.

Poiani, R., Metelli, A. M., and Restelli, M. (2022). Multi-fidelity best-arm identification. In *Advances in Neural Information Processing Systems*, volume 35, pages 17857–17870. Curran Associates, Inc.

Roijers, D. M., Zintgraf, L. M., and Nowé, A. (2017). Interactive thompson sampling for multi-objective multi-armed bandits. In *Algorithmic Decision Theory*, pages 18–34, Cham. Springer International Publishing.

Russac, Y., Katsimerou, C., Bohle, D., Cappé, O., Garivier, A., and Koolen, W. M. (2021). A/b/n testing with control in the presence of subpopulations. In *Advances in Neural Information Processing Systems*, volume 34, pages 25100–25110. Curran Associates, Inc.

Russo, D. (2016). Simple bayesian algorithms for best arm identification. In *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 1417–1418, Columbia University, New York, New York, USA. PMLR.

Seltman, H. J. (2012). Experimental design and analysis.

Senadeera, M., Rana, S., Gupta, S., and Venkatesh, S. (2020). Level set estimation with search space warping. In *Advances in Knowledge Discovery and Data Mining*, pages 827–839, Cham. Springer International Publishing.

Shang, X., Colin, I., Barlier, M., and Cherkaoui, H. (2023). Price of safety in linear best arm identification. *arXiv preprint arXiv:2309.08709*.

Shekhar, S. and Javidi, T. (2019). Multiscale gaussian process level set estimation. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 3283–3291. PMLR.

Siroker, D. and Koomen, P. (2015). *A/B testing: The most powerful way to turn clicks into customers*. John Wiley & Sons.

Student (1908). The probable error of a mean. *Biometrika*, pages 1–25.

Sui, Y., Gotovos, A., Burdick, J., and Krause, A. (2015). Safe exploration for optimization with gaussian processes. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 997–1005, Lille, France. PMLR.

Sui, Y., Zhuang, V., Burdick, J., and Yue, Y. (2018). Stagewise safe Bayesian optimization with Gaussian processes. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4781–4789. PMLR.

Tang, D., Jain, R., Nayyar, A., and Nuzzo, P. (2024). Pure exploration for constrained best mixed arm identification with a fixed budget. *arXiv preprint arXiv:2405.15090*.

Tao, C., Blanco, S., Peng, J., and Zhou, Y. (2019). Thresholding bandit with optimal aggregate regret. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Tsai, T.-H., Tsai, Y.-D., and Lin, S.-D. (2024). lil'hdoc: An algorithm for good arm identification under small threshold gap. In *Advances in Knowledge Discovery and Data Mining: 28th Pacific-Asia Conference on Knowledge Discovery and Data Mining, PAKDD 2024, Taipei, Taiwan, May 7–10, 2024, Proceedings, Part V*, page 78–89, Berlin, Heidelberg. Springer-Verlag.

Tsai, Y.-D., Tsai, T.-H., and Lin, S.-D. (2023). Differential good arm identification. *arXiv preprint arXiv:2303.07154*.

Turchetta, M., Berkenkamp, F., and Krause, A. (2019). Safe exploration for interactive machine learning. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Van Moffaert, K., Van Vaerenbergh, K., Vrancx, P., and Nowe, A. (2014). Multi-objective χ-armed bandits. In *2014 International Joint Conference on Neural Networks (IJCNN)*, pages 2331–2338.

Wang, Z. and Scarlett, J. (2022). Max-min grouped bandits. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(8):8603–8611.

Wang, Z., Wagenmaker, A. J., and Jamieson, K. (2022). Best arm identification with safety constraints. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 9114–9146. PMLR.

Weltz, J., Fiez, T., Volfovsky, A., Laber, E., Mason, B., nassif, h., and Jain, L. (2023). Experimental designs for heteroskedastic variance. In *Advances in Neural Information Processing Systems*, volume 36, pages 65967–66005. Curran Associates, Inc.

Wu, Y., Zheng, Z., and Zhu, T. (2023). Best arm identification with fairness constraints on subpopulations. In *2023 Winter Simulation Conference (WSC)*, pages 540–551.

Yahyaa, S., Drugan, M., and Manderick, B. (2014a). Knowledge gradient for multi-objective multi-armed bandit algorithms. In *ICAART 2014 - Proceedings of the 6th International Conference on Agents and Artificial Intelligence*.

Yahyaa, S. Q., Drugan, M. M., and Manderick, B. (2014b). Annealing-pareto multi-objective multi-armed bandit algorithm. In *2014 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL)*, pages 1–8.

Zanette, A., Zhang, J., and Kochenderfer, M. J. (2019). Robust super-level set estimation using gaussian processes. In *Machine Learning and Knowledge Discovery in Databases*, pages 276–291, Cham. Springer International Publishing.

Zhang, Q. and Ying, L. (2023). Fast and regret optimal best arm identification: Fundamental limits and low-complexity algorithms. In *Advances in Neural Information Processing Systems*, volume 36, pages 16729–16769. Curran Associates, Inc.

Zhao, Y., Stephens, C., Szepesvari, C., and Jun, K.-S. (2023). Revisiting simple regret: Fast rates for returning a good arm. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 42110–42158. PMLR.

Zhou, R. and Tian, C. (2022). Approximate top-m arm identification with heterogeneous reward variances. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 7483–7504. PMLR.

Zintgraf, L. M., Roijers, D. M., Linders, S., Jonker, C. M., and Nowé, A. (2018). Ordered preference elicitation strategies for supporting multi-objective decision making. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '18, page 1477–1485, Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Appendix Table of Contents

A Additional Experiments **17**

A.1 Generality to Heterogeneous Variance and Number of Treatments . . . 17

A.2 Robustness to Unknown Variance . . . 17

A.3 Type I Error . . . 18

B Variant of `SHRVar` with Confidence-Based Elimination **19**

B.1 Confidence-Based Elimination . . . 19

B.2 Theoretical Guarantee . . . 21

C Discussions **22**

C.1 Variance Equalization Compared to Neyman Allocation . . . 22

C.2 Equivalence of Max-Min and Expectation Characterization of Best Treatment . . . 23

C.3 Relaxing Assumptions . . . 24

C.4 Other Discussions . . . 25

D Related Works **26**

D.1 Pure Exploration Objectives . . . 27

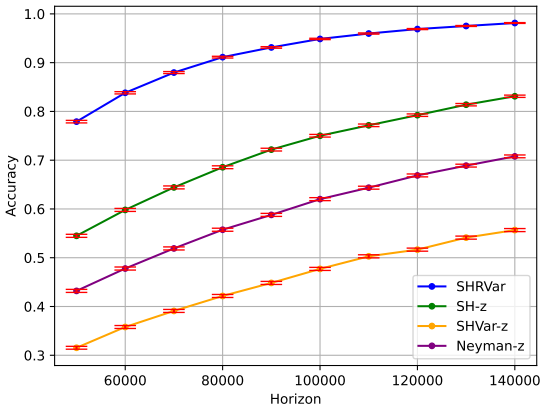
D.2 Bandits with Multiple Metrics . . . 28

D.3 Best Arm Identification Formulations . . . 29

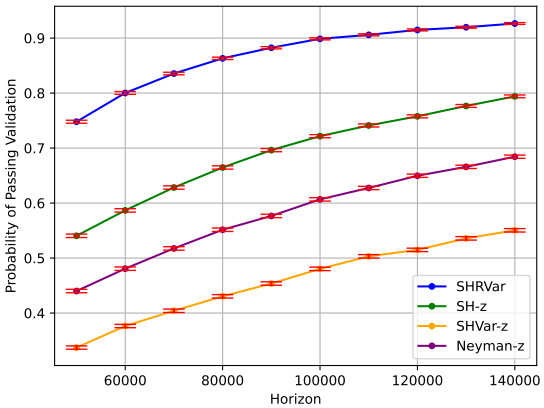
E Proof of Proposition 3.1 **30**

F Proof of Theorem 5.1 **32**

G Proof of Theorem 5.2 **35**



(a) Exploration Accuracy



(b) Validation Success Probability

Figure 3: (a) Exploration Accuracy: rate of identifying the best treatment in exploration as time horizon T increases. (b) Validation Success Probability: rate of passing validation, i.e., successfully validating a better-than-control treatment after both phases. Error bars represent confidence intervals.

A Additional Experiments

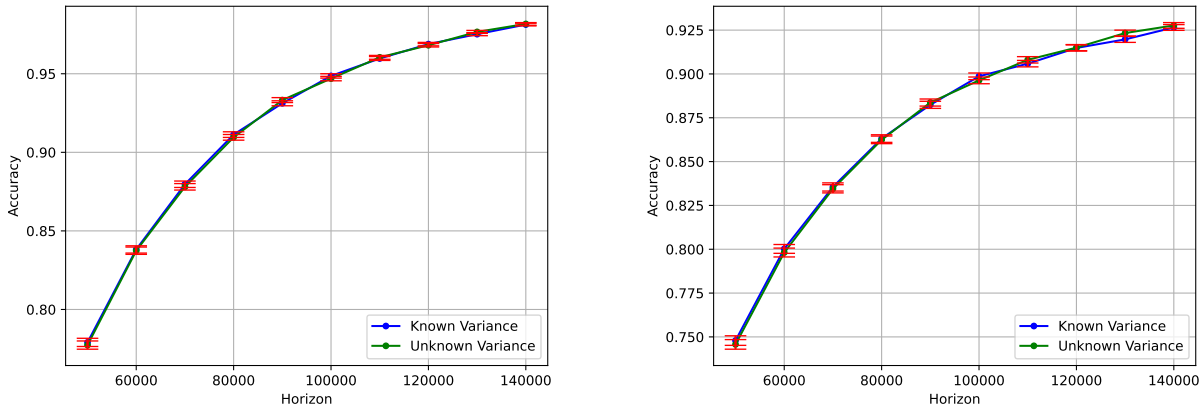
In this section, we provide results from additional numerical experiments that demonstrate our proposed approach's generality against multiple factors such as a large number of treatments, heterogeneous variance, and unknown variances. All results are evaluated over 10^5 repetitions. All experiments in this paper are conducted on a single CPU with less than an hour of runtime.

A.1 Generality to Heterogeneous Variance and Number of Treatments

We consider an `M3AB` problem with $A = 128$ treatments and $M = 3$ metrics, where each treatment-metric pair has a different variance. The reward of control is set to $[0, 0, 0]$ for all metrics, and the variance of control is $[1, 1, 1]$. For treatments, we set the z-value for the best treatment to be $[0.15, 0.15, 0.05]$ for the three metrics, and set the z-value for other treatments to be $[-0.05, 0.25, 0.25]$. Therefore, the only treatment that is better than control for all three metrics is the optimal treatment. The relative variances of treatments for the three metrics are set to $[0.8, 0.5, 0.2]$ plus a random noise uniformly distributed in $[-0.1, 0.1]$. We use the same validation constant $\xi_{a,i}$ for all pairs of treatment and metric. Then, we calculate the standard deviation and the reward mean of each arm inversely from z-values and relative variances, and the variances will be heterogeneous. The validation time horizon $T_{\mathrm{v}} = 2000$. We compare our proposed algorithm `SHRVar` to three baselines, i.e., `SH-z` and `SHVar-z` introduced in Sec. 6, and the baseline `Neyman-z` which replaces the sampling strategy of `SHVar-z` with Neyman allocation (Neyman, 1992). The detailed discussion of Neyman allocation and the variance equalization allocation that we use in our paper will be provided in the appendix Sec. C.1. The results are shown in Fig. 3. The result shows that for both the probability of identifying the best treatment a^* and the probability of passing validation, our proposed `SHRVar` has the best performance compared to all baselines, including Neyman allocation, which is asymptotically optimal in two-armed settings. The trend coincides with the results shown in Sec. 6, which is for a smaller number of arms and variances only heterogeneous among metrics. Therefore, it shows the generality of `SHRVar` and the advantage that `SHRVar` is robust to the number of arms and the level of heterogeneity between metrics and arms.

A.2 Robustness to Unknown Variance

In this section, we incorporate `SHRVar` into the unknown variance setting and compare the performance to show the robustness of the proposed approach. We call the variant `SHRVar-Ada`. We assume the algorithm does not know the variance of each arm at the beginning, and therefore we first sample each arm uniformly and collect $T/(A+1)\log_2 A$ samples for each arm. Then, we estimate the variance of each arm using the sample variance,



(a) Exploration Accuracy

(b) Validation Success Probability

Figure 4: (a) Exploration Accuracy: rate of identifying the best treatment in exploration as time horizon T increases. (b) Validation Success Probability: rate of passing validation, i.e., successfully validating a better-than-control treatment after both phases. Error bars represent confidence intervals.

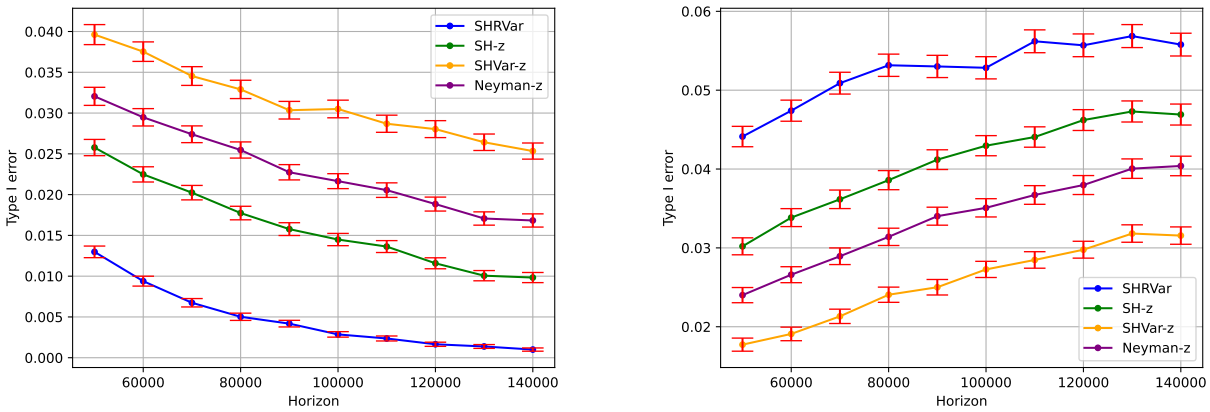
i.e.,

$$\hat{\mu}_0(a,i) = \frac{1}{N_0(a)} \sum_{k=1}^{N_0(a)} \mathbf{x}_{a,k}, \quad \hat{\sigma}^2_{a,i} = \frac{1}{N_0(a)-1} \sum_{k=1}^{N_0(a)} \left(\mathbf{x}_{a,k} - \hat{\mu}_0(a,i)\right),$$

where $\mathbf{x}_{a,i,k}$ is the random reward of metric i and treatment a for the k-th pull, $N_0(a) = T/(A+1)\log_2 A$ for all arm, and $\hat{\mu}_0(a,i)$ is the empirical mean after the uniform pulling round. Then, we plug in the estimated $\hat{\sigma}_{a,i}$ to `SHRVar` with the remaining budget and proceed as if it is the true variance. The probability of finding the best treatment, and the probability of passing validation in the above `M3AB` setting are shown in Fig. 4. It shows that `SHRVar-Ada` in the unknown variance setting has almost the same performance as `SHRVar` of the known variance case, in both the probability of identifying the best treatment a^* and the probability of passing validation. As the confidence intervals are overlapping with each other, given that the length of the interval is already small, we conclude that there is no significant difference between the unknown-variance variant and the original `SHRVar` algorithm. Therefore, it shows the robustness of `SHRVar` towards unknown variance.

A.3 Type I Error

It is also unsatisfactory if a worse-than-control arm is recommended and passes the validation against control, especially in settings where no treatment is better than control in all metrics. This is referred to as the Type I error. In this section, we study the type I error of the proposed algorithm compared to baselines. In addition to the settings mentioned above, where the best treatment is better than control for all metrics, we consider a null setting where all treatments are worse than control in at least one metric. This instance can be constructed from decreasing all the z-values of each treatment by 0.1 in the above example, and the new z-value for the best treatment is $[-0.05, 0.05, 0.05]$, and for other treatments is $[-0.15, 0.15, 0.15]$. The type I error of both instances is shown in Fig. 5. It is shown that if there exists a treatment which is uniformly better than the control, our proposed algorithm `SHRVar` will have a type I error smaller than all baselines. This is because our algorithm will recommend the treatment with a higher probability of passing validation. If there exist treatments uniformly better than control, it will try to recommend the one in this set with the largest positive effect gap over the bottleneck metric. Therefore, it automatically decreases the type I error. If there is no treatment better than the control for all metrics, our algorithm will have a slightly larger type I error. This is because we try to recommend the treatment that has a better effect, i.e., a smaller effect gap to control, even though it is inferior. Naturally, other algorithms have slightly better type I error since they try to recommend treatments with a more negative effect more often, which is easy to distinguish for A/B tests. We emphasize that the type I error is mainly controlled by the A/B test in validation instead of the sampling and recommendation rules designed in `SHRVar`.



(a) Standard

(b) Null

Figure 5: (a) Standard instance: the best treatment is better than control in all metrics. (b) Null instance: no treatment is better than control uniformly.

For example, it is controlled by the required p-value to pass validation in non-Bayesian settings or the posterior probability threshold in Bayesian validation. Therefore, if worse algorithms return treatments with more negative effects, the type I error will be lower since the negative effect gap to control is larger. On the other hand, `SHRVar` returns arms with effects close to zero, so the type I error is close to the nominal rate defined by the p-value. Our experiments show that the type I error of `SHRVar` is not significantly worse than baselines in the case where no treatment is better than control, and we can control the type I error by setting the p-value or the posterior threshold properly.

B Variant of `SHRVar` with Confidence-Based Elimination

In this section, we study and discuss the variant of the proposed `SHRVar` algorithm used in our numerical section, which is the `SHRVar-c` algorithm with confidence-based elimination. The details of the algorithm are summarized in algorithm 2. The only difference between `SHRVar-c` and `SHRVar` is the elimination strategy in Lines 9-12 that computes the confidence level $\delta_s(a)$ for each treatment and uses it to eliminate treatments. We let the confidence bonus be:

$$b_s^\delta(a,i) = 2\sqrt{\left(\frac{\rho_{a,i}^2}{N_s(a)} + \frac{\lambda_{a,i}^2}{N_s(0)}\right)\log\left(\frac{|\mathcal{A}_s|M}{\delta}\right)}. \tag{5}$$

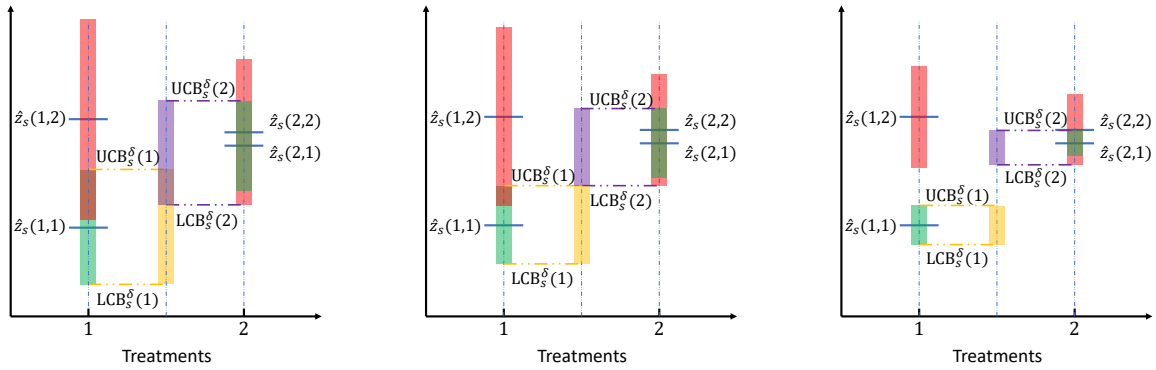
B.1 Confidence-Based Elimination

Intuition. We illustrate the intuition of this elimination rule through a simple example shown in Fig. 6. Recall that for each z value estimate $\hat{z}_s(a,i)$, we can construct its confidence interval so that with probability $1-\delta$, we have:

$$\underbrace{\min_i \left\{\hat{z}_s(a,i) - b_s^\delta(a,i)\right\}}_{\mathrm{LCB}_s^\delta(a)} \le \hat{z}_s(a,i) \le \underbrace{\min_i \left\{\hat{z}_s(a,i) + b_s^\delta(a,i)\right\}}_{\mathrm{UCB}_s^\delta(a)},$$

where the confidence bonus $b_s^\delta(a,i)$ can be explicitly computed as:

$$b_s^\delta(a,i) = 2\sqrt{\left(\frac{\rho_{a,i}^2}{N_s(a)} + \frac{\lambda_{a,i}^2}{N_s(0)}\right)\log\left(\frac{|\mathcal{A}_s|M}{\delta}\right)}.$$



(a) overlap, $\delta = 0.01$ (b) almost overlap, $\delta = 0.05$ (c) separate, $\delta = 0.1$

Figure 6: A toy example to compute $\delta_s(1)$ where treatment 2 has the largest LCB, i.e., $\arg\max_a \mathrm{LCB}_s^\delta(a) = 2$. Empirical z values are marked with solid lines, and confidence intervals are marked with color bars. In (a), δ is too small so that $\mathrm{UCB}_s^\delta(1) > \mathrm{LCB}_s^\delta(2)$. In (c), δ is too large so that $\mathrm{UCB}_s^\delta(1) < \mathrm{LCB}_s^\delta(2)$. In (b), we have $\mathrm{UCB}_s^\delta(1) = \mathrm{LCB}_s^\delta(2)$, meaning $\delta_s(1) = 0.05$.

Then, for any confidence level δ and any treatment a, if there exists another treatment a' that has a lower confidence bound $\mathrm{LCB}_s^\delta(a')$ which is larger than the upper confidence bound $\mathrm{UCB}_s^\delta(a)$ of treatment a, it means the ground-truth z value $z_{a,i}$ is smaller than $z_{a',i}$ with probability at least $1-\delta$. In other words, if $\mathrm{UCB}_s^\delta(a) < \max_{a'} \mathrm{LCB}_s^\delta(a')$, this means treatment a is not the best treatment with high probability and eliminating it would incur an error probability at most δ. This situation is illustrated in Fig. 6(c) where the yellow confidence interval of treatment 1 does not overlap with the purple confidence interval of treatment 2. However, if we decrease δ in this case while keeping the confidence intervals of both treatments non-overlapping, we can find a tighter probability of mistake upper bound. The optimal δ should be the case shown in Fig. 6(b), where the yellow and purple confidence intervals for both treatments almost overlap with each other, and we record this confidence level as $\delta_s(a)$:

$$\delta_s(a) = \inf\left\{\delta \in \mathbb{R}_+ \,\middle|\, \mathrm{UCB}_s^\delta(a) \le \max_{a' \in \mathcal{A}_s} \mathrm{LCB}_s^\delta(a')\right\}.$$

Therefore, $\delta_s(a)$ represents the tightest error probability upper bound if we eliminate treatment a. Then, it becomes natural that we should eliminate the treatments with less probability of making mistakes since we are more confident that they are not the best treatment, which is the treatments with less $\delta_s(a)$.

Underestimation Alleviation. It is shown in our numerical experiment in Fig. 2(a) that `SHRVar` may suffer performance loss due to underestimating the minimum z value $\min_i z_{a,i}$ if the sampling strategy is not optimally designed, which deviates from equalization of the estimate variance. To be specific, since we use $\min_i \hat{z}_s(a,i)$ to estimate the ground-truth $\min_i z_{a,i}$, we have:

$$\mathbb{E}[\min_i \hat{z}_s(a,i)] \le \min_i \mathbb{E}[\hat{z}_s(a,i)] = \min_i z_{a,i}.$$

If all the estimates $\hat{z}_s(a,i)$ have the same variance that has been perfectly balanced through the design of the sampling rule, the difference between the left-hand side and right-hand side of the inequality above will be the same for all treatments a. Then, using `SHRVar` for elimination would not incur bias towards any treatment and would still produce a good performance shown in Fig. 2(a). However, if the variances are not balanced, the underestimation gap would be different for each treatment, and therefore, an implicit bias would occur if we eliminate treatments based on $\min_i \hat{z}_s(a,i)$, which is harmful and increases the probability of error. The confidence-based elimination rule in `SHRVar-c` will alleviate this underestimation issue since the confidence intervals will still be valid even if the variances are not well-balanced, and thus the confidence level $\delta_s(a)$ is not biased. However, the computation of $\delta_s(a)$ involves multiple treatments and may incur additional variance in the elimination process, which explains the slightly worse empirical performance compared to `SHRVar` in Fig. 2(a). One could also use methods such as the double estimator (Hasselt, 2010) to mitigate this issue, but this will also add additional variance to each estimator since only half of the samples are used to identify a^*.

B.2 Theoretical Guarantee

We present the probability of error upper bound guarantee for `SHRVar-c` in algorithm 2 analogous to Theorem 5.1 in this section. We also assume A is a power of 2. Recall that for an arbitrary set $\mathcal{S}$ of treatments, the heterogeneity of relative variance $\kappa_{\mathcal{S},a,i}$ is defined as follows:

$$\kappa_{\mathcal{S},a,i} = \frac{\frac{\rho_{a,i}^2}{\max_i \rho_{a,i}^2}\rho_{\mathcal{S},\Sigma} + \frac{\lambda_{a,i}^2}{\max_{a\in\mathcal{S},i}\lambda_{a,i}^2}\lambda_{\mathcal{S},\Sigma}}{\rho_{\mathcal{S},\Sigma}+\lambda_{\mathcal{S},\Sigma}}.$$

We additionally define another effective gap $\tilde{D}_{\mathcal{S},a}$ similar to the gap $D_{\mathcal{S},a}$ used in Theorem. 5.1 as follows:

$$\tilde{D}_{\mathcal{S},a}^2 = \min_{i\in[M]}\max_{j\in[M]}\min\left\{\frac{[z_{a^*,i}-z_{a,j}]_+^2}{(\kappa_{\mathcal{S},a,j}+\kappa_{\mathcal{S},a^*,i})^2}, A_{\mathcal{S},a,i,j}\right\},$$

where the additional term $A_{\mathcal{S},a,i,j} = +\infty$ if $\kappa_{\mathcal{S},a,j} \le \kappa_{\mathcal{S},a,i}$, and is expressed as follows:

$$A_{\mathcal{S},a,i,j} = \begin{cases} +\infty & \kappa_{\mathcal{S},a,j}, \le \kappa_{\mathcal{S},a,i} \\ \frac{[z_{a^*,i}-z_{a,j}]_+^2}{(\kappa_{\mathcal{S},a,j}-\kappa_{\mathcal{S},a^*,i})^2} + \Delta_{\min}^2 - \frac{8A\log_2^2 A}{T}, & \kappa_{\mathcal{S},a,j} > \kappa_{\mathcal{S},a,i} \end{cases}.$$

Finally, for any set of treatments $\mathcal{S}$, let $\tilde{\mathcal{S}}_c'$ be its subset excluding 1/4 of the treatments with smallest $\tilde{D}_{\mathcal{S},a}$ similar to the definition of $\mathcal{S}_c'$ in Theorem. 5.1. The $\tilde{H}_3$ complexity measure used to characterize the performance of `SHRVar-c` is defined as:

$$\tilde{H}_3^{-1} = \min_{\mathcal{S}:a^*\in\mathcal{S}} \frac{\min_{a\in\tilde{\mathcal{S}}_c'}\tilde{D}_{\mathcal{S},a}^2}{(\rho_{\mathcal{S},\Sigma}+\lambda_{\mathcal{S},\Sigma})^2}.$$

Theorem B.1. *If we use `SHRVar-c` in algorithm 2 to identify the best treatment a^* in our exploration phase, the probability of mistake can be upper bounded as:*

$$\mathbb{P}(\hat{a}\neq a^*) \le 6M\log_2 A\cdot\exp\left(-\frac{T}{2\tilde{H}_3\log_2 A}\right).$$

Proof Roadmap. For Theorem 5.2, we use the same proof framework as the proof of Theorem 5.1 except that we need to analyze the probability that a sub-optimal treatment has lower confidence $\delta_s(a)$ than the best treatment. This is more challenging since it also involves a third treatment, i.e., the treatment that has the largest lower confidence bound. Our solution is to couple the confidence-based elimination rule with the simple elimination rule based on the minimum z value used in `SHRVar`, to only analyze the event where the concentration of the minimum z value estimate does not hold. This can only happen when the confidence $\delta_s(a^*)$ for the best treatment is small, meaning a large deviation happens between the best treatment and the treatment with the largest LCB. From this coupling argument, we will be able to reach the instance-dependent exponent $\tilde{D}_{\mathcal{A}_s,a}$ which is almost the same as $D_{\mathcal{A}_s,a}$ in Theorem 5.1. The complete proof is in the appendix section G

Effective Gap. The effective gap notion $\tilde{D}_{\mathcal{S},a}$ in Theorem. 5.2 is slightly different from $D_{\mathcal{S},a}^2$ in Theorem. 5.1 due to the additional $A_{\mathcal{S},a,i,j}$ term in the case where $\kappa_{\mathcal{A}_s,a,j} > \kappa_{\mathcal{A}_s,a^*,i}$. This means the confidence bonus $b_s^\delta(a,j)$ for the sub-optimal treatment a is larger than the best treatment., which will make the best treatment "vulnerable" to elimination as it is easier to increase δ to create a separation of confidence intervals for the best treatment even if it has a larger empirical z value due to less variance. The additional term $A_{\mathcal{S},a,i,j}$ accounts for the probability that this case happens. When T is as large as $\tilde{\Omega}(A\Delta_{\min}^{-2})$, the two effective gaps will be the same, since:

$$A_{\mathcal{S},a,i,j} \ge \frac{[z_{a^*,i}-z_{a,j}]_+^2}{(\kappa_{\mathcal{S},a,j}-\kappa_{\mathcal{S},a^*,i})^2} \ge \frac{[z_{a^*,i}-z_{a,j}]_+^2}{(\kappa_{\mathcal{S},a,j}+\kappa_{\mathcal{S},a^*,i})^2},$$

and thus will be vacuous when the minimization is taken. This means when the exploration time horizon T is small, `SHRVar-c` may have a slightly larger probability of error compared to `SHRVar`. As T increases, the difference will become smaller and smaller, and finally, the two algorithms will have the same performance when T is large enough. This theoretical insight coincides with our numerical experiment in Fig. 2(a) as the gap between `SHRVar` and `SHRVar-c` becomes smaller as the time horizon increases. Therefore, in the large T regime, `SHRVar-c` will be equivalent to `SHRVar` and enjoys all the characteristics of `SHRVar`.

C Discussions

In this section, we provide further discussions of some aspects of our paper that may provide insights to future research and better engage our findings with the existing literature.

C.1 Variance Equalization Compared to Neyman Allocation

In the best arm identification problems with a fixed budget, the essential goal is to design a sampling strategy that optimally reduces the uncertainty of the estimators and results in the minimum probability of error. The variance equalization design principle is widely used in the best arm identification literature (Weltz et al., 2023; Lalitha et al., 2023; Antos et al., 2010), even though other principles are also studied. In this section, we discuss the reason and advantage of choosing the variance equalization sampling rule in this paper to design `SHRVar`, and compare it to the well-known Neyman allocation.

Neyman Allocation. The Neyman allocation arises from estimation problems in stratified sampling tasks (Neyman, 1992). The essential idea of Neyman allocation is to assign the number of samples for each random variable proportional to its standard deviation. In the fixed budget best arm identification problem, the number of pulls for each arm under Neyman allocation would be proportional to the standard deviation of the reward, i.e., $N(a) \propto \sigma_a$, where $N(a)$ is the number of samples for each arm and σ_a is the standard deviation. The Neyman allocation has been proven optimal in statistical estimation (Glynn and Juneja, 2004) or evaluation tasks (Li et al., 2023), but the performance in decision-making tasks, especially for the setting where there exists multiple arms, is elusive. In (Kaufmann et al., 2016), it has been shown that for a limited setting, i.e., a two-armed fixed-budget best arm identification problem with Gaussian noise, Neyman allocation achieves the optimal asymptotic rate of error as follows and matches the lower bound.

$$\lim_{T\to\infty} -\frac{1}{T}\log \mathbb{P}\left(\hat{a} \neq a^*\right) = \frac{(\mu_1-\mu_2)^2}{(\sigma_1+\sigma_2)^2}, \quad \text{for Neyman allocation.}$$

where $\hat{a}$ is the recommended arm by the algorithm, a^* is the best arm with highest expected reward, and μ_1 and μ_2 are the reward mean of both arms. But the performance for more general settings, such as multiple arms or other reward distributions, is not optimal as shown in our experiment in Sec A.

Variance Equalization Allocation. Our sampling rule design principle is variance equalization, which has been efficiently used in best arm identification problems under both fixed budget (Lalitha et al., 2023) and fixed confidence settings (Weltz et al., 2023; Antos et al., 2010) with multiple arms. Specifically, the principle aims to design and balance the number of pulls for each arm, so that the variance of each estimator for decision-making will be the same. In the classic multiple-armed bandit problem, it will result in an allocation rule where the number of pulls $N(a)$ for each arm a is proportional to the variance σ_a^2 of that arm. The sequential halving algorithm designed based on this rule is optimal up to a logarithmic factor in best arm identification with fixed confidence (Weltz et al., 2023).

Comparison. We show that variance equalization allocation is more general and has a better performance over a larger set of problems compared to Neyman allocation, without losing much optimality in the two-armed Gaussian bandit problem. This justifies why we chose this rule to design `SHRVar`. We consider a multi-armed bandit model with $A+2$ arms. Suppose arm 1 is the best arm with expected reward μ_1, and all other arms have the same reward mean μ_2. Both arms 1 and 2 have large reward variance Σ^2, and all other arms have small reward variance σ^2, and we assume σ is much smaller than Σ. Then, with a large enough time horizon T, the bottleneck of the best arm identification task in this model would be the test between arm 1 and arm 2, and the exponent of error probability for each allocation rule would be as follows:

$$\lim_{T\to\infty} -\frac{1}{T}\log \mathbb{P}\left(\hat{a} \neq a^*\right) = \frac{(\mu_1-\mu_2)^2}{4\Sigma^2+2A\Sigma\sigma}, \quad \text{for Neyman allocation;}$$
$$\lim_{T\to\infty} -\frac{1}{T}\log \mathbb{P}\left(\hat{a} \neq a^*\right) = \frac{(\mu_1-\mu_2)^2}{4\Sigma^2+2A\sigma^2}, \quad \text{for variance equalization allocation.}$$

Therefore, Neyman allocation has a strictly worse performance than variance equalization for $\sigma < \Sigma$, independent of the number of arms. As the number of arms A becomes much larger, especially much larger than Σ^2/σ^2, the

denominator of Neyman allocation can be arbitrarily larger than variance equalization, i.e.,

$$\lim_{A\to\infty}\lim_{T\to\infty}\frac{\log\mathbb{P}\left(\hat{a}\neq a^*|\text{Neyman}\right)}{\log\mathbb{P}\left(\hat{a}\neq a^*|\text{variance equalization}\right)}=\lim_{A\to\infty}\frac{4\Sigma^2+2A\Sigma\sigma}{4\Sigma^2+2A\sigma^2}=\frac{\Sigma}{\sigma}.$$

Since Σ can be arbitrarily larger than σ, the probability of error for Neyman allocation can be arbitrarily larger than that of variance equalization allocation. This is because variance-equalization assigns more samples to large variance arms, which is helpful when the bottleneck involves two arms with large variance. Therefore, variance-equalization is a more general approach that is near-optimal over a wider range of problems. This is also why it has been much more adopted in the best arm identification literature than Neyman allocation (Weltz et al., 2023; Lalitha et al., 2023). On the other hand, in the specialized two-armed Gaussian bandit setting, the performance of variance equalization allocation is still optimal up to a factor of 2, an absolute constant which cannot be arbitrarily worse, i.e.,

$$\lim_{T\to\infty}-\frac{1}{T}\log\mathbb{P}\left(\hat{a}\neq a^*\right)=\frac{(\mu_1-\mu_2)^2}{2\left(\sigma_1^2+\sigma_2^2\right)},\quad\text{for variance equalization allocation.}$$

And we have:

$$\frac{(\mu_1-\mu_2)^2}{2\left(\sigma_1^2+\sigma_2^2\right)}\geq\frac{1}{2}\frac{(\mu_1-\mu_2)^2}{(\sigma_1+\sigma_2)^2},$$

where the left-hand side is half of the exponent for Neyman allocation. This also shows that variance equalization does not suffer much performance loss in the specialized setting and should still achieve near-optimal probability of error. Overall, variance equalization allocation achieves a better performance in our studied `M3AB` problem as shown in our additional experiments in Sec A.

C.2 Equivalence of Max-Min and Expectation Characterization of Best Treatment

In this section, we illustrate and discuss the relation between the max-min formulation studied in our paper in equation 1 and a more natural formulation that maximizes the expected probability of passing validation. Recall that $\mathcal{E}_{\text{v},i}$ is the event that the treatment $\hat{a}$ selected by the exploration phase beats the control in the validation A/B test. The two formulations can be represented as follows:

$$\text{Max-Min Formulation:}\quad a_1^*=\arg\max_{a\in\mathcal{A}/\{0\}}\min_{i\in[M]}\mathbb{P}\left(\mathcal{E}_{\text{v},i}\right);$$

$$\text{Expected Formulation:}\quad a_2^*=\arg\max_{a\in\mathcal{A}/\{0\}}\mathbb{P}\left(\bigcap_{i\in[M]}\mathcal{E}_{\text{v},i}\right);$$

Let's consider the non-Bayesian validation with two treatments $\mathcal{A}=\{0,1,2\}$ and two metrics $M=2$ with $\delta_i=0.5$ for simplicity. Assume both treatments are better than the control for both metrics. The argument can easily be generalized to settings with multiple treatments and metrics. Suppose the recommended treatment to validate is a treatment a. Then, by proposition 3.1, we have:

$$a_1^*=\arg\max_{a\in\{1,2\}}\min_{i\in\{1,2\}}\left(\frac{\mu_{a,i}-\mu_{0,i}}{\sqrt{\sigma_{a,i}^2+\sigma_{0,i}^2}}+\xi_{a,i}\right).$$

Recall that from the proof of proposition 3.1, we have:

$$\mathcal{E}_{\text{v},i}=\left\{\hat{\mu}_{\text{v}}(a,i)-\hat{\mu}_{\text{v}}(0,i)\geq\Phi^{-1}(1-\delta_i)\sqrt{\frac{\sigma_{a,i}^2+\sigma_{0,i}^2}{T_{\text{v}}/2}}\right\}.$$

The probability of failure can be characterized using the cumulative distribution function $\Phi(\cdot)$ of standard normal as follows:

$$\mathbb{P}\left(\mathcal{E}_{\text{v},i}^{\complement}\right)=\Phi\left(-\frac{\mu_{a,i}-\mu_{0,i}}{\sqrt{\sigma_{a,i}^2+\sigma_{0,i}^2}}\sqrt{T_{\text{v}}/2}\right).$$

On the other hand, we consider the expected formulation as follows. First, by the independence among metrics, we have:

$$\begin{aligned}
\mathbb{P}\left(\bigcap_{i\in[M]} \mathcal{E}_{\mathrm{v},i}\right) &= \prod_{i\in[M]} \mathbb{P}\left(\mathcal{E}_{\mathrm{v},i}\right) = \prod_{i\in[M]} \left(1-\mathbb{P}\left(\mathcal{E}_{\mathrm{v},i}^{\complement}\right)\right) \\
&= \left(1-\Phi\left(-\frac{\mu_{a,1}-\mu_{0,1}}{\sqrt{\sigma_{a,1}^2+\sigma_{0,1}^2}}\sqrt{T_{\mathrm{v}}/2}\right)\right)\left(1-\Phi\left(-\frac{\mu_{a,2}-\mu_{0,2}}{\sqrt{\sigma_{a,2}^2+\sigma_{0,2}^2}}\sqrt{T_{\mathrm{v}}/2}\right)\right) \\
&= 1-\underbrace{\Phi\left(-\frac{\mu_{a,1}-\mu_{0,1}}{\sqrt{\sigma_{a,1}^2+\sigma_{0,1}^2}}\sqrt{T_{\mathrm{v}}/2}\right)}_{A_1(a)} - \underbrace{\Phi\left(-\frac{\mu_{a,2}-\mu_{0,2}}{\sqrt{\sigma_{a,2}^2+\sigma_{0,2}^2}}\sqrt{T_{\mathrm{v}}/2}\right)}_{A_2(a)} \\
&\quad + \underbrace{\Phi\left(-\frac{\mu_{a,1}-\mu_{0,1}}{\sqrt{\sigma_{a,1}^2+\sigma_{0,1}^2}}\sqrt{T_{\mathrm{v}}/2}\right)\Phi\left(-\frac{\mu_{a,2}-\mu_{0,2}}{\sqrt{\sigma_{a,2}^2+\sigma_{0,2}^2}}\sqrt{T_{\mathrm{v}}/2}\right)}_{A_{1,2}(a)}.
\end{aligned}$$

So we have:

$$a_2^* = \arg\min_{a\in\{1,2\}} A_1(a) + A_2(a) - A_{1,2}(a).$$

Since both treatments are better than the control, when T_{v} is large, all three terms approach 0. Considering the shape of the tail of the normal CDF, i.e., $\Phi(-\alpha\sqrt{T}) = \Theta(\mathrm{poly}(T)\exp(-\alpha^2 T^2))$, it is not difficult to see that all three terms $A_1(a)$, $A_2(a)$, and $A_{1,2}(a)$ decreases to 0 exponentially fast. Therefore, the exponent α will determine the largest term among the three. Since all terms are decreasing exponentially, the largest term will dominate the choice of the arg min since the difference in other terms among arms will be negligible compared to the largest term. Therefore, we can transform the problem into the following when T_{v} is large enough:

$$\begin{aligned}
a_2^* &= \arg\min_{a\in\{1,2\}} A_1(a) + A_2(a) - A_{1,2}(a) \\
&= \arg\min_{a\in\{1,2\}} \max\left\{A_1(a), A_2(a), A_{1,2}(a)\right\} \\
&= \arg\max_{a\in\{1,2\}} \min\left\{\frac{\mu_{a,1}-\mu_{0,1}}{\sqrt{\sigma_{a,1}^2+\sigma_{0,1}^2}}, \frac{\mu_{a,2}-\mu_{0,2}}{\sqrt{\sigma_{a,2}^2+\sigma_{0,2}^2}}, \frac{\mu_{a,1}-\mu_{0,1}}{\sqrt{\sigma_{a,1}^2+\sigma_{0,1}^2}} + \frac{\mu_{a,2}-\mu_{0,2}}{\sqrt{\sigma_{a,2}^2+\sigma_{0,2}^2}}\right\} \\
&= \arg\max_{a\in\{1,2\}} \min\left\{\frac{\mu_{a,1}-\mu_{0,1}}{\sqrt{\sigma_{a,1}^2+\sigma_{0,1}^2}}, \frac{\mu_{a,2}-\mu_{0,2}}{\sqrt{\sigma_{a,2}^2+\sigma_{0,2}^2}}\right\}.
\end{aligned}$$

Comparing this to the definition of a_1^*, we can easily conclude that $a_1^* = a_2^*$ in the large T_{v} regime, and the two problem formulations are equivalent.

C.3 Relaxing Assumptions

We discuss intuitions and methods to relax our assumptions in the main text, such as known variances, Gaussianity, and independence. We remark that these assumptions are natural in the multi-armed bandit literature when developing algorithms for new models, such as this paper.

Non-Existence of Uniformly Better Treatment. It is possible that no treatment is better than control in all metrics simultaneously, and it would be ideal to detect this circumstance as quickly as possible in the exploration phase. A counterfactual evaluation over the samples from the recommended treatment and the control could be performed after the exploration phase to decide the confidence of whether the recommended treatment is better than the control. For each metric, the probability of failure of A/B validation could be calculated as characterized in our proof for Proposition 3.1. If the overall failure probability is too high, we could decide not to move to A/B

validation and declare the non-existence of a universally better treatment. Even if we indeed move to A/B in this setting, the validation itself serves as a second protective mechanism to prevent a worse-than-control arm slip through. A numerical experiment on this null setting is also conducted in Appendix A.

Unknown Treatment Variance. Even though knowing the variance of a treatment is a common assumption in developing bandit algorithms from theoretical causes, especially under novel problems with unique challenges, such as our paper, it is indeed in practice that the true variances of treatments are largely unknown. An adaptation we use is to use a small portion of the budget to estimate the variances beforehand and treat the estimated variances as the true variances. We are aware of this challenge, so we numerically tested `SHRVar` with this adaptation in Fig. 4 of the appendix, which shows almost no performance loss compared to the known variance case. We believe the variance estimation error is likely to be secondary in most applications, and we also find this method works well in real-world applications. In practical use cases, we observe that different metrics, say revenue versus click rate, can have very different variances, where revenue is subject to a large variance while click rate has a much lower variance. For a given metric, different treatments may exhibit varying variances, but these differences are typically of a much smaller magnitude compared to the ones observed across metrics. Therefore, in our work, we prioritize the multi-objective challenge.

Beyond Sub-Gaussian Rewards. Although the Gaussian distribution is a commonly adopted assumption for developing bandit algorithms in new models, such as the multi-metric problem with control in our paper, it is exciting to see if extension beyond these assumptions is possible. We believe it should not pose too much difficulty to generalize `SHRVar` to other distributions based on a number of reasons from intuition. For example, the z-value objective in Proposition 3.1 is derived by characterizing the failure probability of the validation A/B test, which is shown to have a monotonic relation with the z-value. For other distributions, such as the exponential distribution, one could also characterize this failure probability. The z-value objective would stay the same as long as the failure probability has a similar monotonic relation with the z-value. However, the tail characteristic of the reward distribution will affect the error bound in Theorem 5.1, because our proof relies on Hoeffding's concentration, which is not valid beyond a sub-Gaussian distribution. But we should be able to use other concentration methods to prove a similar error bound, for example, using Bernstein's inequality for sub-exponential distributions.

Incorrect Prior. A correct and good prior for Bayesian validation would possibly speed up the validation process, i.e., using fewer samples to produce the required significance. On the other hand, an incorrect prior could also affect the performance of Bayesian validation in our algorithm, similar to all other Bayesian algorithms. How to choose the prior is a rich literature that is complementary to this work. In our paper, if the prior of a metric is picked too large or too small, it would change the z-value of each arm for that metric, and affect the choice of the best treatment since all metrics are related via the min operator in Proposition 3.1. However, this should only be problematic if the validation time horizon T_v is too small, since the prior variance is weighted by $1/T_v$ in our z-value object.

Metric Correlation. It is an interesting topic to extend our paper to settings with metric correlation, which would be our future work. In most A/B tests with multiple metrics, each metric is evaluated individually, so we believe it still makes sense to follow the objective in Eq. 1 to optimize the minimum passing probability over metrics. Then, the problem can still be transformed into the z-value max-min problem in Proposition 3.1, independent of the existence of correlation. If the correlation is known, it will play an important role in our sampling rule to estimate the z-values. We could follow the same framework of this paper to analyze the covariance matrix of the z-estimators and perform an optimal design based on it. Leveraging this correlation would potentially improve the algorithm's performance, resulting in a better error bound, since assuming independence is essentially assuming the worst case. But on the other hand, knowing the metric correlation is as difficult as knowing the variance of each treatment in practice, if not more difficult.

C.4 Other Discussions

Lower Bounds for Fixed-Budget Problems. Developing tight lower bounds for the `M3AB` problem is a challenging but meaningful problem, which we aim to address as our future direction. It is difficult in the sense that, to our best knowledge, no instance-dependent error probability lower bound has been developed in the literature beyond a two-armed Gaussian setting. Even for classic best arm identification with multiple arms and homogeneous variance, such a tight matching lower bound has not been developed, and our studied `M3AB` problem is much more complex, involving multiple treatments, multiple metrics, and a control arm. Some researchers also question whether an achievable lower bound exists in the fixed budget setting (Degenne and Koolen, 2019), where

the authors prove that if there exists a lower bound for the fixed-budget best arm identification problem, there is no single algorithm that can achieve it on every problem instance. And our problem is even more general. On the other hand, the complexity of our proposed `SHRVar` is a generalization of `SH`, one of the best fixed-budget algorithms in single-metric unit-variance problems, and therefore inherits the advantages of `SH`.

Relative Variance and Validation. One of the major contributions of our paper is to formally introduce the validation phase and the control into the best arm identification framework with multiple metrics for consideration. The problem studied in this paper has unique characteristics which makes classic algorithms such as `SH` not applicable, e.g., (i) the control is special and needs to be considered separately in exploration; (ii) the objective shifts from identifying the treatment that maximizes the mean reward to the treatment that maximizes the probability of passing validation, which reflects a mean-variance trade-off; (iii) multi-metric. Some of the aspects have been studied separately (while others, such as the control, are overlooked), but the combination received little attention. The algorithm under this general setting requires a systematic design instead of collecting separate solutions to each aspect. Our starting point is to maximize the probability of passing validation. We systematically and jointly analyzed the exploration and validation phases and reduced them to identify the treatment with the best mean-variance trade-off, i.e., the largest z value. By studying the z-value estimate with the variance equalization allocation principle, the relative variance naturally arises, and we combine the novel relative-variance-based allocation with a sequential halving framework to propose `SHRVar`. This algorithm is provably efficient and strictly generalizes well-known algorithms such as sequential halving in classic BAI. It also connects to real-world applications, as we cannot control the downstream validation, where the experimentation time and passing criteria are set agnostically.

Importance Differences of Metrics. It is common in many real-world applications, where some metrics (say revenue) are "must-haves," and some (say engagement rate or downstream impact) are "nice-to-haves." If some metrics are less important, one could adjust the parameters of the A/B test related to this metric to make it easier to pass, for example, increase the confidence level δ_i or decrease the required posterior p_i for this metric. If some metrics are more important, one could also adjust the same parameter to make it harder to pass. There may be more than one treatment that surpasses all the constraints. Our current algorithm will pick the best one that is most distinguishable from all constraints, which may not necessarily be the one with the highest value in the "must-have" metric. This is a valuable extension for our future work.

Technical Challenges and Contributions. The high-level framework of designing `SHRVar` is rooted in the adaptive experimental design literature, where the sampling rule strives to equalize the variance of the estimators for different treatments, and the proof of Theorem 5.1 follows a general framework presented Karnin et al. (2013) to prove the theoretical performance of `SH`. However, many technical subtleties need to be addressed before applying this framework:

1. Performing BAI based on our objective $\min_i z_{a,i}$ has two unique challenges, i.e., not knowing the bottleneck metric i_a^* makes it difficult to analyze the variance, and the estimator of each $z_{a,i}$ may also be correlated due to the common control. Therefore, we seek surrogates to decouple the estimators by equalizing a variance upper bound and extending from single-metric to multi-metric settings as discussed in Section 4.

2. In the proof of Theorem 5.1, we generalize the idea from the H_2 complexity for `SH` (Karnin et al., 2013) for heterogeneous and multi-metric cases to form the notion of effective gap, relative variance, and set S_c' to capture the trade-off between gap and variance of each treatment through the complexity term H_3. This step is not direct due to the heterogeneous nature of the problem. As a comparison, for `SHVar` proposed in Lalitha et al. (2023), which also studies BAI in heterogeneous variance cases, only uses the minimum reward gap and the sum of variances for all arms, and thus fails to capture the reward-variance trade-off in each arm. Moreover, due to the correlation between the estimators of treatments and metrics, we analyze the concentration events more carefully instead of taking a direct union bound over arms and metrics (which would result in additional dependence on M and A outside the exponential upper bound).

D Related Works

Adaptive experimental design, as a form of online sequential decision-making problems, has been studied in various contexts, e.g., regret minimization (Auer et al., 2002; Auer and Ortner, 2010) and pure exploration (Audibert and Bubeck, 2010; Garivier and Kaufmann, 2016) for multi-armed bandits, Bayesian optimization, and active

learning. Our work is similar to the best arm identification formulation, where a decision on recommendation is made following an exploration phase. This section reviews existing papers related to our problem formulation.

D.1 Pure Exploration Objectives

Most pure exploration problems strive to model a typical product selection pipeline where an exploration phase is conducted to experiment with the quality of multiple objects (arms or actions). After the exploration, one arm (or multiple products) with the desired quality, measured by the reward, is chosen to launch and handed over to future pipelines. For different practical scenarios, the notion of the desired arm(s) varies in the literature, and we review these notions:

Best Arm Identification. The classic best arm identification problem (Audibert and Bubeck, 2010; Garivier and Kaufmann, 2016) considers only the best arm as the desired arm for recommendation, i.e., the arm with the highest ground-truth reward among all candidates. The best arm identification problem is majorly divided into two strings according to the optimization criteria: the fixed confidence setting (Jamieson and Nowak, 2014; Garivier and Kaufmann, 2016; Zhang and Ying, 2023) where the goal is to minimize the sample complexity (number of rounds) before identifying the best arm with confidence higher than a pre-specified threshold $1 - \delta$, and the fixed budget setting (Bubeck et al., 2011; Audibert and Bubeck, 2010; Russo, 2016; Karnin et al., 2013; Abbasi-Yadkori et al., 2018) where the goal is to minimize the probability of identification error with a fixed number of interactions. Even though the fixed-confidence problem is better studied, algorithms developed under this setting, such as TAS (Garivier and Kaufmann, 2016), are rarely implemented in practice. On the other hand, successive halving (Karnin et al., 2013) and the top two algorithms (Russo, 2016; Jourdan et al., 2022), which arose from the fixed-budget setting, are more popular in practice.

Top Arms Identification. In some scenarios, the production pipeline requires more than one arm to be recommended, which motivated the top K arms identification problem (Kalyanakrishnan and Stone, 2010; Kalyanakrishnan et al., 2012; Kaufmann et al., 2016; Chen et al., 2017a; Zhou and Tian, 2022) where the goal is to identify the set of all K arms with reward ranking from one to K. Even though the top K arms identification problem generalizes best arm identification with broadened practical insights, the algorithms developed under this setting usually mimic best arm identification algorithms such as TAS or successive halving.

ε-Good Arm Identification. In the case where the reward of the best arm is very close to the second-best arm, it usually takes a tremendous number of samples to identify the exact best arm, while the quality improvement is small. Therefore, an intuitive bypass is to settle with a good enough arm, i.e., an arm that has a reward close to the best arm, which motivates the ε-Good Arm Identification problem (Mason et al., 2020; Katz-Samuels and Jamieson, 2020; Zhao et al., 2023). In this problem, the objective is to settle with an arm (or all arms) whose reward is within the ε gap to the best arm. A slightly different problem of similar insight is to find an arm (or all arms) with a reward larger than some fraction of the best arm, which is studied under the level set estimation literature with implicit threshold level (Gotovos et al., 2013). Both formulations require the knowledge of ε for algorithm design and typically stop exploration as soon as the ambiguity between arms is less than this threshold.

Good Arm Identification. The good arm identification problem is an even weaker notion than ε-good arm identification, where here the objective is simply to identify one or all arms that have a reward larger than some fixed threshold (Kano et al., 2019; Degenne and Koolen, 2019), and output them as quickly as possible. In this problem, the exploration and exploitation trade-off lies within the arms that are close to the threshold, i.e., whether to pull an arm that is close to the threshold to reduce confusion or to pull an empirical good arm to increase confidence. Different algorithms have been proposed to solve this problem with anytime characterization (Jourdan and Réda, 2023), data-driven threshold (Tsai et al., 2023), and small gaps settings (Tsai et al., 2024). (Kaufmann et al., 2018) studied an easier problem to identify the existence of a good arm with a reward larger than some threshold. The good arm identification problem is internally closely related to two similar problems, i.e., thresholding bandits (Locatelli et al., 2016; Mukherjee et al., 2017; Ouhamma et al., 2021) and level set estimation with explicit level threshold (Bryan et al., 2005; Gotovos et al., 2013). In thresholding bandits, the goal is to classify the arms into two sets according to the ground truth reward, one with a reward higher than the threshold and one with a reward lower than the threshold. The simple regret performance of the thresholding bandit is also studied in (Tao et al., 2019). The level set estimation problem also intends to classify the input space (arm space) into two sets, one with a function value (reward) larger than the threshold and the other with a function value lower than the threshold. However, level set estimation problems usually deal with continuous input spaces or discredited versions of them (Shekhar and Javidi, 2019; Ngo et al., 2024). Moreover, the function

value is usually assumed to follow a Gaussian process prior (Gotovos et al., 2013), and the algorithm to solve it will make decisions based on the acquisition function (Zanette et al., 2019; Senadeera et al., 2020; Bogunovic et al., 2016), which requires the knowledge of this Gaussian process before computing. The results are also generalized with heterogeneous variance scenarios (Inatsu et al., 2019; Iwazaki et al., 2020). Attempts to reduce the Gaussian process prior to approximations have also been made. In particular, (Mason et al., 2022) proposed an RKHS approximation, and (Ha et al., 2021) used neural networks as an approximation. The theoretical limits of level set estimation are characterized in (Bachoc et al., 2021). Both problems are deeply connected with the broad active learning literature and have applications in classification tasks (Jain and Jamieson, 2019).

Better Arm Than Control. To the best of our knowledge, the only work in the literature that studies recommending an arm with a higher reward than the control arm is (Russac et al., 2021), where the goal is to identify an arm better than the known control arm with reward performance averaged over multiple populations. The sampling strategy distinguishing control and treatment arms is also studied. However, their developed algorithm is inherent to TAS for best arm identification, which requires large computational resources. Moreover, their recommended arm is agnostic to the variance and thus may result in poor validation power. Thirdly, their approach towards multiple metrics is a simple weighted sum, while we aim for the robust worst-case metric.

D.2 Bandits with Multiple Metrics

Our work is also closely related to bandits with multiple objectives since we consider multiple reward metrics. In this model, the arm is associated with a joint high-dimensional distribution that generates a vector of rewards upon arm pulls. We also review the literature on multi-objective bandits.

Pareto Front Identification. The notion of the best arm is vague in the context of multi-dimensional reward, and therefore a large body of works study the Pareto optimality framework where the goal is to identify all arms that lie on the Pareto front in terms of the groud-truth reward (Drugan and Nowe, 2013; Auer et al., 2016; Yahyaa et al., 2014b,a; Crepon et al., 2024). (Van Moffaert et al., 2014) studies the Pareto front identification problem in continuous action settings, and (Drugan and Nowe, 2014) uses a scalarization approach to identify the Pareto front, which is further generalized to infinite horizon bandits (Drugan and Manderick, 2015). Our goal is not to identify the Pareto front. Instead, we identify the one better-than-control arm that has the most testing power against it.

Scalarization: Another approach to defining the best arm with multi-dimensional reward is scalarization with a pre-specified utility function. (Roijers et al., 2017) defines the best arm as a weighted linear combination of different dimensions of the ground-truth reward vector, and then performs best arm identification under the linear transformation. However, the linear weight vector is usually hard to pre-define in practice, and thus (Jeunen et al., 2024) proposes a data-driven method to learn the weight vector. Beyond linear scalarization, (Ararat and Tekin, 2023) proposes to scalarize based on a preference over the dimension, which reduces the problem to a hierarchy of best arm identification problems. A completely data-driven scalarization method is also used in (Zintgraf et al., 2018) where the utility function is learned from preference data. Our approach falls into the general category of scalarization to treat multiple metrics. However, different from the above works, the scalarization criterion depends on the ground-truth reward and thus is unknown to us, and we have no access to auxiliary data as in (Jeunen et al., 2024) for utility function learning. Another similar line of work is feasible arm identification (Katz-Samuels and Scott, 2018), a generalization of good arm identification to multi-dimensional rewards. Here, an arm is considered feasible if the ground-truth reward lies in a known polyhedron, so the scalarization from high dimension to feasibility also depends on the reward of each arm. Follow-up work (Katz-Samuels and Jamieson, 2020) generalizes the problem into the large arm regime, with application to recommendation systems. Our setting is also different from theirs in several ways. First, they don't apply the preference to feasible arms and intend to identify all of them, while we prefer to identify the one that has the most testing power against the constraint set. Second, their constraint set is known, while ours is determined by a control arm with an unknown reward, which is more challenging since the sampling strategy of the control arm should also be considered.

Constrained Optimization: The most widely studied approach to multi-dimensional reward is to view one reward metric as the primary reward metric for optimization while viewing the other reward metrics as constraints that should exhibit quality higher (or lower) than some threshold. Constrained regret minimization has been well-studied in the bandit literature (Pacchiano et al., 2021; Liu et al., 2021), but on the other hand, constrained best arm identification has received less attention. The earliest best arm identification problem considering

constraints during the exploration phase is bandits with knapsacks (Badanidiyuru et al., 2018; Ding et al., 2013), where the exploration is subject to a hard resource budget constraint and therefore requires careful planning. (Kanarios et al., 2024) studied a slightly easier problem, relaxing the hard constraint, which intends to find the best arm while minimizing the constraint violation during exploration.

Constrained Fixed-Confidence Bandits: In the fixed confidence setting, (Chen et al., 2016) studied the best arm identification problem with matroid constraints, and (Katz-Samuels and Scott, 2019) studied the polyhedron constraint. (Wang et al., 2022) studies the constrained best arm identification problem, where, in addition to each arm, a dose-level modeling clinical trial is also required for selection during exploration and is related to the safety constraint. The goal is to find the best safe arm if given the proper dose level. Moreover, the safety constraint is imposed both during exploration and the final arm it identifies. (Camilleri et al., 2022; Shang et al., 2023) studied a generalized linear bandit best arm identification problem with safety constraints on the exploration phase. (Carlsson et al., 2024) studied the problem of identifying the best mixture of arms in the presence of constraints, instead of a single arm, since an arm mixture could potentially have a better performance quality. Best arm identification problems with more practical constraints, such as variance constraints (Hou et al., 2023), risk constraints (David et al., 2018), fairness constraints (Wu et al., 2023), and resource constraints (Li and Chi Cheung, 2024), are also studied. Another line of work that presents a similar objective is safe Bayesian optimization (Sui et al., 2015, 2018), where the goal is to maximize the function value $f(x)$ satisfying the constraint $g(x) < h$ in the fixed-confidence setting with the presence of Gaussian noise. Similar to level set estimation, most works in this field assume the functions are generated from a known Gaussian process, which provides prior information for the arm identification and enables the construction of an acquisition function (Berkenkamp et al., 2016, 2023). Future studies generalize this field of work to improve the exploration efficiency (Turchetta et al., 2019), and to the continuous action (Bottero et al., 2022) and high-dimensional input settings (Kirschner et al., 2019; Duivenvoorden et al., 2017). The performance of safe Bayesian optimization algorithms is also validated in empirical studies (Gelbart et al., 2014; Hernández-Lobato et al., 2016; Gardner et al., 2014). (Bogunovic et al., 2016) studied the similarity between Bayesian optimization and level set estimation, and proposed a unified framework.

Constrained Fixed-Budget Bandits. The fixed-budget setting is less studied compared to the fixed-confidence counterpart. The earliest paper that studies constrained best arm identification in this setting is (Chang, 2020) using an epsilon-greedy algorithm for exploration, which is later shown to be inefficient. An adaptation of successive rejection is proposed in (Faizal and Nair, 2022), which finds the best single arm under constraint, and (Tang et al., 2024) generalized the algorithm to tackle the identification of support of best arm mixtures. (Lindner et al., 2022) studied a slightly different problem where they assume a known reward but an unknown constraint, and the goal is to learn the constraint while identifying the best safe arm. Our work is different from constrained optimization since there is no primary reward metric.

D.3 Best Arm Identification Formulations

Considering the large amount of literature studying online sequential learning with different objectives and formulations, we review the works that present similar goals to our paper.

Min-Max Optimization: The minimax optimization was studied largely in the robust optimization literature (Bertsimas et al., 2010; Chen et al., 2017b) and the game theory literature (Cai and Daskalakis, 2011; Nouiehed et al., 2019), where a natural adversary agnostic to the agent is assumed to pick the worst environment for the agent to perform optimization. This assumption coincides with the risk-averse nature of many applications. The min-max objective is also studied in many works related to multi-armed bandits with structure. For example, (Wang and Scarlett, 2022) assumed the arms belonged to certain groups and intended to maximize the worst arm performance in a group. (Garivier et al., 2016; Marchesi et al., 2019) assumed the reward of each arm would also depend on the action of an adversary and thus formulated a min-max game bandit problem. The objective is to pick the best arm robust to the worst selection of the adversary. Beyond bandit literature, the min-max objective is also commonly studied in multi-class classifications (Nouiehed et al., 2019) and other machine learning settings.

Heterogeneous Variance: One characterization that makes our problem interesting is the heterogeneous variance nature of arms and reward metrics, so we review the best arm identification with fixed-budget literature for works dealing with heterogeneous variance. (Lalitha et al., 2023) generalized the sequential halving algorithm to the heterogeneous variance setting and proposed a variance-aware sampling strategy. If the variance is unknown to the practitioner, they also proposed an unbiased estimation with theoretical performance guarantees.

(Atsidakou et al., 2022) studied the Bayesian bandit setting with heterogeneous variance to understand the relationship between performance and prior. Another line of work that typically studies the influence of variance is multi-fidelity bandits (Poiani et al., 2022, 2024), motivated by the simulation experiments with heterogeneous qualities, where the goal is to identify the experimental result while controlling the cost incurred by using high-fidelity simulators. The heterogeneous variance effect is also studied in generalized adaptive experimental design (Weltz et al., 2023; Lalitha et al., 2023), and other active learning scenarios (Antos et al., 2010; Chaudhuri et al., 2017).

E Proof of Proposition 3.1

We first give the explicit forms of the posterior mean and variances $\hat{\Delta}_{\mathrm{v},i}$ and $\hat{\sigma}^2_{\mathrm{v},i}$ in the Bayesian validation as follows. The explicit expression can be derived easily from (Atsidakou et al., 2022, equation 2)

$$\hat{\sigma}^2_{\mathrm{v},i} = \left(\frac{T_{\mathrm{v}}}{2\left(\sigma^2_{\hat{a},i} + \sigma^2_{0,i}\right)} + \frac{1}{\tau_i^2} \right)^{-1}, \quad \hat{\Delta}_{\mathrm{v},i} = \frac{T_{\mathrm{v}} \hat{\sigma}^2_{\mathrm{v},i}}{2\left(\sigma^2_{\hat{a},i} + \sigma^2_{0,i}\right)} \left(\hat{\mu}_{\mathrm{v}}(\hat{a}, i) - \hat{\mu}_{\mathrm{v}}(0, i)\right). \tag{6}$$

Then, we prove the proposition. In both Bayesian and non-Bayesian validation, we fix a reward metric i and characterize the probability that treatment a will pass the validation A/B test against control, i.e., $\mathcal{E}^{\mathtt{C}}_{\mathrm{v},i}$ when $\hat{a} = a$. We first start with the non-Bayesian validation. Recall that passing the validation means the difference in the empirical mean should be large enough, i.e.,

$$\mathcal{E}_{\mathrm{v},i} = \left\{ \hat{\mu}_{\mathrm{v}}(a, i) - \hat{\mu}_{\mathrm{v}}(0, i) \geq \Phi^{-1}(1 - \delta_i) \sqrt{\frac{\sigma^2_{a,i} + \sigma^2_{0,i}}{T_{\mathrm{v}}/2}} \right\}.$$

Since the reward follows a normal distribution, the empirical estimate will also be normal, and therefore, the probability of failure can be characterized using the cumulative distribution function $\Phi(\cdot)$ of standard normal as follows:

$$\begin{aligned}
\mathbb{P}\left(\mathcal{E}^{\mathtt{C}}_{\mathrm{v},i}\right) =& \mathbb{P}\left(\hat{\mu}_{\mathrm{v}}(a, i) - \hat{\mu}_{\mathrm{v}}(0, i) \leq \Phi^{-1}(1 - \delta_i) \sqrt{\frac{\sigma^2_{a,i} + \sigma^2_{0,i}}{T_{\mathrm{v}}/2}} \right) \\
=& \Phi\left(\frac{\Phi^{-1}(1 - \delta_i) \sqrt{\frac{\sigma^2_{a,i} + \sigma^2_{0,i}}{T_{\mathrm{v}}/2}} - (\mu_{a,i} - \mu_{0,i})}{\sqrt{\mathrm{Var}(\hat{\mu}_{\mathrm{v}}(a, i) - \hat{\mu}_{\mathrm{v}}(0, i))}} \right) \\
=& \Phi\left(\Phi^{-1}(1 - \delta_i) - \frac{\mu_{a,i} - \mu_{0,i}}{\sqrt{\sigma^2_{a,i} + \sigma^2_{0,i}}} \sqrt{T_{\mathrm{v}}/2} \right).
\end{aligned}$$

By the monotonicity of $\Phi(\cdot)$, the treatment that minimizes the probability of failure should:

$$\begin{aligned}
\arg\min_{a \in \mathcal{A}/\{0\}} \max_i \mathbb{P}\left(\mathcal{E}^{\mathtt{C}}_{\mathrm{v},i}\right) &= \arg\min_{a \in \mathcal{A}/\{0\}} \max_i \Phi\left(\Phi^{-1}(1 - \delta_i) - \frac{\mu_{a,i} - \mu_{0,i}}{\sqrt{\sigma^2_{a,i} + \sigma^2_{0,i}}} \sqrt{T_{\mathrm{v}}/2} \right) \\
&= \arg\min_{a \in \mathcal{A}/\{0\}} \max_i \left(\Phi^{-1}(1 - \delta_i) - \frac{\mu_{a,i} - \mu_{0,i}}{\sqrt{\sigma^2_{a,i} + \sigma^2_{0,i}}} \sqrt{T_{\mathrm{v}}/2} \right) \\
&= \arg\max_{a \in \mathcal{A}/\{0\}} \min_i \left(\frac{\mu_{a,i} - \mu_{0,i}}{\sqrt{\sigma^2_{a,i} + \sigma^2_{0,i}}} + \frac{\Phi^{-1}(\delta_i)}{\sqrt{T_{\mathrm{v}}/2}} \right).
\end{aligned}$$

where the last step uses the anti-symmetry of $\Phi(\cdot)$, i.e., $\Phi(x) = 1 - \Phi(-x)$. Therefore, the validation constant in the non-Bayesian test is:

$$\xi_{a,i} = \frac{\Phi^{-1}(\delta_i)}{\sqrt{T_{\mathrm{v}}/2}}.$$

Next, we derive the validation constant $\xi_{a,i}$ in the Bayesian validation. We also fix a reward metric i and characterize the probability that treatment a will pass the validation A/B test against control, i.e., $\mathcal{E}^{\complement}_{\mathrm{v},i}$ when $\hat{a} = a$. Recall that the criteria for passing the validation are that the posterior probability of $\{\mu_{a,i} > \mu_{0,i}\}$ is larger than a given threshold q_i. Since the posterior of $(\mu_{a,i} - \mu_{0,i})$ is a normal distribution, i.e., $\nu_{\mathrm{v},i} = \mathcal{N}(\hat{\Delta}_{\mathrm{v},i}, \hat{\sigma}^2_{\mathrm{v},i})$, this is equivalent to that the cumulative distribution function of $\nu_{\mathrm{v},i}$, evaluated at 0, is smaller than $1 - q_i$, which is also equivalent to the following event:

$$\left\{\Phi\left(\frac{-\hat{\Delta}_{\mathrm{v},i}}{\hat{\sigma}_{\mathrm{v},i}}\right) \le 1 - q_i\right\} = \left\{\frac{-\hat{\Delta}_{\mathrm{v},i}}{\hat{\sigma}_{\mathrm{v},i}} \le \Phi^{-1}(1 - q_i)\right\}.$$

where the equality is from the monotonicity of $\Phi(\cdot)$. Substitute the expression of the posterior mean $\hat{\Delta}_{\mathrm{v},i}$ and variance $\hat{\sigma}^2_{\mathrm{v},i}$, we have:

$$\begin{aligned}
\mathcal{E}_{\mathrm{v},i} &= \left\{\frac{\frac{T_{\mathrm{v}}\hat{\sigma}^2_{\mathrm{v},i}}{2\left(\sigma^2_{a,i}+\sigma^2_{0,i}\right)}\left(\hat{\mu}_{\mathrm{v}}(a,i) - \hat{\mu}_{\mathrm{v}}(0,i)\right)}{\hat{\sigma}_{\mathrm{v},i}} \ge -\Phi^{-1}(1 - q_i)\right\} \\
&= \left\{\hat{\mu}_{\mathrm{v}}(a,i) - \hat{\mu}_{\mathrm{v}}(0,i) \ge \frac{2\left(\sigma^2_{a,i} + \sigma^2_{0,i}\right)}{T_{\mathrm{v}}\hat{\sigma}_{\mathrm{v},i}}\Phi^{-1}(q_i)\right\}.
\end{aligned}$$

Similarly, by the normal nature of the empirical estimate of reward, we can express the probability of failure in validation as follows:

$$\begin{aligned}
\mathbb{P}\left(\mathcal{E}^{\complement}_{\mathrm{v},i}\right) =& \mathbb{P}\left(\hat{\mu}_{\mathrm{v}}(a,i) - \hat{\mu}_{\mathrm{v}}(0,i) \le \frac{2\left(\sigma^2_{a,i} + \sigma^2_{0,i}\right)}{T_{\mathrm{v}}\hat{\sigma}_{\mathrm{v},i}}\Phi^{-1}(q_i)\right) \\
=& \Phi\left(\frac{\Phi^{-1}(q_i)\frac{2\left(\sigma^2_{a,i}+\sigma^2_{0,i}\right)}{T_{\mathrm{v}}\hat{\sigma}_{\mathrm{v},i}} - (\mu_{a,i} - \mu_{0,i})}{\sqrt{\mathrm{Var}(\hat{\mu}_{\mathrm{v}}(a,i) - \hat{\mu}_{\mathrm{v}}(0,i))}}\right) \\
=& \Phi\left(\frac{\Phi^{-1}(q_i)\frac{2\left(\sigma^2_{a,i}+\sigma^2_{0,i}\right)}{T_{\mathrm{v}}\hat{\sigma}_{\mathrm{v},i}} - (\mu_{a,i} - \mu_{0,i})}{\sqrt{\frac{\sigma^2_{a,i}+\sigma^2_{0,i}}{T_{\mathrm{v}}/2}}}\right) \\
=& \Phi\left(\Phi^{-1}(q_i)\sqrt{\frac{2}{T_{\mathrm{v}}}\frac{\sigma^2_{a,i} + \sigma^2_{0,i}}{\hat{\sigma}^2_{\mathrm{v},i}}} - \frac{\mu_{a,i} - \mu_{0,i}}{\sqrt{\sigma^2_{a,i} + \sigma^2_{0,i}}}\sqrt{T_{\mathrm{v}}/2}\right).
\end{aligned}$$

where we substitute the expression of $\hat{\sigma}^2_{\mathrm{v},i}$ as follows:

$$\hat{\sigma}^2_{\mathrm{v},i} = \left(\frac{T_{\mathrm{v}}}{2\left(\sigma^2_{\hat{a},i} + \sigma^2_{0,i}\right)} + \frac{1}{\tau_i^2}\right)^{-1},$$

And then we have:

$$\mathbb{P}\left(\mathcal{E}^{\complement}_{\mathrm{v},i}\right) = \Phi\left(\Phi^{-1}(q_i)\sqrt{1 + \frac{2}{T_{\mathrm{v}}}\frac{\sigma^2_{a,i} + \sigma^2_{0,i}}{\tau_i^2}} - \frac{\mu_{a,i} - \mu_{0,i}}{\sqrt{\sigma^2_{a,i} + \sigma^2_{0,i}}}\sqrt{T_{\mathrm{v}}/2}\right).$$

Therefore, similarly by the monotonicity and anti-symmetry of $\Phi(\cdot)$, our objective becomes:

$$\arg\min_{a\in\mathcal{A}/\{0\}} \max_i \mathbb{P}\left(\mathcal{E}^{\complement}_{\mathrm{v},i}\right) = \arg\max_{a\in\mathcal{A}/\{0\}} \min_i \left(\frac{\mu_{a,i}-\mu_{0,i}}{\sqrt{\sigma^2_{a,i}+\sigma^2_{0,i}}} + \frac{\Phi^{-1}(1-q_i)}{\sqrt{T_{\mathrm{v}}/2}} \sqrt{1+\frac{2}{T_{\mathrm{v}}}\frac{\sigma^2_{a,i}+\sigma^2_{0,i}}{\tau_i^2}} \right),$$

where the validation constant in the Bayesian test is:

$$\xi_{a,i} = \frac{\Phi^{-1}(1-q_i)}{\sqrt{T_{\mathrm{v}}/2}} \sqrt{1+\frac{\sigma^2_{a,i}+\sigma^2_{0,i}}{\tau_i^2(T_v/2)}}.$$

F Proof of Theorem 5.1

Suppose at stage s and fix the active treatment set $\mathcal{A}_s$, we first analyze the probability that the best treatment is eliminated. This can only happen if there exists at least $|\mathcal{A}_s|$ arms with a larger minimum empirical z value than the best treatment, i.e.,

$$\left\{ \sum_{a\in\mathcal{A}_s} \mathbb{1}_{\min_i \hat{z}_s(a,i) > \min_j \hat{z}_s(a^*,j)} \geq \frac{|\mathcal{A}_s|}{2} \right\}.$$

Therefore, it is essential to study the event $\{\min_i \hat{z}_s(a,i) > \min_j \hat{z}_s(a^*,j)\}$ for all sub-optimal arms. We first present the following lemma, which bounds the probability that the minimum z-value estimate of a suboptimal treatment a would be larger than the best treatment.

Lemma F.1. *Fix a stage s and the active treatment set $\mathcal{A}_s$, we have:*

$$\begin{aligned}
&\mathbb{P}\Big(\min_i \hat{z}_s(a,i) > \min_j \hat{z}_s(a^*,j) \Big) \\
\leq& 2M \exp\left(- \frac{T}{2\left(\rho_{\mathcal{A}_s,\Sigma}+\lambda_{\mathcal{A}_s,\Sigma}\right)^2 \log_2 A} \cdot \min_{j\in[M]} \max_{z\leq z_{a^*,j}} \min \left\{ \frac{[z_{a^*,j}-z]_+^2}{\kappa^2_{\mathcal{A}_s,a^*,j}}, \sum_{i:z_{a,i}<z_{a^*,j}} \frac{[z-z_{a,i}]_+^2}{\kappa^2_{\mathcal{A}_s,a,i}} \right\} \right),
\end{aligned}$$

where $(\rho_{\mathcal{A}_s,\Sigma}, \lambda_{\mathcal{A}_s,\Sigma})$ measures the relative variance of the active treatment set $\mathcal{A}_s$, i.e.,

$$\rho^2_{\mathcal{A}_s,\Sigma} = \sum_{a\in\mathcal{A}_s} \max_i \rho^2_{a,i}, \quad \lambda^2_{\mathcal{A}_s,\Sigma} = \max_{a\in\mathcal{A}_s} \max_i \lambda^2_{a,i};$$

and $\kappa^2_{\mathcal{A}_s,a,i}$ measures the heterogeneity of relative variance of arm a and metric i, i.e.,

$$\kappa^2_{\mathcal{A}_s,a,i} = \frac{\rho^2_{a,i}}{\max_i \rho^2_{a,i}} \frac{\rho_{\mathcal{A}_s,\Sigma}}{\rho_{\mathcal{A}_s,\Sigma}+\lambda_{\mathcal{A}_s,\Sigma}} + \frac{\lambda^2_{a,i}}{\max_{a\in\mathcal{A}_s}\max_i \lambda^2_{a,i}} \frac{\lambda_{\mathcal{A}_s,\Sigma}}{\rho_{\mathcal{A}_s,\Sigma}+\lambda_{\mathcal{A}_s,\Sigma}}.$$

Proof. We first analyze the event and separate the reward metrics, i.e.,

$$\{\min_i \hat{z}_s(a,i) > \min_j \hat{z}_s(a^*,j)\} = \bigcup_{j\in[M]} \left\{ \min_i \hat{z}_s(a,i) > \hat{z}_s(a^*,j) \right\} = \bigcup_{j\in[M]} \bigcap_{i\in[M]} \{\hat{z}_s(a,i) > \hat{z}_s(a^*,j)\}.$$

Now, we analyze the intersection of events for a fixed j. First, we ignore the reward metrics i where $z_{a,i} \geq z_{a^*,j}$, and upper bound the event as:

$$\bigcap_{i\in[M]} \{\hat{z}_s(a,i) > \hat{z}_s(a^*,j)\} \subset \bigcap_{i:z_{a,i}<z_{a^*,j}} \{\hat{z}_s(a,i) > \hat{z}_s(a^*,j)\}.$$

For an arbitrary constant $z \leq z_{a^*,j}$, we could separate the two estimates as follows:

$$\begin{aligned}
\{\hat{z}_s(a,i) > \hat{z}_s(a^*,j)\} =& \{\hat{z}_s(a,i) > \hat{z}_s(a^*,j), \hat{z}_s(a^*,j) < z\} \cup \{\hat{z}_s(a,i) > \hat{z}_s(a^*,j), \hat{z}_s(a^*,j) \geq z\} \\
\subset& \{\hat{z}_s(a^*,j) < z\} \cup \{\hat{z}_s(a,i) > z\}.
\end{aligned}$$

So we can upper bound the intersection events as follows:

$$\bigcap_{i:z_{a,i}<z_{a^*,j}} \{\hat{z}_s(a,i) > \hat{z}_s(a^*,j)\} \subset \bigcap_{i:z_{a,i}<z_{a^*,j}} (\{\hat{z}_s(a^*,j) < z\} \cup \{\hat{z}_s(a,i) > z\})$$
$$= \{\hat{z}_s(a^*,j) < z\} \cup \left(\bigcap_{i:z_{a,i}<z_{a^*,j}} \{\hat{z}_s(a,i) > z\} \right).$$

Then, we evaluate the probability of this event using a union bound as follows:

$$\mathbb{P}\left(\bigcap_{i:z_{a,i}<z_{a^*,j}} \{\hat{z}_s(a,i) > \hat{z}_s(a^*,j)\} \right) \leq \mathbb{P}\left(\hat{z}_s(a^*,j) < z\right) + \mathbb{P}\left(\bigcap_{i:z_{a,i}<z_{a^*,j}} \{\hat{z}_s(a,i) > z\} \right)$$
$$= \mathbb{P}\left(\hat{z}_s(a^*,j) < z\right) + \prod_{i:z_{a,i}<z_{a^*,j}} \mathbb{P}\left(\hat{z}_s(a,i) > z\right),$$

where the second equality uses the independence among metrics. For an arbitrary constant z, we can bound both probabilities through Hoeffding's inequality as:

$$\mathbb{P}\left(\hat{z}_s(a^*,j) < z\right) \leq \exp\left(-\frac{[z_{a^*,j} - z]_+^2}{2\mathrm{Var}(\hat{z}_s(a^*,j))} \right),$$

where the variance of the z value estimate can be computed as:

$$\begin{aligned}
\mathrm{Var}(\hat{z}_s(a^*,j)) =& \frac{\mathrm{Var}(\hat{\mu}_s(a^*,j)) + \mathrm{Var}(\hat{\mu}_s(0,j))}{\sigma_{a^*,j}^2 + \sigma_{0,j}^2} \\
=& \left(\sqrt{\sum_{a\in\mathcal{A}_s} \max_i \rho_{a,i}^2} + \max_{a\in\mathcal{A}_s} \max_i \lambda_{a,i} \right) \frac{\log_2 A}{T} \left(\frac{\rho_{a^*,j}^2}{\max_i \rho_{a^*,i}^2} \sqrt{\sum_{a\in\mathcal{A}_s} \max_i \rho_{a,i}^2} + \frac{\lambda_{a^*,j}^2}{\max_{a\in\mathcal{A}_s} \max_i \lambda_{a,i}} \right) \\
=& \left(\rho_{\mathcal{A}_s,\Sigma} + \lambda_{\mathcal{A}_s,\Sigma}\right) \frac{\log_2 A}{T} \left(\frac{\rho_{a^*,j}^2}{\max_i \rho_{a^*,i}^2} \rho_{\mathcal{A}_s,\Sigma} + \frac{\lambda_{a^*,j}^2}{\max_{a\in\mathcal{A}_s} \max_i \lambda_{a,i}^2} \lambda_{\mathcal{A}_s,\Sigma} \right) \\
=& \left(\rho_{\mathcal{A}_s,\Sigma} + \lambda_{\mathcal{A}_s,\Sigma}\right)^2 \frac{\log_2 A}{T} \kappa_{\mathcal{A}_s,a^*,j}^2.
\end{aligned}$$

So the probability is bounded as follows:

$$\mathbb{P}\left(\hat{z}_s(a^*,j) < z\right) \leq \exp\left(-\frac{[z_{a^*,j} - z]_+^2}{2\left(\rho_{\mathcal{A}_s,\Sigma} + \lambda_{\mathcal{A}_s,\Sigma}\right)^2 \kappa_{\mathcal{A}_s,a^*,j}^2} \frac{T}{\log_2 A} \right),$$

Similarly, for the probability regarding treatment a, we could also bound them as:

$$\mathbb{P}\left(\hat{z}_s(a,i) > z\right) \leq \exp\left(-\frac{[z - z_{a,i}]_+^2}{2\left(\rho_{\mathcal{A}_s,\Sigma} + \lambda_{\mathcal{A}_s,\Sigma}\right)^2 \kappa_{\mathcal{A}_s,a,i}^2} \frac{T}{\log_2 A} \right).$$

Therefore, the intersection of events is bounded using independence as follows:

$$\begin{aligned}
&\mathbb{P}\left(\bigcap_{i:z_{a,i}<z_{a^*,j}} \{\hat{z}_s(a,i) > \hat{z}_s(a^*,j)\} \right) \\
\leq& \exp\left(-\frac{T}{2\left(\rho_{\mathcal{A}_s,\Sigma} + \lambda_{\mathcal{A}_s,\Sigma}\right)^2 \log_2 A} \frac{[z_{a^*,j} - z]_+^2}{\kappa_{\mathcal{A}_s,a^*,j}^2} \right) + \exp\left(-\frac{T}{2\left(\rho_{\mathcal{A}_s,\Sigma} + \lambda_{\mathcal{A}_s,\Sigma}\right)^2 \log_2 A} \sum_{i:z_{a,i}<z_{a^*,j}} \frac{[z - z_{a,i}]_+^2}{\kappa_{\mathcal{A}_s,a,i}^2} \right) \\
\leq& 2\exp\left(-\frac{T}{2\left(\rho_{\mathcal{A}_s,\Sigma} + \lambda_{\mathcal{A}_s,\Sigma}\right)^2 \log_2 A} \min\left\{ \frac{[z_{a^*,j} - z]_+^2}{\kappa_{\mathcal{A}_s,a^*,j}^2}, \sum_{i:z_{a,i}<z_{a^*,j}} \frac{[z - z_{a,i}]_+^2}{\kappa_{\mathcal{A}_s,a,i}^2} \right\} \right).
\end{aligned}$$

Since z is arbitrary chosen satisfying $z \leq z_{a^*,j}$. Therefore, choosing z to minimize the exponent, we have:

$$\begin{aligned}
&\mathbb{P}\left(\bigcap_{i:z_{a,i}<z_{a^*,j}} \{\hat{z}_s(a,i) > \hat{z}_s(a^*,j)\}\right) \\
\leq& 2\exp\left(-\frac{T}{2\left(\rho_{\mathcal{A}_s,\Sigma}+\lambda_{\mathcal{A}_s,\Sigma}\right)^2 \log_2 A} \max_{z\leq z_{a^*,j}} \min\left\{\frac{[z_{a^*,j}-z]_+^2}{\kappa^2_{\mathcal{A}_s,a^*,j}}, \sum_{i:z_{a,i}<z_{a^*,j}} \frac{[z-z_{a,i}]_+^2}{\kappa^2_{\mathcal{A}_s,a,i}}\right\}\right).
\end{aligned}$$

Finally, from a union bound, we will be able to bound the probability that a sub-optimal treatment has a better minimum z-value estimate than the best treatment as follows:

$$\begin{aligned}
&\mathbb{P}\Big(\min_i \hat{z}_s(a,i) > \min_j \hat{z}_s(a^*,j)\Big) \\
\leq& \sum_{j\in[M]} \mathbb{P}\left(\min_i \hat{z}_s(a,i) > \hat{z}_s(a^*,j)\right) \\
\leq& 2\sum_{j\in[M]} \exp\left(-\frac{T}{2\left(\rho_{\mathcal{A}_s,\Sigma}+\lambda_{\mathcal{A}_s,\Sigma}\right)^2 \log_2 A} \max_{z\leq z_{a^*,j}} \min\left\{\frac{[z_{a^*,j}-z]_+^2}{\kappa^2_{\mathcal{A}_s,a^*,j}}, \sum_{i:z_{a,i}<z_{a^*,j}} \frac{[z-z_{a,i}]_+^2}{\kappa^2_{\mathcal{A}_s,a,i}}\right\}\right) \\
\leq& 2M\exp\left(-\frac{T}{2\left(\rho_{\mathcal{A}_s,\Sigma}+\lambda_{\mathcal{A}_s,\Sigma}\right)^2 \log_2 A} \min_{j\in[M]} \max_{z\leq z_{a^*,j}} \min\left\{\frac{[z_{a^*,j}-z]_+^2}{\kappa^2_{\mathcal{A}_s,a^*,j}}, \sum_{i:z_{a,i}<z_{a^*,j}} \frac{[z-z_{a,i}]_+^2}{\kappa^2_{\mathcal{A}_s,a,i}}\right\}\right).
\end{aligned}$$

This concludes the proof. □

Before we proceed, we first take a deeper look at the exponent of the probability bound, specifically, the instance-dependent hardness constant, which measures how difficult to distinguish a sub-optimal treatment from the optimal treatment as follows:

$$\min_{j\in[M]} \max_{z\leq z_{a^*,j}} \min\left\{\frac{[z_{a^*,j}-z]_+^2}{\kappa^2_{\mathcal{A}_s,a^*,j}}, \sum_{i:z_{a,i}<z_{a^*,j}} \frac{[z-z_{a,i}]_+^2}{\kappa^2_{\mathcal{A}_s,a,i}}\right\}.$$

We can replace the summation with any reward metric i, which results in the following lower bound:

$$\min_{j\in[M]} \max_{i\in[M]} \max_{z\leq z_{a^*,j}} \min\left\{\frac{[z_{a^*,j}-z]_+^2}{\kappa^2_{\mathcal{A}_s,a^*,j}}, \frac{[z-z_{a,i}]_+^2}{\kappa^2_{\mathcal{A}_s,a,i}}\right\} = \min_{j\in[M]} \max_{i\in[M]} \frac{[z_{a^*,j}-z_{a,i}]_+^2}{\left(\kappa_{\mathcal{A}_s,a,i}+\kappa_{\mathcal{A}_s,a^*,j}\right)^2}.$$

For simplicity of notation, we use the complexity notation as follows:

$$D_{\mathcal{A}_s,a} = \min_{i\in[M]} D_{\mathcal{A}_s,a,i} = \min_{i\in[M]} \max_{j\in[M]} \frac{[z_{a^*,i}-z_{a,j}]_+^2}{\left(\kappa_{\mathcal{A}_s,a,j}+\kappa_{\mathcal{A}_s,a^*,i}\right)^2}.$$

Then, we bound the probability that the algorithm makes a mistake and excludes the best arm on stage s, summarized in the following lemma.

Lemma F.2. *Fix a stage s and suppose the best treatment a^* has not been eliminated in previous stages, the probability that it is eliminated at this stage is bounded as follows:*

$$\mathbb{P}\left(a^* \notin \mathcal{A}_{s+1}\right) \leq 6M\exp\left(-\frac{T}{2\log_2 A} \min_{\mathcal{S}\subset\mathcal{A}:a^*\in\mathcal{S},|\mathcal{S}|=\frac{|\mathcal{A}|}{2^{s-1}}} \frac{\min_{a\in\mathcal{S}'_c} D^2_{\mathcal{S},a}}{\left(\rho_{\mathcal{S},\Sigma}+\lambda_{\mathcal{S},\Sigma}\right)^2}\right),$$

where $\mathcal{S}$ is an arbitrary subset of all treatments that contains the best arm with a fixed cardinality, and $\mathcal{S}'_c$ is the subset of $\mathcal{S}$ defined as follows:

- *When $|\mathcal{S}| \geq 4$, $\mathcal{S}'_c$ is the subset excluding at most $\frac{1}{4}|\mathcal{S}|$ treatments with minimum $D_{\mathcal{S},a}$.*
- *When $|\mathcal{S}| \leq 3$, $\mathcal{S}'_c$ is the subset of $\mathcal{S}$ containing all sub-optimal treatments.*

Proof. Fix an arbitrary $\mathcal{A}_s$ which contains the best treatment a^*. Suppose it has at least 4 treatments. Let $\mathcal{A}'_s$ as defined be the set of treatments in $\mathcal{A}_s$ excluding at most $\frac{1}{4}|\mathcal{A}_s|$ treatments with minimum $D_{\mathcal{A}_s,a}$. If the best treatment is eliminated on round s, it must be the case that at least half of the treatments in $\mathcal{A}_s$ have an empirical minimum z value larger than the best treatment. In particular, at least $\frac{1}{3}$ of the treatments in $\mathcal{A}'_s$ will have a larger empirical minimum z value. So we have:

$$\begin{aligned}\mathbb{E}\left[\sum_{a\in\mathcal{A}'_s}\mathbb{1}_{\min_i \hat{z}_s(a,i)>\min_j \hat{z}_s(a^*,j)}\right] =& \sum_{a\in\mathcal{A}'_s}\mathbb{P}\left(\min_i \hat{z}_s(a,i) > \min_j \hat{z}_s(a^*,j)\right)\\ \leq& 2M\sum_{a\in\mathcal{A}'_s}\exp\left(-\frac{T}{2\log_2 A}\frac{D^2_{\mathcal{A}_s,a}}{\left(\rho_{\mathcal{A}_s,\Sigma}+\lambda_{\mathcal{A}_s,\Sigma}\right)^2}\right)\\ \leq& 2M|\mathcal{A}'_s|\exp\left(-\frac{T}{2\log_2 A}\frac{\min_{a\in\mathcal{A}'_s} D^2_{\mathcal{A}_s,a}}{\left(\rho_{\mathcal{A}_s,\Sigma}+\lambda_{\mathcal{A}_s,\Sigma}\right)^2}\right).\end{aligned}$$

Therefore, by Markov's inequality, we will be able to establish an upper bound for the event that the best treatment is eliminated:

$$\begin{aligned}\mathbb{P}\left(a^*\notin\mathcal{A}_{s+1}\right) =& \mathbb{P}\left(\sum_{a\in\mathcal{A}'_s}\mathbb{1}_{\min_i \hat{z}_s(a,i)>\min_j \hat{z}_s(a^*,j)} \geq \frac{|\mathcal{A}'_s|}{3}\right)\\ \leq& \frac{3}{|\mathcal{A}'_s|}\mathbb{E}\left[\sum_{a\in\mathcal{A}'_s}\mathbb{1}_{\min_i \hat{z}_s(a,i)>\min_j \hat{z}_s(a^*,j)}\right]\\ \leq& 6M\exp\left(-\frac{T}{2\log_2 A}\frac{\min_{a\in\mathcal{A}'_s} D^2_{\mathcal{A}_s,a}}{\left(\rho_{\mathcal{A}_s,\Sigma}+\lambda_{\mathcal{A}_s,\Sigma}\right)^2}\right).\end{aligned}$$

However, this bound is valid for a fixed $\mathcal{A}_s$ which contains the best treatment, and is not valid if $\mathcal{A}_s$ is a stochastic set. Therefore, we take a maximum over the probability upper bound shown above to make it compatible with stochastic sets as follows:

$$\mathbb{P}\left(a^*\notin\mathcal{A}_{s+1}\right) \leq 6M\exp\left(-\frac{T}{2\log_2 A}\min_{\mathcal{S}:a^*\in\mathcal{S},|\mathcal{S}|=\frac{|\mathcal{A}|}{2^{s-1}}}\frac{\min_{a\in\mathcal{S}'_c} D^2_{\mathcal{S},a}}{\left(\rho_{\mathcal{S},\Sigma}+\lambda_{\mathcal{S},\Sigma}\right)^2}\right),$$

where $\mathcal{S}'_c$ is defined in the statement of lemma. □

Finally, to prove Theorem 5.1, the best treatment a^* will only be recommended at the end of the exploration phase if it survives all $\log_2 A$ stages. With a union bound, we have:

$$\begin{aligned}\mathbb{P}\left(\hat{a}\neq a^*\right) \leq& \sum_{s=1}^{\log_2 A}\mathbb{P}\left(a^* \text{ is eliminated at stage } s\right)\\ \leq& 6M\sum_{s=1}^{\log_2 A}\exp\left(-\frac{T}{2\log_2 A}\min_{\mathcal{S}:a^*\in\mathcal{S},|\mathcal{S}|=\frac{|\mathcal{A}|}{2^{s-1}}}\frac{\min_{a\in\mathcal{S}'_c} D^2_{\mathcal{S},a}}{\left(\rho_{\mathcal{S},\Sigma}+\lambda_{\mathcal{S},\Sigma}\right)^2}\right)\\ \leq& 6M\log_2 A\cdot\exp\left(-\frac{T}{2\log_2 A}\min_{\mathcal{S}:a^*\in\mathcal{S}}\frac{\min_{a\in\mathcal{S}'_c} D^2_{\mathcal{S},a}}{\left(\rho_{\mathcal{S},\Sigma}+\lambda_{\mathcal{S},\Sigma}\right)^2}\right).\end{aligned}$$

Replacing $\rho_{\mathcal{S},\Sigma}$ and $\lambda_{\mathcal{S},\Sigma}$ with their definitions respectively, we obtain Theorem 5.1.

G Proof of Theorem 5.2

The proof of the confidence-based elimination strategy is similar to the proof of Theorem 5.1 except that a new upper bound for the probability that some sub-optimal treatment a will have a larger confidence $\delta_s(a)$ than the

best treatment needs to be established. So following the proof of Theorem 5.1, suppose we fix a stage s and the active treatment set $\mathcal{A}_s$, we analyze the probability that a treatment a has a larger confidence $\delta_s(a)$ than the best treatment a^* in the following lemma, i.e.,

Lemma G.1. *Fix a stage s and the active treatment set $\mathcal{A}_s$, we have:*

$$\mathbb{P}\left(\delta_s(a) > \delta_s(a^*)\right) \leq 2M \exp\left(-\frac{T}{8\log_2 A\left(\rho_{\mathcal{A}_s,\Sigma} + \lambda_{\mathcal{A}_s,\Sigma}\right)^2}\tilde{D}^2_{\mathcal{A}_s,a}\right),$$

where $(\rho_{\mathcal{A}_s,\Sigma}, \lambda_{\mathcal{A}_s,\Sigma})$ measures the relative variance of the active treatment set $\mathcal{A}_s$, i.e.,

$$\rho^2_{\mathcal{A}_s,\Sigma} = \sum_{a\in\mathcal{A}_s} \max_i \rho^2_{a,i}, \quad \lambda^2_{\mathcal{A}_s,\Sigma} = \max_{a\in\mathcal{A}_s}\max_i \lambda^2_{a,i};$$

and $\kappa^2_{\mathcal{A}_s,a,i}$ measures the heterogeneity of relative variance of arm a and metric i, i.e.,

$$\kappa^2_{\mathcal{A}_s,a,i} = \frac{\rho^2_{a,i}}{\max_i \rho^2_{a,i}}\frac{\rho_{\mathcal{A}_s,\Sigma}}{\rho_{\mathcal{A}_s,\Sigma}+\lambda_{\mathcal{A}_s,\Sigma}} + \frac{\lambda^2_{a,i}}{\max_{a\in\mathcal{A}_s}\max_i \lambda^2_{a,i}}\frac{\lambda_{\mathcal{A}_s,\Sigma}}{\rho_{\mathcal{A}_s,\Sigma}+\lambda_{\mathcal{A}_s,\Sigma}}.$$

The instance-dependent complexity $\tilde{D}^2_{\mathcal{A}_s,a}$ measures the hardness of identifying the best treatment a^* against the sub-optimal treatment a, which is defined as: if $\kappa_{\mathcal{A}_s,a,j} > \kappa_{\mathcal{A}_s,a^*,i}$,

$$\tilde{D}^2_{\mathcal{A}_s,a,i,j} = \min\left\{\frac{[z_{a^*,i}-z_{a,j}]^2_+}{(\kappa_{\mathcal{A}_s,a,j}+\kappa_{\mathcal{A}_s,a^*,i})^2}, \frac{[z_{a^*,i}-z_{a,j}]^2_+}{\left(\kappa_{\mathcal{A}_s,a,j}-\kappa_{\mathcal{A}_s,a^*,i}\right)^2} + \Delta^2_{\min} - \frac{8A\log_2^2 A}{T}\right\},$$

And if $\kappa_{\mathcal{A}_s,a,j} \leq \kappa_{\mathcal{A}_s,a^*,i}$,

$$\tilde{D}^2_{\mathcal{A}_s,a,i,j} = \frac{[z_{a^*,i}-z_{a,j}]^2_+}{(\kappa_{\mathcal{A}_s,a,j}+\kappa_{\mathcal{A}_s,a^*,i})^2}.$$

And we define:

$$\tilde{D}^2_{\mathcal{A}_s,a} = \min_i \tilde{D}^2_{\mathcal{A}_s,a,i} = \min_{i\in[M]}\max_{j\in[M]} \tilde{D}^2_{\mathcal{A}_s,a,i,j}.$$

Proof. Recall the definition of confidence $\delta_s(a)$, if $\delta_s(a) > \delta_s(a^*)$, it must be the case where

$$\mathrm{UCB}^{\delta_s(a^*)}(a^*) = \max_{a'\in\mathcal{A}_s}\mathrm{LCB}^{\delta_s(a^*)}_s(a') \leq \max_{a'\in\mathcal{A}_s}\mathrm{LCB}^{\delta_s(a)}_s(a') = \mathrm{UCB}^{\delta_s(a)}(a) \leq \mathrm{UCB}^{\delta_s(a^*)}(a),$$

where the inequality is because the lower confidence bound $\mathrm{LCB}^{\delta}_s(a)$ is monotonically non-decreasing concerning the confidence δ. The second inequality is because the upper confidence bound $\mathrm{UCB}^{\delta}_s(a)$ is monotonically non-increasing with the confidence δ. Then, this will induce the following event:

$$\begin{aligned}
\{\delta_s(a) > \delta_s(a^*)\} &\subset \left\{\min_i\left\{\hat{z}_s(a,i) + b^{\delta_s(a^*)}_s(a,i)\right\} \geq \min_j\left\{\hat{z}_s(a^*,j) + b^{\delta_s(a^*)}_s(a^*,j)\right\}\right\} \\
&= \bigcup_{j\in[M]}\left\{\min_i\left\{\hat{z}_s(a,i) + b^{\delta_s(a^*)}_s(a,i)\right\} \geq \hat{z}_s(a^*,j) + b^{\delta_s(a^*)}_s(a^*,j)\right\} \\
&= \bigcup_{j\in[M]}\bigcap_{i\in[M]}\underbrace{\left\{\hat{z}_s(a,i) + b^{\delta_s(a^*)}_s(a,i) \geq \hat{z}_s(a^*,j) + b^{\delta_s(a^*)}_s(a^*,j)\right\}}_{\mathcal{E}_{j,i}}.
\end{aligned}$$

Then, it suffices to analyze each event. We first take a deeper look at the confidence bonus with the relative-variance-based sampling strategy as follows:

$$b^{\delta_s(a^*)}_s(a,i) = 2\sqrt{\left(\frac{\rho^2_{a,i}}{N_s(a)} + \frac{\lambda^2_{a,i}}{N_s(0)}\right)\log\left(\frac{|\mathcal{A}_s|M}{\delta}\right)}.$$

Specifically, we look at the variance terms as follows:

$$\begin{aligned}\frac{\rho_{a,i}^2}{N_s(a)}+\frac{\lambda_{a,i}^2}{N_s(0)} =& \frac{\log_2 A}{T}\left(\sqrt{\sum_{a\in\mathcal{A}_s}\max_i \rho_{a,i}^2}+\max_{a\in\mathcal{A}_s}\max_i \lambda_{a,i}\right)\left(\frac{\rho_{a,i}^2}{\max_i \rho_{a,i}^2}\sqrt{\sum_{a\in\mathcal{A}_s}\max_i \rho_{a,i}^2}+\frac{\lambda_{a,i}^2}{\max_{a\in\mathcal{A}_s}\max_i \lambda_{a,i}}\right)\\ =& \frac{\log_2 A}{T}\left(\rho_{\mathcal{A}_s,\Sigma}+\lambda_{\mathcal{A}_s,\Sigma}\right)\left(\frac{\rho_{a,i}^2}{\max_i \rho_{a,i}^2}\rho_{\mathcal{A}_s,\Sigma}+\frac{\lambda_{a,i}^2}{\max_{a\in\mathcal{A}_s}\max_i \lambda_{a,i}^2}\lambda_{\mathcal{A}_s,\Sigma}\right)\\ =& \frac{\log_2 A}{T}\left(\rho_{\mathcal{A}_s,\Sigma}+\lambda_{\mathcal{A}_s,\Sigma}\right)^2\kappa_{\mathcal{A}_s,a,i}^2,\end{aligned}$$

where we use the definition:

$$\kappa_{\mathcal{A}_s,a,i}^2=\frac{\rho_{a,i}^2}{\max_i \rho_{a,i}^2}\frac{\rho_{\mathcal{A}_s,\Sigma}}{\rho_{\mathcal{A}_s,\Sigma}+\lambda_{\mathcal{A}_s,\Sigma}}+\frac{\lambda_{a,i}^2}{\max_{a\in\mathcal{A}_s}\max_i \lambda_{a,i}^2}\frac{\lambda_{\mathcal{A}_s,\Sigma}}{\rho_{\mathcal{A}_s,\Sigma}+\lambda_{\mathcal{A}_s,\Sigma}}.$$

Then, let $\hat{c}_s(a)=\sqrt{\log\left(\frac{|\mathcal{A}_s|M}{\delta_s(a)}\right)}$ and $\Gamma_s^2=\frac{\log_2 A}{T}\left(\rho_{\mathcal{A}_s,\Sigma}+\lambda_{\mathcal{A}_s,\Sigma}\right)^2$, each event $\mathcal{E}_{j,i}$ becomes:

$$\mathcal{E}_{j,i}=\left\{\hat{z}_s(a,i)+2\kappa_{\mathcal{A}_s,a,i}\Gamma_s\hat{c}_s(a^*)\geq \hat{z}_s(a^*,j)+2\kappa_{\mathcal{A}_s,a^*,j}\Gamma_s\hat{c}_s(a^*)\right\}.$$

We divide the metrics i into two groups based on the relation between $\kappa_{\mathcal{A}_s,a,i}$ and $\kappa_{\mathcal{A}_s,a^*,j}$.

Case 1. $\kappa_{\mathcal{A}_s,a,i}\leq\kappa_{\mathcal{A}_s,a^*,j}$. For the first group, in order for event $\mathcal{E}_{j,i}$ to hold, we must have:

$$\mathcal{E}_{j,i}\subset\left\{\hat{z}_s(a,i)\geq\hat{z}_s(a^*,j)\right\}.$$

By Hoeffding's inequality, we can bound the probability of this event as follows:

$$\mathbb{P}\left(\hat{z}_s(a,i)\geq\hat{z}_s(a^*,j)\right)\leq\exp\left(-\frac{[z_{a^*,j}-z_{a,i}]_+^2}{2\left(\mathrm{Var}(\hat{z}_s(a,i))+\mathrm{Var}(\hat{z}_s(a^*,j))\right)}\right).$$

The variance of estimate $\hat{z}_s(a^*,j)$ can be analyzed as follows:

$$\begin{aligned}\mathrm{Var}(\hat{z}_s(a^*,j))=&\frac{\mathrm{Var}(\hat{\mu}_s(a^*,j))+\mathrm{Var}(\hat{\mu}_s(0,j))}{\sigma_{a^*,j}^2+\sigma_{0,j}^2}\\ =&\left(\sqrt{\sum_{a\in\mathcal{A}_s}\max_i \rho_{a,i}^2}+\max_{a\in\mathcal{A}_s}\max_i \lambda_{a,i}\right)\frac{\log_2 A}{T}\left(\frac{\rho_{a^*,j}^2}{\max_i \rho_{a^*,i}^2}\sqrt{\sum_{a\in\mathcal{A}_s}\max_i \rho_{a,i}^2}+\frac{\lambda_{a^*,j}^2}{\max_{a\in\mathcal{A}_s}\max_i \lambda_{a,i}}\right)\\ =&\left(\rho_{\mathcal{A}_s,\Sigma}+\lambda_{\mathcal{A}_s,\Sigma}\right)\frac{\log_2 A}{T}\left(\frac{\rho_{a^*,j}^2}{\max_i \rho_{a^*,i}^2}\rho_{\mathcal{A}_s,\Sigma}+\frac{\lambda_{a^*,j}^2}{\max_{a\in\mathcal{A}_s}\max_i \lambda_{a,i}^2}\lambda_{\mathcal{A}_s,\Sigma}\right)\\ =&\left(\rho_{\mathcal{A}_s,\Sigma}+\lambda_{\mathcal{A}_s,\Sigma}\right)^2\frac{\log_2 A}{T}\kappa_{\mathcal{A}_s,a^*,j}^2.\end{aligned}$$

Similarly, the variance of estimate $\hat{z}_s(a,i)$ can be analyzed as follows:

$$\mathrm{Var}(\hat{z}_s(a,i))=\left(\rho_{\mathcal{A}_s,\Sigma}+\lambda_{\mathcal{A}_s,\Sigma}\right)^2\frac{\log_2 A}{T}\kappa_{\mathcal{A}_s,a,i}^2.$$

Therefore, we have the following bound:

$$\begin{aligned}\mathbb{P}\left(\mathcal{E}_{j,i}\right)\leq&\mathbb{P}\left(\hat{z}_s(a,i)\geq\hat{z}_s(a^*,j)\right)\\ \leq&\exp\left(-\frac{T}{2\log_2 A\left(\rho_{\mathcal{A}_s,\Sigma}+\lambda_{\mathcal{A}_s,\Sigma}\right)^2}\frac{[z_{a^*,j}-z_{a,i}]_+^2}{\kappa_{\mathcal{A}_s,a,i}^2+\kappa_{\mathcal{A}_s,a^*,j}^2}\right)\\ \leq&\exp\left(-\frac{T}{2\log_2 A\left(\rho_{\mathcal{A}_s,\Sigma}+\lambda_{\mathcal{A}_s,\Sigma}\right)^2}\frac{[z_{a^*,j}-z_{a,i}]_+^2}{\left(\kappa_{\mathcal{A}_s,a,i}+\kappa_{\mathcal{A}_s,a^*,j}\right)^2}\right)\\ =&\exp\left(-\frac{T}{2\log_2 A\left(\rho_{\mathcal{A}_s,\Sigma}+\lambda_{\mathcal{A}_s,\Sigma}\right)^2}D_{\mathcal{A}_s,a,j,i}^2\right).\end{aligned}$$

Case 2. $\kappa_{\mathcal{A}_s,a,i} > \kappa_{\mathcal{A}_s,a^*,j}$. For the second group, if $z_{a^*,j} - z_{a,i} < 0$, we can simply bound the probability of this event by 1. Assume $z_{a^*,j} > z_{a,i}$, and we can bound the event $\mathcal{E}_{j,i}$ as:

$$\mathcal{E}_{j,i} \subset \underbrace{\left\{\hat{z}_s(a,i) - \hat{z}_s(a^*,j) \geq -\frac{z_{a^*,j} - z_{a,i}}{2}\right\}}_{\mathcal{E}_{j,i,1}} \cup \underbrace{\left\{\hat{c}_s(a^*) \geq \frac{z_{a^*,j} - z_{a,i}}{4\left(\kappa_{\mathcal{A}_s,a,i} - \kappa_{\mathcal{A}_s,a^*,j}\right)\Gamma_s}\right\}}_{\mathcal{E}_{j,i,2}}.$$

The probability of the first event $\mathcal{E}_{j,i,1}$ can be bounded with Hoeffding's inequality, similar to Case 1, as follows:

$$\begin{aligned}
\mathbb{P}\left(\hat{z}_s(a,i) - \hat{z}_s(a^*,j) \geq -\frac{z_{a^*,j} - z_{a,i}}{2}\right) &\leq \exp\left(-\frac{\left(\frac{z_{a^*,j} - z_{a,i}}{2}\right)^2}{2\left(\mathrm{Var}(\hat{z}_s(a,i)) + \mathrm{Var}(\hat{z}_s(a^*,j))\right)}\right) \\
&\leq \exp\left(-\frac{T}{8\log_2 A\left(\rho_{\mathcal{A}_s,\Sigma} + \lambda_{\mathcal{A}_s,\Sigma}\right)^2}\frac{[z_{a^*,j} - z_{a,i}]_+^2}{(\kappa_{\mathcal{A}_s,a,i} + \kappa_{\mathcal{A}_s,a^*,j})^2}\right) \\
&= \exp\left(-\frac{T}{8\log_2 A\left(\rho_{\mathcal{A}_s,\Sigma} + \lambda_{\mathcal{A}_s,\Sigma}\right)^2} D^2_{\mathcal{A}_s,a,j,i}\right).
\end{aligned}$$

The second event $\mathcal{E}_{j,i,2}$ induces a lower bound on the confidence to eliminate the best treatment a^*. Therefore, there must exist another arm a', whose lower confidence bound is larger than the upper confidence bound of a^* when the confidence bonus $b_s^{\delta}(a,i)$ is constructed with $\delta = \delta_s(a^*)$, i.e., under $\mathcal{E}_{j,i,2}$, we have:

$$\left\{\exists a' \in \mathcal{A}_s, \min_i \left\{\hat{z}_s(a',i) - b_s^{\delta_s(a^*)}(a',i)\right\} \geq \min_{j'}\left\{\hat{z}_s(a^*,j') + b_s^{\delta_s(a^*)}(a^*,j')\right\}\right\}.$$

Notice that in event $\mathcal{E}_{j,i,2}$, reward metric j is the metric which achieves the minimum in the right-hand side, so we can further rewrite this event as:

$$\left\{\exists a' \in \mathcal{A}_s, \forall i \in [M], \hat{z}_s(a',i) - b_s^{\delta_s(a^*)}(a',i) \geq \hat{z}_s(a^*,j) + b_s^{\delta_s(a^*)}(a^*,j)\right\}.$$

Similarly, we start from the event for a fixed a', a fixed metric i' and a fixed j, i.e.,

$$\begin{aligned}
\mathcal{E}'_{j,i'} &= \left\{\hat{z}_s(a',i') - b_s^{\delta_s(a^*)}(a',i') \geq \hat{z}_s(a^*,j) + b_s^{\delta_s(a^*)}(a^*,j)\right\} \\
&= \{\hat{z}_s(a',i') - 2\kappa_{\mathcal{A}_s,a',i'}\Gamma_s\hat{c}_s(a^*) \geq \hat{z}_s(a^*,j) + 2\kappa_{\mathcal{A}_s,a^*,j}\Gamma_s\hat{c}_s(a^*)\} \\
&= \{\hat{z}_s(a',i') - \hat{z}_s(a^*,j) \geq 2\left(\kappa_{\mathcal{A}_s,a',i'} + \kappa_{\mathcal{A}_s,a^*,j}\right)\Gamma_s\hat{c}_s(a^*)\}.
\end{aligned}$$

Then, under event $\mathcal{E}_{j,i,2}$, we can lower bound the additional confidence based bonus term $\hat{c}_s(a^*)$ and obtain:

$$\mathcal{E}'_{j,i'} \subset \left\{\hat{z}_s(a',i') - \hat{z}_s(a^*,j) \geq \frac{\kappa_{\mathcal{A}_s,a',i'} + \kappa_{\mathcal{A}_s,a^*,j}}{2\left(\kappa_{\mathcal{A}_s,a,i} - \kappa_{\mathcal{A}_s,a^*,j}\right)}(z_{a^*,j} - z_{a,i})\right\}.$$

From Hoeffding's inequality, the probability of this event can be bounded by:

$$\mathbb{P}\left(\mathcal{E}'_{j,i}\right) \leq \exp\left(-\frac{\left(\frac{\kappa_{\mathcal{A}_s,a',i'} + \kappa_{\mathcal{A}_s,a^*,j}}{2\left(\kappa_{\mathcal{A}_s,a,i} - \kappa_{\mathcal{A}_s,a^*,j}\right)}(z_{a^*,j} - z_{a,i}) + [z_{a^*,j} - z_{a',i'}]_+\right)^2}{2\left(\mathrm{Var}(\hat{z}_s(a',i')) + \mathrm{Var}(\hat{z}_s(a^*,j))\right)}\right),$$

where, following similar arguments from previous analysis, we can derive:

$$\begin{aligned}
\mathrm{Var}(\hat{z}_s(a',i')) + \mathrm{Var}(\hat{z}_s(a^*,j)) &= \left(\rho_{\mathcal{A}_s,\Sigma} + \lambda_{\mathcal{A}_s,\Sigma}\right)^2 \frac{\log_2 A}{T}\left(\kappa^2_{\mathcal{A}_s,a^*,j} + \kappa^2_{\mathcal{A}_s,a',i'}\right) \\
&\leq \left(\rho_{\mathcal{A}_s,\Sigma} + \lambda_{\mathcal{A}_s,\Sigma}\right)^2 \frac{\log_2 A}{T}\left(\kappa_{\mathcal{A}_s,a^*,j} + \kappa_{\mathcal{A}_s,a',i'}\right)^2.
\end{aligned}$$

So substitute this result in, we have:

$$\begin{aligned}\mathbb{P}\left(\mathcal{E}'_{j,i}\right) \leq & \exp\left(-\frac{T}{8\log_2 A\left(\rho_{\mathcal{A}_s,\Sigma}+\lambda_{\mathcal{A}_s,\Sigma}\right)^2}\left(\frac{\left(z_{a^*,j}-z_{a,i}\right)^2}{\left(\kappa_{\mathcal{A}_s,a,i}-\kappa_{\mathcal{A}_s,a^*,j}\right)^2}+\frac{\left[z_{a^*,j}-z_{a',i'}\right]_+^2}{\left(\kappa_{\mathcal{A}_s,a^*,j}+\kappa_{\mathcal{A}_s,a',i'}\right)^2}\right)\right)\\ = & \exp\left(-\frac{T}{8\log_2 A\left(\rho_{\mathcal{A}_s,\Sigma}+\lambda_{\mathcal{A}_s,\Sigma}\right)^2}\left(\frac{\left(z_{a^*,j}-z_{a,i}\right)^2}{\left(\kappa_{\mathcal{A}_s,a,i}-\kappa_{\mathcal{A}_s,a^*,j}\right)^2}+D^2_{\mathcal{A}_s,a',j,i'}\right)\right).\end{aligned}$$

Finally, we use a union bound over a' and choose the best i' to minimize the probability upper bound as follows:

$$\begin{aligned}\mathbb{P}\left(\mathcal{E}_{j,i,2}\right) \leq & \sum_{a'\neq a^*}\exp\left(-\frac{T}{8\log_2 A\left(\rho_{\mathcal{A}_s,\Sigma}+\lambda_{\mathcal{A}_s,\Sigma}\right)^2}\left(\frac{\left(z_{a^*,j}-z_{a,i}\right)^2}{\left(\kappa_{\mathcal{A}_s,a,i}-\kappa_{\mathcal{A}_s,a^*,j}\right)^2}+\max_{i'}D^2_{\mathcal{A}_s,a',j,i'}\right)\right)\\ \leq & A\exp\left(-\frac{T}{8\log_2 A\left(\rho_{\mathcal{A}_s,\Sigma}+\lambda_{\mathcal{A}_s,\Sigma}\right)^2}\left(\frac{\left(z_{a^*,j}-z_{a,i}\right)^2}{\left(\kappa_{\mathcal{A}_s,a,i}-\kappa_{\mathcal{A}_s,a^*,j}\right)^2}+\min_{a'\neq a^*}\max_{i'}D^2_{\mathcal{A}_s,a',j,i'}\right)\right).\end{aligned}$$

Notice that the second term:

$$\min_{a'\neq a^*}\max_{i'}D^2_{\mathcal{A}_s,a',j,i'} \geq \min_{a\neq a^*}\max_{i}[z_{a^*,j}-z_{a,i}]_+^2 \geq \min_{a\neq a^*}(z_{a^*,j}-z_{a,i^*})^2 \geq \underbrace{\min_{a\neq a^*}(z_{a^*,i^*}-z_{a,i^*})^2}_{\Delta^2_{\min}}.$$

So the bound can be rewritten as follows:

$$\mathbb{P}\left(\mathcal{E}_{j,i,2}\right) \leq \exp\left(-\frac{T}{8\log_2 A\left(\rho_{\mathcal{A}_s,\Sigma}+\lambda_{\mathcal{A}_s,\Sigma}\right)^2}\left(\frac{\left(z_{a^*,j}-z_{a,i}\right)^2}{\left(\kappa_{\mathcal{A}_s,a,i}-\kappa_{\mathcal{A}_s,a^*,j}\right)^2}+\Delta^2_{\min}-\frac{8A\log_2^2 A}{T}\right)\right).$$

Therefore, we can use a union bound to derive the probability of event $\mathcal{E}_{j,i}$ as follows:

$$\begin{aligned}\mathbb{P}\left(\mathcal{E}_{j,i}\right) \leq & \exp\left(-\frac{T}{8\log_2 A\left(\rho_{\mathcal{A}_s,\Sigma}+\lambda_{\mathcal{A}_s,\Sigma}\right)^2}\frac{\left[z_{a^*,j}-z_{a,i}\right]_+^2}{\left(\kappa_{\mathcal{A}_s,a,i}+\kappa_{\mathcal{A}_s,a^*,j}\right)^2}\right)\\ & +\exp\left(-\frac{T}{8\log_2 A\left(\rho_{\mathcal{A}_s,\Sigma}+\lambda_{\mathcal{A}_s,\Sigma}\right)^2}\left(\frac{\left[z_{a^*,j}-z_{a,i}\right]_+^2}{\left(\kappa_{\mathcal{A}_s,a,i}-\kappa_{\mathcal{A}_s,a^*,j}\right)^2}+\Delta^2_{\min}-\frac{8A\log_2^2 A}{T}\right)\right)\\ \leq & 2\exp\left(-\frac{T}{8\log_2 A\left(\rho_{\mathcal{A}_s,\Sigma}+\lambda_{\mathcal{A}_s,\Sigma}\right)^2}\tilde{D}^2_{\mathcal{A}_s,a,j,i}\right),\end{aligned}$$

where we recall the definition that for i where $\kappa_{\mathcal{A}_s,a,i} > \kappa_{\mathcal{A}_s,a^*,j}$:

$$\tilde{D}^2_{\mathcal{A}_s,a,j,i} = \min\left\{\frac{\left[z_{a^*,j}-z_{a,i}\right]_+^2}{\left(\kappa_{\mathcal{A}_s,a,i}+\kappa_{\mathcal{A}_s,a^*,j}\right)^2}, \frac{\left[z_{a^*,j}-z_{a,i}\right]_+^2}{\left(\kappa_{\mathcal{A}_s,a,i}-\kappa_{\mathcal{A}_s,a^*,j}\right)^2}+\Delta^2_{\min}-\frac{8A\log_2^2 A}{T}\right\}.$$

And for i where $\kappa_{\mathcal{A}_s,a,i} \leq \kappa_{\mathcal{A}_s,a^*,j}$:

$$\tilde{D}^2_{\mathcal{A}_s,a,j,i} = \frac{\left[z_{a^*,j}-z_{a,i}\right]_+^2}{\left(\kappa_{\mathcal{A}_s,a,i}+\kappa_{\mathcal{A}_s,a^*,j}\right)^2}.$$

So, with both cases analyzed, the intersection of events can be bounded with the event that minimizes the probability upper bound, i.e.,

$$\begin{aligned}\mathbb{P}\left(\bigcap_{i\in[M]}\mathcal{E}_{j,i}\right) \leq & \min_{i\in[M]}\mathbb{P}\left(\mathcal{E}_{j,i}\right)\\ = & \min\left\{\min_{i:\kappa_{\mathcal{A}_s,a,i}\leq\kappa_{\mathcal{A}_s,a^*,j}}\mathbb{P}\left(\mathcal{E}_{j,i}\right), \min_{i:\kappa_{\mathcal{A}_s,a,i}>\kappa_{\mathcal{A}_s,a^*,j}}\mathbb{P}\left(\mathcal{E}_{j,i}\right)\right\}\\ \leq & 2\exp\left(-\frac{T}{8\log_2 A\left(\rho_{\mathcal{A}_s,\Sigma}+\lambda_{\mathcal{A}_s,\Sigma}\right)^2}\max_{i\in[M]}\tilde{D}^2_{\mathcal{A}_s,a,j,i}\right).\end{aligned}$$

Then, the probability that a sub-optimal arm has better $\delta_s(a)$ than the optimal arm can be bounded with a union bound as follows:

$$\begin{aligned}\mathbb{P}\left(\delta_s(a) > \delta_s(a^*)\right) \leq & \sum_{j\in[M]} \mathbb{P}\left(\bigcap_{i\in[M]} \mathcal{E}_{j,i}\right) \\ \leq & 2\sum_{j\in[M]} \exp\left(-\frac{T}{8\log_2 A\left(\rho_{\mathcal{A}_s,\Sigma}+\lambda_{\mathcal{A}_s,\Sigma}\right)^2}\max_{i\in[M]}\tilde{D}^2_{\mathcal{A}_s,a,j,i}\right) \\ \leq & 2M\exp\left(-\frac{T}{8\log_2 A\left(\rho_{\mathcal{A}_s,\Sigma}+\lambda_{\mathcal{A}_s,\Sigma}\right)^2}\min_{j\in[M]}\max_{i\in[M]}\tilde{D}^2_{\mathcal{A}_s,a,j,i}\right) \\ = & 2M\exp\left(-\frac{T}{8\log_2 A\left(\rho_{\mathcal{A}_s,\Sigma}+\lambda_{\mathcal{A}_s,\Sigma}\right)^2}\tilde{D}^2_{\mathcal{A}_s,a}\right)\end{aligned}$$

This concludes the proof. □

Then, we bound the probability that the algorithm makes a mistake and excludes the best treatment on stage s, summarized in the following lemma.

Lemma G.2. *Fix a stage s and suppose the best treatment a^* has not been eliminated in previous stages, the probability that it is eliminated at this stage is bounded as follows:*

$$\mathbb{P}\left(a^* \notin \mathcal{A}_{s+1}\right) \leq 6M\exp\left(-\frac{T}{2\log_2 A}\min_{\mathcal{S}\subset\mathcal{A}:a^*\in\mathcal{S},|\mathcal{S}|=\frac{|\mathcal{A}|}{2^{s-1}}}\frac{\min_{a\in\tilde{\mathcal{S}}'_c}\tilde{D}^2_{\mathcal{S},a}}{\left(\rho_{\mathcal{S},\Sigma}+\lambda_{\mathcal{S},\Sigma}\right)^2}\right),$$

where $\mathcal{S}$ is an arbitrary subset of all treatments that contains the best arm with a fixed cardinality, and $\tilde{\mathcal{S}}'_c$ is the subset of $\mathcal{S}$ defined as follows:

- *When $|\mathcal{S}| \geq 4$, $\tilde{\mathcal{S}}'_c$ is the subset excluding at most $\frac{1}{4}|\mathcal{S}|$ treatments with minimum $\tilde{D}_{\mathcal{S},a}$.*
- *When $|\mathcal{S}| \leq 3$, $\tilde{\mathcal{S}}'_c$ is the subset of $\mathcal{S}$ containing all sub-optimal treatments.*

Proof. Fix an arbitrary $\mathcal{A}_s$ which contains the best treatment a^*. Suppose it has at least 4 treatments. Let $\mathcal{A}'_s$ as defined be the set of treatments in $\mathcal{A}_s$ excluding at most $\frac{1}{4}|\mathcal{A}_s|$ treatments with minimum $\tilde{D}_{\mathcal{A}_s,a}$. If the best treatment is eliminated on round s, it must be the case that at least half of the treatments in $\mathcal{A}_s$ have a smaller confidence of mistake $\delta_s(a)$ larger than the best treatment. In particular, at least $\frac{1}{3}$ of the treatments in $\mathcal{A}'_s$ will have a smaller confidence. So we have:

$$\begin{aligned}\mathbb{E}\left[\sum_{a\in\mathcal{A}'_s}\mathbb{1}_{\delta_s(a)>\delta_s(a^*)}\right] = & \sum_{a\in\mathcal{A}'_s}\mathbb{P}\left(\delta_s(a) > \delta_s(a^*)\right) \\ \leq & 2M\sum_{a\in\mathcal{A}'_s}\exp\left(-\frac{T}{8\log_2 A}\frac{\tilde{D}^2_{\mathcal{A}_s,a}}{\left(\rho_{\mathcal{A}_s,\Sigma}+\lambda_{\mathcal{A}_s,\Sigma}\right)^2}\right) \\ \leq & 2M|\mathcal{A}'_s|\exp\left(-\frac{T}{8\log_2 A}\frac{\min_{a\in\mathcal{A}'_s}\tilde{D}^2_{\mathcal{A}_s,a}}{\left(\rho_{\mathcal{A}_s,\Sigma}+\lambda_{\mathcal{A}_s,\Sigma}\right)^2}\right).\end{aligned}$$

Therefore, by Markov's inequality, we will be able to establish an upper bound for the event that the best treatment is eliminated:

$$\begin{aligned}\mathbb{P}\left(a^* \notin \mathcal{A}_{s+1}\right) = & \mathbb{P}\left(\sum_{a\in\mathcal{A}'_s}\mathbb{1}_{\delta_s(a)>\delta_s(a^*)} \geq \frac{|\mathcal{A}'_s|}{3}\right) \leq \frac{3}{|\mathcal{A}'_s|}\mathbb{E}\left[\sum_{a\in\mathcal{A}'_s}\mathbb{1}_{\delta_s(a)>\delta_s(a^*)}\right] \\ \leq & 6M\exp\left(-\frac{T}{8\log_2 A}\frac{\min_{a\in\mathcal{A}'_s}\tilde{D}^2_{\mathcal{A}_s,a}}{\left(\rho_{\mathcal{A}_s,\Sigma}+\lambda_{\mathcal{A}_s,\Sigma}\right)^2}\right).\end{aligned}$$

However, this bound is valid for a fixed $\mathcal{A}_s$ which contains the best treatment, and is not valid if $\mathcal{A}_s$ is a stochastic set. Therefore, we take a maximum over the probability upper bound shown above to make it compatible with stochastic sets as follows:

$$\mathbb{P}\left(a^* \notin \mathcal{A}_{s+1}\right) \leq 6M \exp\left(-\frac{T}{8\log_2 A} \min_{\mathcal{S}: a^* \in \mathcal{S}, |\mathcal{S}| = \frac{|\mathcal{A}|}{2^{s-1}}} \frac{\min_{a \in \tilde{\mathcal{S}}'_c} \tilde{D}^2_{\mathcal{S},a}}{\left(\rho_{\mathcal{S},\Sigma} + \lambda_{\mathcal{S},\Sigma}\right)^2}\right),$$

where $\tilde{\mathcal{S}}'_c$ is defined in the statement of lemma. □

Finally, to prove Theorem 5.2, the best treatment a^* will only be recommended at the end of the exploration phase if it survives all $\log_2 A$ stages. With a union bound, we have:

$$\begin{aligned}
\mathbb{P}\left(\hat{a} \neq a^*\right) \leq & \sum_{s=1}^{\log_2 A} \mathbb{P}\left(a^* \text{ is eliminated at stage } s\right) \\
\leq & 6M \sum_{s=1}^{\log_2 A} \exp\left(-\frac{T}{8\log_2 A} \min_{\mathcal{S}: a^* \in \mathcal{S}, |\mathcal{S}| = \frac{|\mathcal{A}|}{2^{s-1}}} \frac{\min_{a \in \tilde{\mathcal{S}}'_c} \tilde{D}^2_{\mathcal{S},a}}{\left(\rho_{\mathcal{S},\Sigma} + \lambda_{\mathcal{S},\Sigma}\right)^2}\right) \\
\leq & 6M \log_2 A \cdot \exp\left(-\frac{T}{8\log_2 A} \min_{\mathcal{S}: a^* \in \mathcal{S}} \frac{\min_{a \in \tilde{\mathcal{S}}'_c} \tilde{D}^2_{\mathcal{S},a}}{\left(\rho_{\mathcal{S},\Sigma} + \lambda_{\mathcal{S},\Sigma}\right)^2}\right).
\end{aligned}$$

Replacing $\rho_{\mathcal{S},\Sigma}$ and $\lambda_{\mathcal{S},\Sigma}$ with their definitions respectively, we obtain Theorem 5.2.