Worst-Case Optimal Multi-Armed Gaussian Best Arm Identification with a Fixed Budget

Anonymous authors
Paper under double-blind review

Abstract

Experimental design is crucial in evidence-based decision-making with multiple treatment arms, such as online advertisements and medical treatments. This study investigates the problem of identifying the treatment arm with the highest expected outcome, referred to as *the best treatment arm*, while minimizing the probability of misidentification. This problem has been studied across numerous research fields, including *best arm identification* (BAI) and ordinal optimization. In our experiments, the number of treatment-allocation rounds is fixed. During each round, a decision-maker allocates a treatment arm to an experimental unit and observes a corresponding outcome, which follows a Gaussian distribution with variances that can differ among the treatment arms. At the end of the experiment, we recommend one of the treatment arms as an estimate of the best treatment arm based on the observations. To design an experiment, we first discuss lower bounds for the probability of misidentification through an information-theoretic approach. Our analysis highlights that the available information on the outcome distribution for each treatment arm, such as means (expected outcomes), variances, and the choice of the best treatment arm, significantly influences the lower bounds. In scenarios where available information is limited, we develop a lower bound that are valid under the unknown means and the unknown choice of the best treatment arm, which are referred to as the worst-case lower bound. We demonstrate that the worst-case lower bound depends solely on the variances of the outcomes. Then, under the assumption that the variances are known, we propose the *Generalized-Neyman-Allocation (GNA)-empirical-best-arm (EBA) strategy*, an extension of the Neyman allocation proposed by Neyman (1934). We show that the GNA-EBA strategy is asymptotically optimal in the sense that its probability of misidentification aligns with the lower bounds as the sample size increases indefinitely and the differences between the expected outcomes of the best and other suboptimal arms converge to a uniform value. We refer to such strategies as asymptotically worst-case optimal.

1 Introduction

Experimental design is crucial in decision-making (Fisher, 1935; Robbins, 1952). This study investigates scenarios involving multiple *treatment arms*[1], such as online advertisements, slot machine arms, diverse therapeutic strategies, and assorted unemployment assistance programs. The objective of an experiment is to identify the treatment arm that yields the highest expected outcome (the *best treatment arm*), while minimizing the probability of misidentification. This problem has been examined in various research areas under a range of names, including *best arm identification* (BAI, Audibert et al., 2010), *ordinal optimization* (Ho et al., 1992)[2], *optimal budget allocation* (Chen et al., 2000), and *policy choice* (Kasy & Sautmann, 2021).

[1]The term 'treatment arm' is used in clinical trials (Nair, 2019) and economics (Athey & Imbens, 2017). There are various names for this concept, including 'arms' (Lattimore & Szepesvári, 2020), 'policies' (Kasy & Sautmann, 2021), 'treatments' (Hahn et al., 2011), 'designs' (Chen et al., 2000), 'populations' (Glynn & Juneja, 2004), and 'alternatives' (Shin et al., 2018).

[2]Ordinal optimization often considers non-adaptive experiments, and BAI primarily addresses adaptive experiments. However, there are also studies on adaptive experiments in ordinal optimization and non-adaptive experiments in BAI.

We mainly follow the terminologies in BAI. BAI has two formulations: fixed-budget and fixed-confidence BAI; this study focuses on the fixed-budget BAI[3].

1.1 Problem Setting

We consider a decision-maker who conducts an experiment with a fixed number of rounds T, referred to as a sample size or a *budget*, and a fixed set of treatment arms $[K] := 1, 2, \ldots, K$. In each round $t \in [T] := 1, 2, \ldots, T$, the decision-maker allocates a treatment arm $A_t \in [K]$ to an experimental unit and immediately observes an outcome Y_t linked to the allocated treatment arm A_t. The decision-maker's goal is to identify the treatment arm with the highest expected outcome, minimizing the probability of misidentification after observing the outcomes at round T.

Potential outcomes. To describe the data-generating process, we introduce potential outcomes following the Neyman-Rubin model (Neyman, 1923; Rubin, 1974). Let P be a joint distribution of K-potential outcomes $(Y^1, Y^2, \ldots, Y^K)$. For P, let $\mathbb{P}P$ and $\mathbb{E}P$ be the probability and expectation under P, respectively, and let $\mu^a(P) = \mathbb{E}P[Y^a]$ be the expected outcome. Throughout this study, we assume that $(Y^1, Y^2, \ldots, Y^K)$ follows a multivariate Gaussian distribution with a unique best treatment arm $a^{\in}[K]$. Formally, we define a set of joint distributions P as

$$\mathcal{P}(a^*, \underline{\Delta}, \overline{\Delta}) :=$$
$$\left\{ P = \left(\mathcal{N}\left(\mu^a, (\sigma^a)^2\right)\right)_{a\in[K]} \mid (\mu^a)_{a\in[K]} \in \mathbb{R}^K, \quad (\sigma^a)^2_{a\in[K]} \in [\underline{C}, \overline{C}]^K, \quad \forall a \in [K]\backslash\{a^*\}\ \underline{\Delta} \leq \mu^{a^*} - \mu^a \leq \overline{\Delta} \right\},$$

where $\mathcal{N}(\mu, v)$ is a Gaussian distribution with a mean (expected outcome) $\mu \in \mathbb{R}$ and a variance $v > 0$, $\underline{\Delta}, \overline{\Delta}$ are constants independent of T such that they are lower and upper bounds for the gap $\mu^{a^*} - \max_{a\in[K]\backslash\{a^*\}} \mu^a$ and $0 < \underline{\Delta} < \overline{\Delta} < \infty$ holds, and $\underline{C}, \overline{C}$ are *unknown* universal constants such that $0 < \underline{C} < \overline{C} < \infty$. Note that $\underline{C}$ and $\overline{C}$ are just introduced for a technical purpose to assume that $(\sigma^a)^2$ is bounded, and we do not use them in designing algorithms. We refer to $\mathcal{P}(a^*, \underline{\Delta}, \overline{\Delta})$ as a *Gaussian bandit model*.

Experiment. Let $P_0 \in \mathcal{P}(a^*, \underline{\Delta}, \overline{\Delta})$ be an instance of bandit models that generates potential outcomes in an experiment, which is decided in advance of the experiment, fixed throughout the experiment, and unknown for the decision-maker except for the variances. An outcome in round $t \in [T]$ is $Y_t = \sum_{a\in[K]} \mathbb{1}[A_t = a] Y_t^a$, where $Y_t^a \in \mathbb{R}$ is a potential independent outcome (random variable), and $(Y_t^1, Y_t^2, \ldots, Y_t^K)$ be an independent (i.i.d.) copy of $(Y^1, Y^2, \ldots, Y^K)$ at round $t \in [T]$ under P_0. Then, we consider an experiment with the following procedure of a decision-maker at each round $t \in [T]$:

1. A potential outcome $(Y_t^1, Y_t^2, \ldots, Y_t^K)$ is drawn from P_0.
2. The decision-maker allocates a treatment arm $A_t \in \mathcal{A}$ based on past observations $\{(Y_s, A_s)\}_{s=1}^{t-1}$.
3. The decision-maker observes a corresponding outcome $Y_t = \sum_{a\in\mathcal{A}} \mathbb{1}[A_t = a] Y_t^a$

At the end of the experiment, the decision-maker estimates a^*, denoted by $\widehat{a}_T \in [K]$. Here, an outcome in round $t \in [T]$ is $Y_t = \sum_{a\in[K]} \mathbb{1}[A_t = a] Y_t^a$.

Probability of misidentification. Our goal is to minimize the *probability of misidentification*, defined as

$$\mathbb{P}_{P_0}(\widehat{a}_T \neq a^*).$$

It is known that for each fixed $P_0 \in \mathcal{P}(a^*, \underline{\Delta}, \overline{\Delta})$, when a^* is unique, $\mathbb{P}_{P_0}(\widehat{a}_T \neq a^*)$ converges to zero with an exponential speed as $T \to \infty$. Therefore, to evaluate the exponential speed, we employ the following measure, called the *complexity*:

$$-\frac{1}{T} \log \mathbb{P}_{P_0}(\widehat{a}_T \neq a^*).$$

The complexity $-\frac{1}{T} \log \mathbb{P}_{P_0}(\widehat{a}_T \neq a^*)$ is widely referenced in the literature on ordinal optimization and BAI (Glynn & Juneja, 2004; Kaufmann et al., 2016). In hypothesis testing, Bahadur (1960) suggests the use of a similar measure to assess performances of methods in hypothesis testing. Also see Section A.3.

[3]The fixed-confidence BAI formulation resembles sequential testing, where the sample size is a random stopping time.

Strategy. We define a *strategy* of a decision-maker as a pair of $((A_t)_{t\in[K]}, \widehat{a}_T)$, where $(A_t)_{t\in[K]}$ is the allocation rule, and $\widehat{a}_T$ is the recommendation rule. Formally, with the sigma-algebras $\mathcal{F}_t = \sigma(A_1, Y_1, \ldots, A_t, Y_t)$, a strategy is a pair $((A_t)_{t\in[T]}, \widehat{a}_T)$, where

- $(A_t)_{t\in[T]}$ is an allocation rule, which is $\mathcal{F}_{t-1}$-measurable and allocates a treatment arm $A_t \in [K]$ in each round t using observations up to round $t-1$.
- $\widehat{a}_T$ is a recommendation rule, which is an $\mathcal{F}_T$-measurable estimator of the best treatment arm a^* using observations up to round T.

We denote a strategy by π. We also denote A_t and $\widehat{a}_T$ by A_t^π and $\widehat{a}_T^\pi$ when we emphasize that A_t and $\widehat{a}_T$ depend on π.

This definition of strategies allows us to design adaptive experiments where we can decide A_t using past observations. In this study, although we develop lower bounds that work for both adaptive and non-adaptive experiments[4], our proposed strategy is non-adaptive; that is, A_t is decided without using observations obtained in an experiment. As we show later, our lower bounds depend only on variances of potential outcomes. By assuming that the variances are known, we design a non-adaptive strategy that is asymptotically optimal in the sense that its probability of misidentification aligns with the lower bounds. If the variances are unknown, we may consider estimating them during an experiment and using $\mathcal{F}_{t-1}$-measurable variance estimators at each round t. However, it is unknown whether an optimal strategy exists when we estimate variances during an experiment. We leave it as an open issue (Section 7).

Notation. When emphasizing the dependency of the best treatment arm on P, we denote it by $a^*(P) := \arg\max_{a\in[K]} \mu^a(P)$. Let $\Delta^a(P) := \mu^{a^*(P)}(P) - \mu^a(P)$. For $P \in \mathcal{P}$, let P^a be a distribution of a reward of treatment arm $a \in [K]$.

1.2 Existence of Optimal Strategies

The existence of optimal strategies in fixed-budget BAI has been a longstanding issue. For example, in fixed-budget BAI, tight lower bounds for the probability of misidentification are unknown. Although there are several conjectures regarding tight lower bounds, it remains unclear whether an optimal strategy exists such that its probability of misidentification aligns with these conjectured lower bounds.

Kaufmann et al. (2016) derives lower bounds for both fixed-budget and fixed-confidence BAI. Garivier & Kaufmann (2016) presents an optimal strategy for fixed-confidence BAI and conjectures an optimal strategy for fixed-budget BAI. Summarizing these early discussions, Kaufmann (2020) clarifies the problem. Following these studies, the existence of optimal strategies is discussed in economics by Kasy & Sautmann (2021) and Ariu et al. (2021). Ariu et al. (2021) provides an instance where the lower bound is larger than a conjectured lower bound from Kaufmann et al. (2016)'s result, implying that there is no strategy such that its upper bound aligns with the lower bounds conjectured by Kaufmann et al. (2016) under any distribution P_0. Qin (2022) summarizes these arguments as an open problem.

To discuss tight lower bounds, we consider restricting the class of strategies. Specifically, we focus on consistent and asymptotically invariant strategies. Consistent strategies recommend the true best arm with probability one as $T \to \infty$ (see Definition 2.1). Asymptotically invariant strategies allocate treatment arms in the same proportion under any distribution $P_0 \in \mathcal{P}(a^*, \underline{\Delta}, \overline{\Delta})$ as T approaches infinity (see Definition 2.3).

One key question is whether there are strategies whose probability of misidentification is smaller than those using uniform sampling (allocating treatment arms to experimental units with equal sample sizes; that is, $\sum_{t=1}^T \mathbb{1}[A_t = a] = T/K$ for all $a \in [K]$). For consistent strategies, Kaufmann et al. (2016) notes that a strategy using uniform sampling is nearly optimal when outcomes follow a distribution in the two-armed one-parameter exponential family. Furthermore, they show that a strategy allocating treatment arms in proportion to the standard deviations of outcomes is asymptotically optimal in two-armed Gaussian bandits if

[4]Non-adaptive experiments are also referred to as static experiments. The difference between adaptive and non-adaptive experiments is the dependency on the past observations. In non-adaptive experiments, we first fix $\{A_t\}_{t\in[T]}$ at the beginning of an experiment and do not change it. Both in adaptive and non-adaptive experiments, $\widehat{a}_T$ depends on observations $\{(A_t, Y_t)\}_{t=1}^T$.

the standard deviations are known. Such a strategy, known as the *Neyman allocation*, is conjectured or shown to be optimal in some sense under Gaussian bandits (Neyman, 1934; Hahn et al., 2011; Glynn & Juneja, 2004; Kaufmann et al., 2016; Adusumilli, 2022).

For asymptotically invariant strategies, we show that a strategy proposed by Glynn & Juneja (2004) is asymptotically optimal for multi-armed bandits with general distributions, which is also independently shown by Degenne (2023). However, the strategy of Glynn & Juneja (2004) is feasible only when the distribution is completely known. For example, in Gaussian bandits, knowledge of the mean and variance parameters, including which arm is the best treatment arm, is necessary. Therefore, the strategy of Glynn & Juneja (2004) is usually infeasible in practice, especially when mean parameters are unknown.

1.3 Main Results

This study investigates the existence of optimal strategies by providing the following results:

1. Tight worst-case lower bounds for multi-armed Gaussian bandits that are applicable to both adaptive and non-adaptive strategies.
2. An asymptotically optimal non-adaptive strategy whose upper bound aligns with the lower bound as the budget approaches infinity and the difference between $\underline{\Delta}$ and $\overline{\Delta}$ approaches zero.

In Section 2, we derive lower bounds for strategies based on available information. Specifically, we focus on how the available information affects lower bounds and the existence of strategies whose probability of misidentification aligns with these lower bounds. We find that lower bounds depend significantly on the amount of information available regarding the distribution of rewards for treatment arms prior to the experiment.

From the information theory, we can relate the lower bounds to the Kullback–Leibler (KL) divergence $\mathrm{KL}(Q^a, P_0^a)$ between $P_0 \in \mathcal{P}(a^*, \underline{\Delta}, \overline{\Delta})$ and an alternative hypothesis $Q \in \cup_{b \neq a^*} \mathcal{P}(b, \underline{\Delta}, \overline{\Delta})$ (Lai & Robbins, 1985; Kaufmann et al., 2016). From the lower bounds, we can compute an ideal expected number of times treatment arms are allocated to each experimental unit; that is, $\mathbb{E}_{P_0}\left[\frac{1}{T}\sum_{t=1}^{T} \mathbb{1}[A_t = a]\right]$. When the lower bounds are linked to the KL divergence, the corresponding ideal sample allocation rule also depends on the KL divergence (Glynn & Juneja, 2004).

If we know the distributions of treatment arms' outcomes completely, we can compute the KL divergence, which allows us to design a strategy whose probability of misidentification matches the lower bounds of Kaufmann et al. (2016) as $T \to \infty$ (Glynn & Juneja, 2004; Chen et al., 2000; Gärtner, 1977; Ellis, 1984).

However, it is common to encounter scenarios where there is either partial or no distributional knowledge available. Since optimal strategies are characterized by distributional information, the lack of complete information hinders us from designing asymptotically optimal strategies. Therefore, we reflect the limitation by considering the worst cases regarding the mean parameters and the choice of the best treatment arm. Specifically, we consider the worst-case lower bound defined as

$$\min_{a^* \in [K]} \inf_{P_0 \in \mathcal{P}(a^*, \underline{\Delta}, \overline{\Delta})} \limsup_{T \to \infty} -\frac{1}{T} \log \mathbb{P}_{P_0}(\widehat{a}_T^{\pi} \neq a^*).$$

While the lower bounds with complete information are characterized by the KL divergence (Lai & Robbins, 1985; Kaufmann et al., 2016), the worst-case lower bounds are characterized by the variances of potential outcomes. Hence, knowledge of at least the variances is sufficient to design worst-case optimal strategies.

Let $\mathcal{W}$ be a set of functions $w : [K] \to (0, 1)$ such that $\sum_{a \in [K]} w(a) = 1$, defined as $\mathcal{W} = \left\{ w : [K] \to (0, 1) \mid \sum_{a \in [K]} w(a) = 1 \right\}$. Then, Theorem 2.6 of Section 2 provides the following lower bound[5]:

$$\min_{a^* \in [K]} \inf_{P_0 \in \mathcal{P}(a^*, \underline{\Delta}, \overline{\Delta})} \limsup_{T \to \infty} -\frac{1}{T} \log \mathbb{P}_{P_0}(\widehat{a}_T^{\pi} \neq a^*) \leq \mathrm{LowerBound}(\overline{\Delta}) := \max_{w \in \mathcal{W}} \min_{a^* \in [K],\ a \in [K] \setminus \{a^*\}} \frac{\overline{\Delta}^2}{2\Omega^{a^*, a}(w)},$$

[5]Note that lower bounds (resp. upper bounds) for $\mathbb{P}_{P_0}(\widehat{a}_T^{\pi} \neq a^*)$ corresponds to upper bounds (resp. lower bounds) for $-\frac{1}{T} \log \mathbb{P}_{P_0}(\widehat{a}_T^{\pi} \neq a^*)$.

where $\Omega^{a^*,a}(w) = \frac{\left(\sigma^{a^*}\right)^2}{w(a^*)} + \frac{\left(\sigma^{a}\right)^2}{w(a)}$.

In Section 3, based on the lower bounds, we design a strategy and show that its probability of misidentification aligns with the lower bounds. In the experimental design, we assume that the variances of outcomes are *known*. Then, we propose the Generalized-Neyman-Allocation (GNA)-empirical-best-arm (EBA) strategy, which can be interpreted as a generalization of the Neyman allocation proposed by Neyman (1934) . Under the strategy, we allocate each treatment arm a to $\lceil w^*(a)T \rceil$ units, where

$$w^{\mathrm{GNA}} := \left(w^{\mathrm{GNA}}(1), w^{\mathrm{GNA}}(2), \ldots, w^{\mathrm{GNA}}(K)\right) = \arg\max_{w \in \mathcal{W}} \min_{a^* \in [K]} \min_{a \in [K] \setminus \{a^*\}} \frac{1}{2\Omega^{a^*,a}(w)},$$

In Theorem 4.2 of Section 4, for the GNA-EBA strategy, we show that the upper bound of the probability of misidentification is

$$\min_{a^* \in [K]} \inf_{P_0 \in \mathcal{P}(a^*, \underline{\Delta}, \overline{\Delta})} \liminf_{T \to \infty} -\frac{1}{T} \log \mathbb{P}_{P_0}\left(\widehat{a}_T^{\mathrm{EBA}} \neq a^*\right) \geq \mathrm{UpperBound}(\underline{\Delta}) := \max_{w \in \mathcal{W}} \min_{a^* \in [K],\ a \in [K] \setminus \{a^*\}} \frac{\underline{\Delta}^2}{2\Omega^{a^*,a}(w)}.$$

This upper bound implies that the probability of misidentification matches the lower bound as $T \to \infty$ and $\underline{\Delta} - \overline{\Delta} \to 0$; that is, $\mathrm{LowerBound}(\overline{\Delta}) - \mathrm{UpperBound}(\underline{\Delta}) \to 0$ as $\underline{\Delta} - \overline{\Delta} \to 0$. This case implies a situation in which the difference between the expected outcome of the best treatment arm and the next best treatment arm is equal to the difference between the expected outcome of the best treatment arm and the worst treatment arm. In such a situation, it is possible to construct an optimal algorithm even if the expected outcomes are unknown. This concept is illustrated in Figure 1.

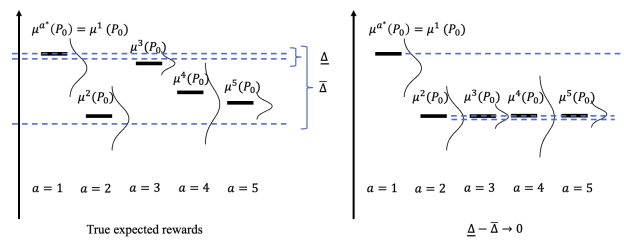

Figure 1: An idea in the derivation of the lower bounds. To lower bound the probability of misidentification (upper bound $-\frac{1}{T}\log \mathbb{P}_{P_0}(\widehat{a}_T^{\pi} \neq a^*)$) it is sufficient to consider a case in the right figure.

As a side-product of our study, we propose a novel setting called hypothesis BAI (HBAI) in Section 5. The results shown above consider a lower bound that is valid for any choices of the mean parameters and the best treatment arm, reflecting a situation where they are unknown. However, we can consider a situation where there is a conjecture of the best treatment arm prior to an experiment. For example, as well as hypothesis testings, we can set null and alternative hypotheses such that a conjectured best treatment arm $\widetilde{a} \in [K]$ is truly best (alternative) or not (null). Under this setting, if we set $\widetilde{a}$ as a proxy of a^* and conduct an experiment, then the probability of minimization is minimized when $\widetilde{a} = a^*$ (the alternative hypothesis is true). As an analogy of hypothesis testing, we call this setting HBAI.

Results of simulation studies are shown in Section 6.

Related work and open problems. Here, we briefly introduce related work. A more comprehensive related work introduction is shown in Appendix A.

As discussed above, the existence of asymptotically optimal strategies for minimization of the probability of misidentification is a longstanding open problem in the community (Kaufmann et al., 2016; Kaufmann, 2020; Ariu et al., 2021; Qin, 2022; Degenne, 2023). Komiyama et al. (2022) provides the minimax optimal strategy and Atsidakou et al. (2023) develops the Bayes optimal strategy. For the expected simple regret minimization, Bubeck et al. (2009; 2011) discuss the worst-case optimality of the uniform allocation, and Komiyama et al. (2022) presents the Bayes optimal strategy. In the Bayesian analysis Russo (2020) shows the posterior convergence optimality of their proposed Bayesian strategies, which do not directly imply the optimality in our problem.

Extending our result, Kato et al. (2023b) discusses that we can characterize the lower bounds by using the variance when the gaps between the best and suboptimal arms approach zero (small-gap regime). This

is because we can approximate the KL divergence of a wide range of distributions by the semiparametric influence function (a semiparametric analog of the Fisher information in parametric models) under the small-gap regime. Based on this finding, Kato et al. (2023b) points out that under the small-gap regime,

- when $K = 2$, and outcomes follow a distribution in the two-armed one-parameter exponential family, the uniform sampling is asymptotically optimal for consistent strategies.
- when $K \geq 3$, and outcomes follow a distribution in the two-armed one-parameter exponential family, the uniform sampling is asymptotically optimal for consistent and asymptotically invariant strategies.

This finding corresponds to a refinement of a comment by Kaufmann et al. (2016). Furthermore, these observations are further explored by Wang et al. (2023) to support the optimality of the uniform sampling.

There remain several open problems, e.g., the existence of optimal strategies that do not assume known variances.

2 Lower Bounds

We first develop lower bounds based on an information-theoretic approach. Then, we find that different available information yields different lower bounds. Finally, we develop the worst-case lower bound for multi-armed Gaussian bandits.

2.1 Existence of Asymptotically Optimal Strategies

The existence of asymptotically optimal strategies is a longstanding open problem (Kaufmann, 2020; Ariu et al., 2021; Degenne, 2023). Kaufmann et al. (2016) provide a lower bound for two-armed bandits and propose an asymptotically optimal strategy for two-armed Gaussian bandits using known variances. However, when the number of treatment arms is three or more ($K \geq 3$), even the lower bounds remain unknown.

There are multiple reasons why this problem is challenging, and it is difficult to offer a single clear explanation. For instance, factors affecting the upper bounds of strategies include: (i) the estimation error of distributional information, (ii) the class of strategies, and (iii) the dependency of optimal sample allocation ratios on the best treatment arm.

To address this problem, we develop novel lower bounds by extending those shown by Kaufmann et al. (2016), focusing on the lack of distributional information.

2.2 Transportation Lemma

To derive lower bounds, we first restrict a class of our strategies. Specifically, we consider consistent strategies, which are also considered in Kaufmann et al. (2016).

Definition 2.1 (Consistent strategy)**.** We say that a strategy π is consistent if $\mathbb{P}_P(\widehat{a}_T^\pi = a^*) \to 1$ as $T \to \infty$ for each $P \in \mathcal{P}(a^*, \underline{\Delta}, \overline{\Delta})$ such that a^* is unique.

For an instance $P \in \mathcal{P}(a^*, \underline{\Delta}, \overline{\Delta})$ and a strategy $\pi \in \Pi$, let us define an average sample allocation ratio $\kappa_{T,P}^\pi : [K] \to (0, 1)$ as $\kappa_{T,P}^\pi(a) = \mathbb{E}_P\left[\frac{1}{T}\sum_{t=1}^T \mathbb{1}[A_t^\pi = a]\right]$, which satisfies $\sum_{a\in[K]} \kappa_{T,P}^\pi(a) = 1$. This quantity represents the average sample allocation to each treatment arm a over a distribution $P \in \mathcal{P}(a^*, \underline{\Delta}, \overline{\Delta})$ under a strategy π.

Then, Kaufmann et al. (2016) presents the following lower bound for $\mathbb{P}_{P_0}(\widehat{a}_T^\pi \neq a^*)$.

Lemma 2.2 (Lower bound given known distributions)**.** *For each $P_0 \in \mathcal{P}(a^*, \underline{\Delta}, \overline{\Delta})$, any consistent (Definition 2.1) strategy π satisfies*

$$\limsup_{T\to\infty} -\frac{1}{T}\log \mathbb{P}_{P_0}(\widehat{a}_T^\pi \neq a^*) \leq \limsup_{T\to\infty} \inf_{Q\in\cup_{b\neq a^*}\mathcal{P}(b,\underline{\Delta},\overline{\Delta})} \sum_{a\in[K]} \kappa_{T,Q}^\pi(a)\mathrm{KL}(Q^a, P_0^a).$$

Here, Q is an alternative hypothesis that is used for deriving lower bounds and not an actual distribution. Note that upper (resp. lower) bounds for $-\frac{1}{T}\log \mathbb{P}_{P_0}(\widehat{a}^\pi_T \neq a^*)$ corresponds to lower (resp. upper) bounds for $\mathbb{P}_{P_0}(\widehat{a}^\pi_T \neq a^*)$.

For two-armed Gaussian bandits, the lower bound can be simplified (See Theorem 12 in Kaufmann et al. (2016)). In this case, it is known that by allocating treatment arm 1 and 2 with sample sizes $\frac{\sigma^1}{\sigma^1+\sigma^2}T$ and $\frac{\sigma^2}{\sigma^1+\sigma^2}T$, we can design asymptotically optimal strategy. Strategies using this allocation rule are called the *Neyman allocation* (Neyman, 1934).

However, to the best of our knowledge, for general distributions with $K \geq 3$, the existence of lower bounds is still an open problem (Kaufmann, 2020; Ariu et al., 2021; Degenne, 2023).

2.3 Lower Bound given Known Distributions

One of the difficulties comes from the open problem that the term $\kappa^\pi_{T,Q}(a)$ does not correspond to sample allocation under P_0 (Kaufmann, 2020). To derive lower bounds, we connect $\kappa^\pi_{T,Q}(a)$ to $\kappa^\pi_{T,P_0}(a)$ by restricting strategies.

In this study, we consider restricting strategies to ones such that the limit of $\kappa^\pi_{T,P}(a)$ ($\lim_{T\to\infty}\kappa^\pi_{T,P}(a)$) is the same across $P \in \mathcal{P}(a^*, \underline{\Delta}, \overline{\Delta})$.

Definition 2.3 (Asymptotically invariant strategy)**.** A strategy π is called asymptotically invariant if there exists $w^\pi \in \mathcal{W}$ such that for any $P \in \mathcal{P}(a^*, \underline{\Delta}, \overline{\Delta})$, and all $a \in [K]$,

$$\kappa^\pi_{T,P}(a) = w^\pi(a) + o(1) \tag{1}$$

holds as $T \to \infty$.

Note that $\kappa^\pi_{T,P}$ is a deterministic value without randomness because it is an expected value of $\frac{1}{T}\sum_{t=1}^T \mathbb{1}[A^\pi_t = a]$. For simplicity, we omit π from $\kappa^\pi_{T,P}$ and $w^\pi(a)$ if the dependency is obvious from the context[6].

A typical example of this class of strategies is one using uniform sampling, such as the Uniform-EBA strategy (Bubeck et al., 2011). Another example is a strategy using the allocation rule only based on variances, such as the Neyman allocation (Neyman, 1934).

Given an asymptotically invariant strategy π, there exists $w^\pi \in \mathcal{W}$ such that for all $P \in \mathcal{P}(a^*, \underline{\Delta}, \overline{\Delta})$, and $a \in [K]$, $\left| w^\pi(a) - \frac{1}{T}\sum_{t=1}^T \mathbb{E}_P\left[\mathbb{1}[A_t = a]\right]\right| \to 0$ holds.

For any consistent and asymptotically invariant strategy π, the following lower bounds hold.

Lemma 2.4 (Lower bound given known distributions)**.** *For each $P_0 \in \mathcal{P}(a^*, \underline{\Delta}, \overline{\Delta})$, any consistent (Definition 2.1) and asymptotically invariant (Definition 2.3) strategy π satisfies*

$$\limsup_{T\to\infty} -\frac{1}{T}\log \mathbb{P}_{P_0}(\widehat{a}^\pi_T \neq a^*) \leq \sup_{w\in\mathcal{W}} \min_{a\in[K]\setminus\{a^*\}} \frac{\left(\Delta^a(P_0)\right)^2}{2\Omega^{a^*,a}(w)}.$$

The proof is shown in Appendix B.

We refer to a limit of the average sample allocation deduced from lower bounds as the *target allocation ratio* and denote it by w^*. We can derive various w^* in different lower bounds. For example, in Lemma 2.4, because $\max_{w\in\mathcal{W}}\min_{a\in[K]\setminus\{a^*\}}\inf_{\substack{(\mu^b)\in\mathbb{R}^K\\ \mu^a>\mu^{a^*}}} w(a)\frac{(\mu^a-\mu^a(P_0))^2}{2(\sigma^a)^2}$ exists, we define the target allocation ratio as

$$w^* = \arg\max_{w\in\mathcal{W}} \min_{a\in[K]\setminus\{a^*\}} \frac{\left(\Delta^a(P_0)\right)^2}{2\Omega^{a^*,a}(w)}. \tag{2}$$

The target allocation ratio w^* works as a conjecture about optimal sample allocation under which the probability of misidentification matches the lower bounds. Here, note that the average sample allocation ratio is linked to an actual strategy, and we can compute w^* independently of each instance P_0.

[6]Degenne (2023) proposes a similar restriction independently of us.

For the asymptotically invariant strategy, we can show that the strategy proposed by Glynn & Juneja (2004) is feasible if we can compute $\mathrm{KL}(Q^a, P_0^a)$, and under the strategy, the probability of misidentification aligns with the lower bound with asymptotically invariant strategies. This result is also shown by Degenne (2023) for more general distributions.

2.4 Uniform Lower Bound

As discussed by Glynn & Juneja (2004) and us, when we know distributional information completely, we can obtain an asymptotically optimal strategy whose probability of misidentification matches the lower bounds in Lemma 2.4. However, when we do not have complete information, related work such as (Ariu et al., 2021) find that there exists $P_0 \in \mathcal{P}(a^*, \underline{\Delta}, \overline{\Delta})$ whose lower bound is larger than that of Kaufmann et al. (2016).

For example, in Lemma 2.4, the target allocation ratio is given as equation 2. However, the target allocation ratio depends on unknown mean parameters $\mu^a(P_0)$ and the true best treatment arm a^*. Therefore, strategies using the target allocation ratio are infeasible if we do not know those values.

We elucidate this problem by examining how our available information affects the lower bounds. Specifically, we consider lower bounds that are uniformly valid regardless of the missing information. First, we consider characterizing the lower bound in Lemma 2.4 by $\overline{\Delta}$, an upper bound of $\Delta^a(P_0)$ as the following lemma.

Theorem 2.5 (Uniform lower bound)**.** *For each $P_0 \in \mathcal{P}(a^*, \underline{\Delta}, \overline{\Delta})$, any consistent (Definition 2.1) and asymptotically invariant (Definition 2.3) strategy π satisfies*

$$\limsup_{T\to\infty} -\frac{1}{T}\log \mathbb{P}_{P_0}(\widehat{a}_T^\pi \neq a^*) \leq \frac{\overline{\Delta}^2}{2\left(\sigma^{a^*} + \sqrt{\sum_{a\in[K]\setminus\{a^*\}} \left(\sigma^a\right)^2}\right)^2}.$$

The proof is shown in Appendix C. Here, the target allocation ratio is given as

$$w^*(a^*) = \frac{\sigma^{a^*}}{\sigma^{a^*} + \sqrt{\sum_{b\in[K]\setminus\{a^*\}} \left(\sigma^b\right)^2}}, \tag{3}$$

$$w^*(a) = \frac{\left(\sigma^b\right)^2 / \sqrt{\sum_{b\in[K]\setminus\{a^*\}} \left(\sigma^b\right)^2}}{\sigma^{a^*} + \sqrt{\sum_{b\in[K]\setminus\{a^*\}} \left(\sigma^b\right)^2}} = (1 - w^*(a^*))\frac{\left(\sigma^a\right)^2}{\sum_{b\in[K]\setminus\{a^*\}} \left(\sigma^b\right)^2}, \quad \forall a \in [K]\setminus\{a^*\}.$$

Note that the lower bounds are characterized by the variances and the true best treatment arm a^*. If we know them, we can design optimal strategies whose upper bound aligns with these lower bounds.

When designing strategies, variances and the best treatment arm are required to construct the target allocation ratio. However, assuming that the best treatment arm a^* is known is unrealistic. We also cannot estimate it during an experiment because such a strategy violates the assumption of asymptotically invariant strategies. For example, Shin et al. (2018) estimates a^* during an experiment under the framework of Glynn & Juneja (2004), but such a strategy does not satisfy the asymptotically invariant strategies because a^* and w^* can differ across the choice of P_0. When considering asymptotically invariant strategies, we need to know a^* before starting an experiment.

To circumvent this issue, one approach is to fix $\widetilde{a}$, independent of P_0, before an experiment. We then construct a target allocation ratio as $w^\dagger(\tilde{a}) = \frac{\sigma^{\tilde{a}}}{\sigma^{\tilde{a}} + \sqrt{\sum_{b\in[K]\setminus\tilde{a}} (\sigma^b)^2}}$ and $w^\dagger(a) = (1 - w^\dagger(\tilde{a}))\frac{(\sigma^a)^2}{\sum_{b\in[K]\setminus\tilde{a}} (\sigma^b)^2}$ for all $a \in [K]\setminus\tilde{a}$. Treatment arms are then allocated following this target allocation ratio. Under a strategy using such an allocation rule, if $\widetilde{a}$ equals a^*, the target allocation ratio $w^\dagger$ aligns with w^* in equation 3. We refer to this setting as *hypothesis BAI* (HBAI), as the formulation resembles hypothesis testing. For details, see Section 5.

However, when $\widetilde{a}$ is *not* equal to a^*, the target allocation ratio $w^\dagger$ does not equal w^*, under which the strategy becomes suboptimal. In the following section, we consider a best-arm agnostic lower bound by contemplating the worst-case scenario for the choice of a^*.

Remark 1. When $K = 2$, the target allocation ratio has a closed-form such that $w^{\mathrm{GNA}}(a) = \frac{\sigma^a}{\sigma^1+\sigma^2}$ for $a \in [K] = \{1, 2\}$. Note that the target allocation ratio becomes independent of a^* in this case. Allocation rules following this target allocation ratio are referred to as the Neyman allocation (Neyman, 1934).

2.5 Best-Arm-Worst-Case Lower Bound

As discussed above, a^* is unknown in practice, and we consider the worst-case analysis regarding the choice of the best treatment arm. We represent the worst-case for all possible best treatment arms by using the following metric:

$$\min_{a^* \in [K]} \inf_{P_0 \in \mathcal{P}(a^*, \underline{\Delta}, \overline{\Delta})} \limsup_{T \to \infty} -\frac{1}{T} \log \mathbb{P}_{P_0}(\widehat{a}^{\pi}_T \neq a^*).$$

Thus, by taking the worst case over P_0, we obtain the following lower bound.

Theorem 2.6 (Best-arm-worst-case lower bound)**.** *Any consistent (Definition 2.1) and asymptotically invariant (Definition 2.3) strategy $\pi \in \Pi$ satisfies*

$$\min_{a^* \in [K]} \inf_{P_0 \in \mathcal{P}(a^*, \underline{\Delta}, \overline{\Delta})} \limsup_{T \to \infty} -\frac{1}{T} \log \mathbb{P}_{P_0}(\widehat{a}^{\pi}_T \neq a^*) \leq \mathrm{LowerBound}(\overline{\Delta}).$$

The target allocation ratio deduced from this lower bound is

$$w^* = \arg\max_{w \in \mathcal{W}} \min_{a^* \in [K],\ a \in [K] \setminus \{a^*\}} \frac{1}{2\Omega^{a^*, a}(w)}. \tag{4}$$

Here, note that $\max_{w \in \mathcal{W}} \max_{a^* \in [K]} \min_{a \in [K] \setminus \{a^*\}} \frac{\underline{\Delta}^2}{2\Omega^{a^*, a}(w)}$ does not have a closed-form solution and requires numerical computations.

Note that when $K = 2$, the target allocation ratio has a closed-form solution such that $w^*(a) = \frac{\sigma^a}{\sigma^1+\sigma^2}$ for $a \in [2]$. Additionally, the lower bound is given as

$$\min_{a^* \in [K]} \inf_{P_0 \in \mathcal{P}(a^*, \underline{\Delta}, \overline{\Delta})} \lim_{\overline{\Delta}(P) \to 0} \limsup_{T \to \infty} -\frac{1}{T} \log \mathbb{P}_{P_0}(\widehat{a}^{\pi}_T \neq a^*) \leq \frac{\overline{\Delta}^2}{2\left(\sigma^1 + \sigma^2\right)^2}.$$

This target allocation ratio and lower bound are equal to those in the lower bound for two-armed Gaussian bandits shown by Kaufmann et al. (2016).

The target allocation ratio is independent of a^*. Therefore, we can avoid the issue of dependency on a^*, which cannot be estimated in an experiment.

Remark 2 (Inequalities of lower bounds)**.** Note that for the derived lower bounds, $\mathrm{LowerBound}(\overline{\Delta}) \leq \frac{\overline{\Delta}^2}{2\left(\sigma^{a^*} + \sqrt{\sum_{a \in [K] \setminus \{a^*\}} (\sigma^a)^2}\right)^2}$ holds. The larger lower bounds imply tighter lower bounds; that is, $\mathbb{P}_{P_0}(\widehat{a}^{\pi}_T \neq a^*)$ is smaller as the lower bounds become larger.

Remark 3 (Two-armed Gaussian bandits)**.** When $K = 2$, the above serial arguments about the lower bound can be simplified. The reason why we cannot use Theorem 2.5 is because the target allocation ratio depends on a^*. However, when $K = 2$, the target allocation ratio is independent of a^* and given as the ratio of the standard deviations. This is because comparison between the best and suboptimal treatment arms plays an important role, which requires the best treatment arm a^*. However, when $K = 2$, a pair of comparisons is unique; that is, we always allocate treatment arms comparing arms 1 and 2, regardless of which arm is best. Therefore, we can simplify the lower bounds when $K = 2$. Specifically, optimal strategies just allocate treatment with the ratio of the standard deviation, which is also referred to as the Neyman allocation (Neyman, 1934). Also, see Theorem 12 in Kaufmann et al. (2016) for details.

3 The GNA-EBA Strategy

We develop the Generalized-Neyman-Allocation (GNA)-Empirical-Best-Arm (EBA) strategy, which is a generalization of the Neyman allocation (Neyman, 1934). The pseudo-code is shown in Algorithm 1.

Algorithm 1 GNA-EBA strategy

Parameter: Fixed budget T.
Allocation rule: generalized Neyman allocation.
Allocate $A_t = 1$ if $t \leq \lceil w^{\mathrm{GNA}}(1)T \rceil$ and $A_t = a$ if $\left\lceil \sum_{b=1}^{a-1} w^{\mathrm{GNA}}(b)T \right\rceil < t \leq \lceil \sum_{b=1}^{a} w^{\mathrm{GNA}}(b)T \rceil$ for $a \in [K] \backslash \{1\}$.
Recommendation rule: empirical best treatment arm.
Recommend $\widehat{a}_T^{\mathrm{EBA}}$ following equation 5.

Allocation rule: Generalized Neyman Allocation (GNA). First, we define a target allocation ratio, which is used to determine our allocation rule, as follows:

$$w^{\mathrm{GNA}} = \arg\max_{w \in \mathcal{W}} \min_{a^* \in [K], a \in [K] \backslash \{a^*\}} \frac{1}{2\Omega^{a^*,a}(w)},$$

which is identical to that in equation 4. Then, we allocate treatment arms to experimental units as follows:

$$A_t = \begin{cases} 1 & \text{if} \quad t \leq \lceil w^{\mathrm{GNA}}(1)T \rceil \\ 2 & \text{if} \quad \lceil w^{\mathrm{GNA}}(1)T \rceil < t \leq \left\lceil \sum_{b=1}^{2} w^{\mathrm{GNA}}(b)T \right\rceil \\ \vdots & \\ K & \text{if} \quad \left\lceil \sum_{b=1}^{K-1} w^{\mathrm{GNA}}(b)T \right\rceil < t \leq T \end{cases}.$$

Recommendation rule: Empirical Best Arm (EBA). After the final round T, we recommend $\widehat{a}_T \in [K]$, an estimate of the best treatment arm, defined as

$$\widehat{a}_T^{\mathrm{EBA}} = \arg\max_{a \in [K]} \widehat{\mu}_T^a, \qquad \widehat{\mu}_T^a = \frac{1}{\lceil w^{\mathrm{GNA}}(a)T \rceil} \sum_{t=1}^{T} \mathbb{1}[A_t = a] Y_t. \tag{5}$$

Our strategy generalizes the Neyman allocation because for w^{GNA} in the GNA allocation rule, $w^{\mathrm{GNA}} = \left(\frac{\sigma^1}{\sigma^1 + \sigma^2}, \frac{\sigma^2}{\sigma^1 + \sigma^2} \right)$ when $K = 2$, which is a target allocation ratio of the Neyman allocation. The EBA recommendation rule is one of the typical recommendation rules and used in other strategies in fixed-budget BAI, such as the Uniform-EBA strategy (Bubeck et al., 2009; 2011).

4 Probability of Misidentification of the GNA-EBA strategy and its Asymptotic Optimality

This section shows an upper bound of the GNA-EBA strategy and its asymptotic optimality.

4.1 Upper Bound for the Probability of Misidentification

First, we show the upper bound for the probability of misidentification. The proof is shown in Appendix D.

Theorem 4.1 (Upper Bound of the GNA-EBA strategy)**.** *For each $P_0 \in \mathcal{P}(a^*, \underline{\Delta}, \overline{\Delta})$, the GNA-EBA strategy satisfies*

$$\liminf_{T \to \infty} -\frac{1}{T} \log \mathbb{P}_{P_0} \left(\widehat{a}_T^{\mathrm{EBA}} \neq a^* \right) \geq \min_{a \neq a^*} \frac{\underline{\Delta}^2}{2\Omega^{a^*,a}(w^{\mathrm{GNA}})}.$$

Then, the worst-case upper bound is given as the following theorem.

Theorem 4.2 (Worst-case upper bound of the GNA-EBA strategy)**.** *The GNA-EBA strategy satisfies*

$$\min_{a^* \in [K]} \inf_{P_0 \in \mathcal{P}(a^*, \underline{\Delta}, \overline{\Delta})} \liminf_{T \to \infty} -\frac{1}{T} \log \mathbb{P}_{P_0} \left(\widehat{a}_T^{\mathrm{EBA}} \neq a^* \right) \geq \mathrm{UpperBound}(\underline{\Delta}).$$

4.2 Asymptotic Optimality

By comparing the lower bound in Theorem 2.6 and the upper bound in Theorem 4.2, we show that they match when $\underline{\Delta} - \overline{\Delta} \to 0$ as the following theorem. Thus, this theorem implies the asymptotic optimality of our proposed GNA-EBA strategy under a situation where $\underline{\Delta} - \overline{\Delta} \to 0$.

Theorem 4.3. *As* $\underline{\Delta} - \overline{\Delta} \to 0$*, we have*

$$\text{LowerBound}(\overline{\Delta}) - \text{UpperBound}(\underline{\Delta}) \to 0.$$

Related work, such as that by Ariu et al. (2021), discusses the non-existence of optimal strategies in fixed-budget Best Arm Identification (BAI) given the lower bound shown by Kaufmann et al. (2016). Our findings do not present a contradiction to their results because we restrict the strategies to asymptotically invariant ones. This finding is also reported by Degenne (2023). In our study, we consider an asymptotically optimal strategy that does not require knowledge of the mean parameters and prior information about the choice of the best treatment arm. When $\underline{\Delta} - \overline{\Delta} \to 0$ and the variances are known, we design an asymptotically worst-case optimal strategy, which does *not* use uniform allocation.

5 Hypothesis BAI

Based on the arguments in Section 2.4, we design the Hypothesis GNA-EBA (H-GNA-EBA) strategy that utilizes a conjecture $\widetilde{a} \in [K]$ of a^*, instead of considering the worst-case for a^*. We refer to $\widetilde{a}$ as a hypothetical best treatment arm.

In the H-GNA-EBA strategy, we define a target allocation ratio as $w^{\text{H-GNA}}(\widetilde{a}) = \frac{\sigma^{\tilde{a}}}{\sigma^{\tilde{a}} + \sqrt{\sum_{b \in [K] \setminus \{\tilde{a}\}} (\sigma^b)^2}}$ and $w^{\text{H-GNA}}(a) = \left(1 - w^{\text{H-GNA}}(\widetilde{a})\right) \frac{(\sigma^a)^2}{\sum_{b \in [K] \setminus \{\tilde{a}\}} (\sigma^b)^2}$ for all $a \in [K] \setminus \{\widetilde{a}\}$. If $\widetilde{a}$ is equal to a^*, the upper bound of the H-GNA-EBA strategy for the probability of misidentification aligns with the lower bound in Theorem 2.5.

This strategy is more suitable under a setting different from BAI, where there are null and alternative hypotheses such that $H_0 : a^* \neq \widetilde{a} \in [K]$ and $H_1 : a^* = \widetilde{a}$; that is, the null hypothesis corresponds to a situation where the hypothetical best treatment arm is *not* the best. In contrast, the alternative hypothesis posits that the hypothetical best treatment arm is the best. Then, we consider minimizing the probability of misidentification when the alternative hypothesis is true. This probability corresponds to power in hypothesis testing. Our aim is to minimize the misidentification probability when the null hypothesis is false, corresponding to the *power of the test*. We refer to this setting as Hypothesis BAI (HBAI). We present two examples for this setting.

Example 1 (Online advertisement)**.** Let $\widetilde{a} \in [K]$ be a treatment arm corresponding to a new advertisement. Our null hypothesis $a^* \neq \widetilde{a}$ implies that the existing advertisements $a \in [K] \setminus \{\widetilde{a}\}$ are superior to the new advertisement. Our goal is to reject the null hypothesis with a maximal probability when the null hypothesis is not correct; that is, the new hypothesis is better than the others.

Example 2 (Clinical trial)**.** Let $\widetilde{a} \in [K]$ be a new drug. Our null hypothesis $a^* \neq \widetilde{a}$ implies that the existing drug $a \in [K] \setminus \{\widetilde{a}\}$ is superior to the new drug (equivalently, the new drug is not good as the existing drugs). Our goal is to reject the null hypothesis with a maximal probability when the new drug is better than the others.

The asymptotic efficiency of hypothesis testing is referred to as the *Bahadur efficiency* (Bahadur, 1960; 1967; 1971) of the test[7].

6 Simulation Studies

We investigate the performances of our strategies in the settings of the GNA-EBA and the H-GNA-EBA, and the existing Uniform-EBA strategy (Uniform, Bubeck et al., 2011) using simulation studies, which allocates

[7]Note that the Bahadur efficiency of a test evaluates P-values (random variable), not the probability of misidentification (non-random variable). However, these are closely related. See Bahadur (1967).

treatment arms with the same allocation ratio $(1/K)$. Let $K \in \{2, 5, 10\}$. The best treatment arm is arm 1 and $\mu^1(P_0) = 1$. The expected outcomes of suboptimal treatment arms are drawn from a uniform distribution with support $[0.75, 0.90]$ for $a \in [K] \backslash \{1, 2\}$, while $\mu^2(P_0) = 0.75$. The variances are drawn from a uniform distribution with support $[0.5, 5]$. We continue the strategies until $T = 10,000$. We conduct 100 independent trials for each setting. For each $T \in 100, 500, 1000, \cdots, 9500, 10000$, we plot the empirical probability of misidentification in Figure 2. From the results, as the theory predicts, strategies that use more information can achieve a lower probability of misidentification.

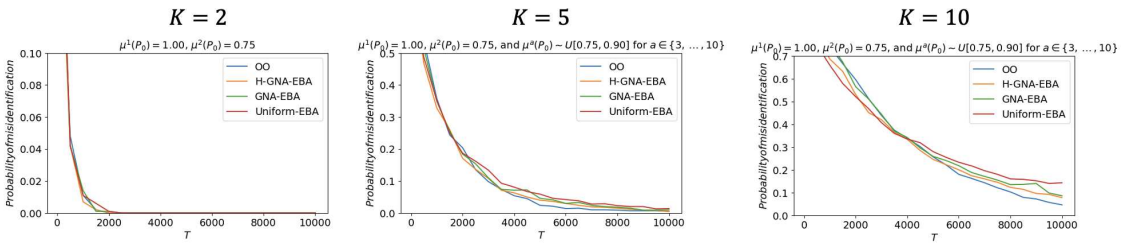

Figure 2: Experimental results. The y-axis and x-axis denote the probability of misidentification and T, respectively.

7 Conclusion

We investigated scenarios in experimental design, known as BAI or ordinal optimization. We found that the optimality of strategies significantly depends on the information available prior to an experiment. Based on our findings, we developed the novel worst-case lower bounds for the probability of misidentification and then proposed a strategy whose worst-case probability of misidentification matches these worst-case lower bounds.

Lastly, we propose three important extensions of our results. The first extension is to remove the restriction of the asymptotically invariant strategies. One promising approach for this question is to apply the minimax large deviation analysis, as performed by Otsu (2008) and Komiyama et al. (2022).

Second, we suggest the potential extension of our results to BAI with general distributions. In this study, we developed lower bounds for multi-armed Gaussian bandits. By approximating the KL divergence of distributions in a broader class under the small-gap regime, we can apply our results to various distributions, including Bernoulli distributions and general nonparametric models. In fact, Kato et al. (2023b) extends our results, developing lower bounds for more general distributions using semiparametric theory (Bickel et al., 1998; van der Vaart, 1998). These extended results, which refine those of Kato et al. (2023b), will be published later. It is important to note that in the small-gap regime, asymptotically optimal strategies for one-parameter bandit models, such as Bernoulli distributions, utilize uniform sampling because variances of outcomes become equal as gaps approach zero[8]. The results from Kato et al. (2023b) align with the conjecture in Kaufmann et al. (2016). From the results of Kato et al. (2023b), we can interpret the GNA-EBA strategy as an extension of the Uniform-EBA strategy proposed by Bubeck et al. (2009; 2011), which recommends uniform sampling in bandit models with bounded supports.

The third question concerns the existence of optimal strategies that estimate variances during an experiment, instead of assuming known variances. If variances are estimated during an experiment, the estimation error affects the probability of misidentification. We conjecture that an optimal strategy exists that estimates variances during an experiment because similar results can be obtained under the central limit theorem (van der Laan, 2008; Hahn et al., 2011; Hadad et al., 2021; Kato et al., 2020; 2023a). However, its existence remains unproven due to the difficulty in evaluating tail probabilities. Related work, such as that by Kato et al. (2023a) and Jourdan et al. (2023), may help to resolve this issue.

[8]Under the small-gap regime, as $\Delta^a(P_0) \to 0$, variances in Bernoulli bandits become equal because the variances are $\mu^a(1 - \mu^a)$. For details, see Kato et al. (2023b).

References

Karun Adusumilli. Neyman allocation is minimax optimal for best arm identification with two arms, 2022. arXiv:2204.05527.

Karun Adusumilli. Risk and optimal policies in bandit experiments, 2023. arXiv:2112.06363.

Kaito Ariu, Masahiro Kato, Junpei Komiyama, Kenichiro McAlinn, and Chao Qin. Policy choice and best arm identification: Asymptotic analysis of exploration sampling, 2021. arXiv:2109.08229.

Timothy B. Armstrong. Asymptotic efficiency bounds for a class of experimental designs, 2022. arXiv:2205.02726.

S. Athey and G.W. Imbens. Chapter 3 - the econometrics of randomized experimentsa. In *Handbook of Field Experiments*, volume 1 of *Handbook of Economic Field Experiments*, pp. 73–140. North-Holland, 2017.

Alexia Atsidakou, Sumeet Katariya, Sujay Sanghavi, and Branislav Kveton. Bayesian fixed-budget best-arm identification, 2023. arXiv:2211.08572.

Jean-Yves Audibert, Sébastien Bubeck, and Remi Munos. Best arm identification in multi-armed bandits. In *Conference on Learning Theory*, pp. 41–53, 2010.

R. R. Bahadur. Stochastic Comparison of Tests. *The Annals of Mathematical Statistics*, 31(2):276 – 295, 1960.

R. R. Bahadur. Rates of Convergence of Estimates and Test Statistics. *The Annals of Mathematical Statistics*, 38(2):303 – 324, 1967.

R. R. Bahadur. *1. Some Limit Theorems in Statistics*, pp. 1–40. Society for Industrial and Applied Mathematics, 1971.

R.E. Bechhofer, J. Kiefer, and M. Sobel. *Sequential Identification and Ranking Procedures: With Special Reference to Koopman-Darmois Populations*. University of Chicago Press, 1968.

Robert E. Bechhofer. A Single-Sample Multiple Decision Procedure for Ranking Means of Normal Populations with known Variances. *The Annals of Mathematical Statistics*, 25(1):16 – 39, 1954.

P. J. Bickel, C. A. J. Klaassen, Y. Ritov, and J. A. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*. Springer, 1998.

Jürgen Branke, Stephen E. Chick, and Christian Schmidt. Selecting a selection procedure. *Management Science*, 53(12):1916–1932, 2007.

Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. Pure exploration in multi-armed bandits problems. In *Algorithmic Learning Theory*, pp. 23–37. Springer Berlin Heidelberg, 2009.

Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. Pure exploration in finitely-armed and continuous-armed bandits. *Theoretical Computer Science*, 2011.

Alexandra Carpentier and Andrea Locatelli. Tight (lower) bounds for the fixed budget best arm identification bandit problem. In *COLT*, 2016.

CBER CDER. Adaptive designs for clinical trials of drugs and biologics guidance for industry draft guidance, 05 2018.

Chun-Hung Chen, Jianwu Lin, Enver Yücesan, and Stephen E. Chick. Simulation budget allocation for further enhancing theefficiency of ordinal optimization. *Discrete Event Dynamic Systems*, 10(3):251–270, 2000.

Herman Chernoff. Sequential Design of Experiments. *The Annals of Mathematical Statistics*, 30(3):755 – 770, 1959.

Shein-Chung Chow and Mark Chang. *Adaptive Design Methods in Clinical Trials*. Chapman and Hall/CRC, 2 edition, 2011.

Harald Cramér. Sur un nouveau théorème-limite de la théorie des probabilités. In *Colloque consacré à la théorie des probabilités*, volume 736, pp. 2–23. Hermann, 1938.

Rémy Degenne. On the existence of a complexity in fixed budget bandit identification. In *Conference on Learning Theory*, volume 195, pp. 1131–1154. PMLR, 2023.

A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications*. Stochastic Modelling and Applied Probability. Springer Berlin Heidelberg, 2009.

Richard S. Ellis. Large Deviations for a General Class of Random Vectors. *The Annals of Probability*, 12(1): 1 – 12, 1984.

Eyal Even-Dar, Shie Mannor, Yishay Mansour, and Sridhar Mahadevan. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of machine learning research*, 2006.

R. A. Fisher. *The Design of Experiments*. Oliver and Boyd, Edinburgh, 1935.

Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *Conference on Learning Theory*, 2016.

Jürgen Gärtner. On large deviations from the invariant measure. *Theory of Probability & Its Applications*, 22(1):24–39, 1977.

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

Shantanu Gupta, Zachary Chase Lipton, and David Childers. Efficient online estimation of causal effects by deciding what to observe. In *Advances in Neural Information Processing Systems*, 2021.

S.S. Gupta. *On a Decision Rule for a Problem in Ranking Means*. University of North Carolina at Chapel Hill, 1956.

Vitor Hadad, David A. Hirshberg, Ruohan Zhan, Stefan Wager, and Susan Athey. Confidence intervals for policy evaluation in adaptive experiments. *Proceedings of the National Academy of Sciences*, 118(15), 2021.

Jinyong Hahn. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, 66(2):315–331, 1998.

Jinyong Hahn, Keisuke Hirano, and Dean Karlan. Adaptive experimental design using the propensity score. *Journal of Business and Economic Statistics*, 2011.

Keisuke Hirano and Jack R. Porter. Asymptotics for statistical treatment rules. *Econometrica*, 77(5): 1683–1701, 2009.

Keisuke Hirano and Jack R. Porter. Asymptotic representations for sequential decisions, adaptive experiments, and batched bandits, 2023. arXiv:2302.03117.

Y. C. Ho, R. S. Sreenivas, and P. Vakili. Ordinal optimization of deds. *Discrete Event Dynamic Systems: Theory and Applications*, 2(1):61–88, July 1992.

L. Jeff Hong, Weiwei Fan, and Jun Luo. Review on ranking and selection: A new perspective. *Frontiers of Engineering Management*, 8(3):321–343, mar 2021.

Marc Jourdan, Rémy Degenne, Dorian Baudry, Rianne de Heide, and Emilie Kaufmann. Top two algorithms revisited, 2022.

Marc Jourdan, Degenne Rémy, and Kaufmann Emilie. Dealing with unknown variances in best-arm identification. In *Proceedings of The 34th International Conference on Algorithmic Learning Theory*, volume 201, pp. 776–849, 2023.

Dean Karlan and Daniel H Wood. The effect of effectiveness: Donor response to aid effectiveness in a direct mail fundraising experiment. Working Paper 20047, National Bureau of Economic Research, April 2014.

Maximilian Kasy and Anja Sautmann. Adaptive treatment assignment in experiments for policy choice. *Econometrica*, 89(1):113–132, 2021.

Masahiro Kato, Takuya Ishihara, Junya Honda, and Yusuke Narita. Efficient adaptive experimental design for average treatment effect estimation, 2020. arXiv:2002.05308.

Masahiro Kato, Masaaki Imaizumi, Takuya Ishihara, and Toru Kitagawa. Asymptotically minimax optimal fixed-budget best arm identification for expected simple regret minimization, 2023a. arXiv:2302.02988.

Masahiro Kato, Masaaki Imaizumi, Takuya Ishihara, and Toru Kitagawa. Fixed-budget hypothesis best arm identification: On the information loss in experimental design. In *ICML Workshop on New Frontiers in Learning, Control, and Dynamical Systems*, 2023b.

Emilie Kaufmann. *Contributions to the Optimal Solution of Several Bandits Problems*. Habilitation á Diriger des Recherches, Université de Lille, 2020. URL `https://emiliekaufmann.github.io/HDR_EmilieKaufmann.pdf`.

Emilie Kaufmann and Wouter M. Koolen. Mixture martingales revisited with applications to sequential tests and confidence intervals. *Journal of Machine Learning Research*, 22(246):1–44, 2021.

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

Junpei Komiyama, Kaito Ariu, Masahiro Kato, and Chao Qin. Optimal simple regret in bayesian best arm identification, 2021.

Junpei Komiyama, Taira Tsuchiya, and Junya Honda. Minimax optimal algorithms for fixed-budget best arm identification. In *Advances in Neural Information Processing Systems*, 2022.

T.L Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 1985.

Anusha Lalitha, Kousha Kalantari, Yifei Ma, Anoop Deoras, and Branislav Kveton. Fixed-budget best-arm identification with heterogeneous reward variances. In *Conference on Uncertainty in Artificial Intelligence*, 2023.

T. Lattimore and C. Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

L Le Cam. *Locally Asymptotically Normal Families of Distributions. Certain Approximations to Families of Distributions and Their Use in the Theory of Estimation and Testing Hypotheses*. University of California Publications in Statistics. vol. 3. no. 2. Berkeley & Los Angeles, 1960.

L Le Cam. Limits of experiments. In *Theory of Statistics*, pp. 245–282. University of California Press, 1972.

Lucien Le Cam. *Asymptotic Methods in Statistical Decision Theory (Springer Series in Statistics)*. Springer, 1986.

Pinyan Lu, Chao Tao, and Xiaojin Zhang. Variance-dependent best arm identification. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161, pp. 1120–1129, 2021.

Charles Manski. Identification problems and decisions under ambiguity: Empirical analysis of treatment response and normative analysis of treatment choice. *Journal of Econometrics*, 95(2):415–442, 2000.

Charles F. Manski. Treatment choice under ambiguity induced by inferential problems. *Journal of Statistical Planning and Inference*, 105(1):67–82, 2002.

Charles F. Manski. Statistical treatment rules for heterogeneous populations. *Econometrica*, 72(4):1221–1246, 2004.

Brijesh Nair. Clinical trial designs. *Indian Dermatol. Online J.*, 10(2):193–201, March 2019.

Jerzy Neyman. Sur les applications de la theorie des probabilites aux experiences agricoles: Essai des principes. *Statistical Science*, 5:463–472, 1923.

Jerzy Neyman. On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97:123–150, 1934.

Taisuke Otsu. Large deviation asymptotics for statistical treatment rules. *Economics Letters*, 101(1):53–56, 2008.

Edward Paulson. A Sequential Procedure for Selecting the Population with the Largest Mean from k Normal Populations. *The Annals of Mathematical Statistics*, 1964.

C. S. Peirce and Joseph Jastrow. On small differences in sensation. *Memoirs of the National Academy of Sciences*, 3:75–83, 1884.

Charles Sanders Peirce and Cornelis de Waal. *Illustrations of the Logic of Science*. Chicago, Illinois: Open Court, 1887.

A. Pong and S.-C Chow. *Handbook of adaptive designs in pharmaceutical and clinical development*. Chapman and Hall/CRC, 04 2016.

Chao Qin. Open problem: Optimal best arm identification with fixed-budget. In *Conference on Learning Theory*, 2022.

Chao Qin, Diego Klabjan, and Daniel Russo. Improving the expected improvement algorithm. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Herbert Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 1952.

Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 1974.

Daniel Russo. Simple bayesian algorithms for best-arm identification. *Operations Research*, 68(6):1625–1647, 2020.

Luigi Sauro. Rapidly finding the best arm using variance. In *European Conference on Artificial Intelligence*, 2020.

Xuedong Shang, Rianne de Heide, Pierre Menard, Emilie Kaufmann, and Michal Valko. Fixed-confidence guarantees for bayesian best-arm identification. In *International Conference on Artificial Intelligence and Statistics*, volume 108, pp. 1823–1832, 2020.

Dongwook Shin, Mark Broadie, and Assaf Zeevi. Tractable sampling strategies for ordinal optimization. *Operations Research*, 66(6):1693–1712, 2018.

Max Tabord-Meehan. Stratification Trees for Adaptive Randomization in Randomized Controlled Trials. *The Review of Economic Studies*, 12 2022.

William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 1933.

Mark J. van der Laan. The construction and analysis of adaptive group sequential designs, 2008. URL `https://biostats.bepress.com/ucbbiostat/paper232`.

A.W. van der Vaart. An asymptotic representation theorem. *International Statistical Review / Revue Internationale de Statistique*, 59(1):97–121, 1991.

A.W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

Stefan Wager and Kuang Xu. Diffusion asymptotics for sequential experiments, 2021. arXiv:2101.09855.

A. Wald. Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16(2):117–186, 1945.

Abraham Wald. Statistical Decision Functions. *The Annals of Mathematical Statistics*, 20(2):165 – 205, 1949.

Po-An Wang, Kaito Ariu, and Alexandre Proutiere. On uniformly optimal algorithms for best arm identification in two-armed bandits with fixed budget, 2023. arXiv:2308.12000.

A Related Work

This section introduces related work.

A.1 Historical Background of Ordinal Optimization and BAI

Researchers have acknowledged the importance of statistical inference and experimental approaches as essential scientific tools (Peirce & Jastrow, 1884; Peirce & de Waal, 1887). With the advancement of these statistical methodologies, experimental design also began attracting attention. Fisher (1935) develops the groundwork for the principles of experimental design. Wald (1949) establishes fundamental theories for statistical decision-making, bridging statistical inference and decision-making. These methodologies have been investigated across various disciplines, such as medicine, epidemiology, economics, operations research, and computer science, transcending their origins in statistics.

Ordinal Optimization. Ordinal optimization involves sample allocation to each treatment arm and selects a certain treatment arm based on a decision-making criterion; therefore, this problem is also known as the optimal computing budget allocation problem. The development of ordinal optimization is closely related to ranking and selection problems in simulation, originating from agricultural and clinical applications in the 1950s (Gupta, 1956; Bechhofer, 1954; Paulson, 1964; Branke et al., 2007; Hong et al., 2021). A modern formulation of ordinal optimization was established in the early 2000s (Chen et al., 2000; Glynn & Juneja, 2004). Existing research has found that the probability of misidentification converges at an exponential rate for a large set of problems. By employing large deviation principles (Cramér, 1938; Ellis, 1984; Gärtner, 1977; Dembo & Zeitouni, 2009), Glynn & Juneja (2004) propose asymptotically optimal algorithms for ordinal optimization.

BAI. A promising idea for enhancing the efficiency of strategies is adaptive experimental design. In this approach, information from past trials can be utilized to optimize the allocation of samples in subsequent trials. The concept of adaptive experimental design dates back to the 1970s Pong & Chow (2016). Presently, its significance is acknowledged (CDER, 2018; Chow & Chang, 2011). Adaptive strategies have also been studied within the domain of machine learning, and the multi-armed bandit (MAB) problem (Thompson, 1933; Robbins, 1952; Lai & Robbins, 1985) is an instance. The Best Arm Identification (BAI) is a paradigm of this problem (Even-Dar et al., 2006; Audibert et al., 2010; Bubeck et al., 2011), influenced by sequential testing, ranking, selection problems, and ordinal optimization (Bechhofer et al., 1968). There are two formulations in BAI: fixed-confidence (Garivier & Kaufmann, 2016) and fixed-budget BAI. In the former, the sample size (budget) is a random variable, and a decision-maker stops an experiment when a certain criterion is satisfied, similar to sequential testing Wald (1945); Chernoff (1959). In contrast, the latter fixes the sample size (budget) and minimizes a certain criterion given the sample size. BAI in this study corresponds to fixed-budget BAI (Bubeck et al., 2011; Audibert et al., 2010; Bubeck et al., 2011). There is no strict distinction between ordinal optimization and BAI.

A.2 Optimal Strategies for BAI

We introduce arguments about optimal strategies for BAI.

Optimal Strategies for Fixed-Confidence BAI. In fixed-confidence BAI, several optimal strategies have been proposed, whose expected stopping time aligns with lower bounds shown by Kaufmann et al. (2016). One of the remarkable studies is Garivier & Kaufmann (2016), which proposes an optimal strategy called Track-and-Stop by extending the Chernoff stopping rule. The Track-and-Stop strategy is refined and extended by the following studies, including Kaufmann & Koolen (2021), Jourdan et al. (2022), and Jourdan et al. (2023). From a Bayesian perspective, Russo (2020), Qin et al. (2017), and Shang et al. (2020) propose Bayesian BAI strategies that are optimal in terms of posterior convergence rate.

Optimal Strategies for Fixed-Budget BAI. Fixed-budget BAI has also been extensively studied, but the asymptotic optimality has open issues. Kaufmann et al. (2016) and Carpentier & Locatelli (2016)

conjecture lower bounds. Garivier & Kaufmann (2016), Kaufmann (2020), Ariu et al. (2021), and Qin (2022) discuss and summarize the problem. In parallel with our study, Komiyama et al. (2022), Degenne (2023), Atsidakou et al. (2023), and Wang et al. (2023) further discuss the problem.

Instead of a tight evaluation of the probability of misidentification, several studies focus on evaluating the expected simple regret. Bubeck et al. (2011) discusses the optimality of uniform sampling. Kato et al. (2023a) demonstrates the asymptotic optimality of a variance-dependent strategy. Komiyama et al. (2021) develops a Bayes optimal strategy. Kato et al. (2023a) shows that variance-dependent allocation improves the expected simple regret compared to uniform sampling in Bubeck et al. (2011). However, there is a constant gap between the lower and upper bounds in Kato et al. (2023a), and it is still unknown whether allocation rules in optimal strategies depend on variances when we consider tighter evaluation in the probability of misidentification.

A.3 Complexity of strategies and bahadur efficiency

The complexity $-\frac{1}{T}\log \mathbb{P}_{P_0}(\widehat{a}_T^{\pi} \neq a^*)$, has been widely adopted in the literature of ordinal optimization and BAI (Glynn & Juneja, 2004; Kaufmann et al., 2016). In the field of hypothesis testing, Bahadur (1960) suggests the use of a similar measure to assess statistics in hypothesis testing. The efficiency of a test under the criterion proposed by Bahadur (1960) is known as Bahadur efficiency, and the complexity is referred to as the Bahadur slope. Although our problem is not hypothesis testing, it can be considered that our asymptotic optimality of strategies corresponds to the concept of Bahadur efficiency. Moreover, our global asymptotic optimality parallels global Bahadur efficiency.

A.4 Efficient Average Treatment Effect Estimation

Efficient estimation of ATE via adaptive experiments constitutes another area of related literature. van der Laan (2008) and Hahn et al. (2011) propose experimental design methods for more efficient estimation of ATE by utilizing covariate information in treatment assignment. Despite the marginalization of covariates, their methods are able to reduce the asymptotic variance of estimators. Karlan & Wood (2014) applies the method of Hahn et al. (2011) to examine the response of donors to new information regarding the effectiveness of a charity. Subsequently, Tabord-Meehan (2022) and Kato et al. (2020) have sought to improve upon these studies, and more recently, Gupta et al. (2021) has proposed the use of instrumental variables in this context.

A.5 Other related work.

Our problem has close ties to theories of statistical decision-making (Wald, 1949; Manski, 2000; 2002; 2004), limits of experiments (Le Cam, 1972; van der Vaart, 1998), and semiparametric theory (Hahn, 1998). The semiparametric theory is particularly crucial as it enables the characterization of lower bounds through the semiparametric analog of Fisher information (van der Vaart, 1998).

Adusumilli (2022; 2023) present a minimax evaluation of bandit strategies for both regret minimization and BAI, which is based on a formulation utilizing a diffusion process proposed by Wager & Xu (2021). Furthermore, Armstrong (2022) and Hirano & Porter (2023) extend the results of Hirano & Porter (2009) to a setting of adaptive experiments. The results of Adusumilli (2022; 2023) and Armstrong (2022) employ arguments on local asymptotic normality (Le Cam, 1960; 1972; 1986; van der Vaart, 1991; 1998), where the class of alternative hypotheses comprises "local models," in which parameters of interest converge to true parameters at a rate of $1/\sqrt{T}$

Variance-dependent strategies have garnered attention in BAI. In fixed-confidence BAI, Jourdan et al. (2023) provides a detailed discussion about BAI strategies with unknown variances. There are also studies using variances in strategies, including Sauro (2020), Lu et al. (2021), and Lalitha et al. (2023). However, they do not discuss optimality based on the arguments of Kaufmann et al. (2016). In fact, our proposed strategy differs from the one in Lalitha et al. (2023), and our results imply that at least the worst-case optimal strategy is not the one in Lalitha et al. (2023).

B Proof of Lemma 2.4

Proof of Lemma 2.4. From Lemma 2.2 and equation 1 in Definition 2.3, there exists w^π such that

$$\limsup_{T\to\infty} -\frac{1}{T}\log \mathbb{P}_{P_0}(\widehat{a}^\pi_T \neq a^*) \leq \inf_{Q\in\cup_{b\neq a^*}\mathcal{P}(b,\underline{\Delta},\overline{\Delta})} \sum_{a\in[K]} w^\pi(a)\mathrm{KL}(Q^a, P_0^a).$$

Then, we bound the probability as

$$\limsup_{T\to\infty} -\frac{1}{T}\log \mathbb{P}_{P_0}(\widehat{a}^\pi_T \neq a^*) \leq \sup_{w\in\mathcal{W}} \inf_{Q\in\cup_{b\neq a^*}\mathcal{P}(b,\underline{\Delta},\overline{\Delta})} \sum_{a\in[K]} w(a)\mathrm{KL}(Q^a, P_0^a).$$

Note that $w^\pi(a)$ is independent of Q. Because $\mathrm{KL}(Q^a, P_0^a)$ is given as $\frac{(\mu^a(Q)-\mu^a(P_0))^2}{2(\sigma^a)^2}$ for $P, Q \in \cup_{b\neq a^*}\mathcal{P}(b,\underline{\Delta},\overline{\Delta})$, we obtain

$$\limsup_{T\to\infty} -\frac{1}{T}\log \mathbb{P}_{P_0}(\widehat{a}^\pi_T \neq a^*) \leq \sup_{w\in\mathcal{W}} \inf_{\substack{(\mu^b)\in\mathbb{R}^K:\\ \arg\max_{b\in[K]}\mu^b\neq a^*}} \sum_{a\in[K]} w(a)\frac{(\mu^a-\mu^a(P_0))^2}{2\left(\sigma^a\right)^2}.$$

Here, we have

$$\begin{aligned}
&\inf_{\substack{(\mu^b)\in\mathbb{R}^K:\\ \arg\max_{b\in[K]}\mu^b\neq a^*}} \sum_{a\in[K]} w(a)\frac{(\mu^a-\mu^a(P_0))^2}{2\left(\sigma^a\right)^2} \\
&= \min_{a\in[K]\setminus\{a^*\}} \inf_{\substack{(\mu^b)\in\mathbb{R}^K:\\ \mu^a>\mu^{a^*}}} \sum_{a\in[K]} w(a)\frac{(\mu^a-\mu^a(P_0))^2}{2\left(\sigma^a\right)^2} \\
&= \min_{a\in[K]\setminus\{a^*\}} \inf_{\substack{(\mu^b)\in\mathbb{R}^K:\\ \mu^a>\mu^{a^*},\ \mu^c=\mu^c(P_0)}} \sum_{a\in[K]} w(a)\frac{(\mu^a-\mu^a(P_0))^2}{2\left(\sigma^a\right)^2} \\
&= \min_{a\in[K]\setminus\{a^*\}} \inf_{\substack{(\mu^{a^*},\mu^a)\in\mathbb{R}^K:\\ \mu^a>\mu^{a^*}}} \left\{ w(a^*)\frac{\left(\mu^{a^*}-\mu^{a^*}(P_0)\right)^2}{2\left(\sigma^{a^*}\right)^2} + w(a)\frac{(\mu^a-\mu^a(P_0))^2}{2\left(\sigma^a\right)^2} \right\} \\
&= \min_{a\in[K]\setminus\{a^*\}} \min_{\mu\in\left[\mu^a(P_0),\mu^{a^*}\right]} \left\{ w(a^*)\frac{\left(\mu-\mu^{a^*}(P_0)\right)^2}{2\left(\sigma^{a^*}\right)^2} + w(a)\frac{(\mu-\mu^a(P_0))^2}{2\left(\sigma^a\right)^2} \right\}.
\end{aligned}$$

Then, by solving the optimization problem, we obtain

$$\begin{aligned}
&\min_{a\in[K]\setminus\{a^*\}} \min_{\mu\in\left[\mu^a(P_0),\mu^{a^*}\right]} \left\{ w(a^*)\frac{\left(\mu-\mu^{a^*}(P_0)\right)^2}{2\left(\sigma^{a^*}\right)^2} + w(a)\frac{(\mu-\mu^a(P_0))^2}{2\left(\sigma^a\right)^2} \right\} \\
&= \min_{a\in[K]\setminus\{a^*\}} \frac{\left(\mu^{a^*}(P_0)-\mu^a(P_0)\right)^2}{2\left(\frac{\left(\sigma^{a^*}\right)^2}{w(a^*)}+\frac{\left(\sigma^a\right)^2}{w(a)}\right)}.
\end{aligned}$$

Thus, we complete the proof. □

C Proofs of Theorem 2.5

Proof. Because there exists $\overline{\Delta}$ such that $\mu^{a^*}(P_0)-\mu^a(P_0)\leq\overline{\Delta}$ for all $a\in[K]$, the lower bound is given as

$$\min_{a\in[K]\setminus\{a^*\}} \frac{(\Delta^a(P_0))^2}{2\left(\frac{\left(\sigma^{a^*}\right)^2}{w(a^*)}+\frac{\left(\sigma^a\right)^2}{w(a)}\right)} \leq \min_{a\in[K]\setminus\{a^*\}} \frac{\overline{\Delta}^2}{2\left(\frac{\left(\sigma^{a^*}\right)^2}{w(a^*)}+\frac{\left(\sigma^a\right)^2}{w(a)}\right)}.$$

Therefore, we consider solving

$$\max_{w\in\mathcal{W}} \min_{a\neq a^*} \frac{1}{\frac{\left(\sigma^{a^*}\right)^2}{w(a^*)} + \frac{(\sigma^a)^2}{w(a)}}.$$

We consider maximising $R > 0$ such that $R \leq 1/\left\{\frac{\left(\sigma^{a^*}\right)^2}{w(a^*)} + \frac{(\sigma^a)^2}{w(a)}\right\}$ for all $a \in [K]\backslash\{a^*\}$ by optimizing $w \in \mathcal{W}$. That is, we consider the following non-linear programming:

$$\begin{aligned}
&\max_{R>0, \boldsymbol{w}=\{w(1),w(2)\ldots,w(K)\}\in(0,1)^K} \quad R \\
\text{s.t.} \quad & R\left(\frac{\left(\sigma^{a^*}\right)^2}{w(a^*)} + \frac{(\sigma^a)^2}{w(a)}\right)\zeta - 1 \leq 0 \qquad \forall a \in [K]\backslash\{a^*\}, \\
& \sum_{a\in[K]} w(a) - 1 = 0, \\
& w(a) > 0 \qquad \forall a \in [K].
\end{aligned}$$

The maximum of R in the constraint optimization is equal to $\max_{w\in\mathcal{W}} \min_{a\neq a^*} \frac{1}{\frac{\left(\sigma^{a^*}\right)^2}{w(a^*)} + \frac{(\sigma^a)^2}{w(a)}}$.

Then, for $(K-1)$ Lagrangian multiplies $\boldsymbol{\lambda} = \{\lambda^a\}_{a\in[K]\backslash\{a^*\}}$ and γ such that $\lambda^a \leq 0$ and $\gamma \in \mathbb{R}$, we define the following Lagrangian function:

$$L(\boldsymbol{\lambda}, \boldsymbol{\gamma}; R, \boldsymbol{w}) = R + \sum_{a\in[K]\backslash\{a^*\}} \lambda^a \left\{R\left(\frac{\left(\sigma^{a^*}\right)^2}{w(a^*)} + \frac{(\sigma^a)^2}{w(a)}\right) - 1\right\} - \gamma\left\{\sum_{a\in[K]} w(a) - 1\right\}.$$

Note that the objective (R) and constraints $\left(R\left(\frac{\left(\sigma^{a^*}\right)^2}{w(a^*)} + \frac{(\sigma^a)^2}{w(a)}\right) - 1 \leq 0 \text{ and } \sum_{a\in[K]} w(a) - 1 = 0\right)$ are differentiable convex functions for R and $\boldsymbol{w}$. Therefore, the global optimizer $R^\dagger$ and $\boldsymbol{w}^\dagger = \{w^\dagger(a)\} \in (0,1)^{KN}$ satisfies the KKT condition; that is, there are Lagrangian multipliers $\lambda^{a\dagger}$, $\gamma^\dagger$, and $R^\dagger$ such that

$$1 + \sum_{a\in[K]\backslash\{a^*\}} \lambda^{a\dagger}\left(\frac{\left(\sigma^{a^*}\right)^2}{w^\dagger(a^*)} + \frac{(\sigma^a)^2}{w^\dagger(a)}\right) = 0 \tag{6}$$

$$-2\sum_{a\in[K]\backslash\{a^*\}} \lambda^{a\dagger} R^\dagger \frac{\left(\sigma^{a^*}\right)^2}{(w^\dagger(a^*))^2} = \gamma^\dagger \tag{7}$$

$$-2\lambda^{a\dagger} R^\dagger \frac{(\sigma^a)^2}{(w^\dagger(a))^2} = \gamma^\dagger \qquad \forall a \in [K]\backslash\{a^*\} \tag{8}$$

$$\lambda^{a\dagger}\left\{R^\dagger\left(\frac{\left(\sigma^{a^*}\right)^2}{w^\dagger(a^*)} + \frac{(\sigma^a)^2}{w^\dagger(a)}\right) - 1\right\} = 0 \qquad \forall a \in [K]\backslash\{a^*\} \tag{9}$$

$$\gamma^\dagger\left\{\sum_{c\in[K]} w^\dagger(c) - 1\right\} = 0$$

$$\lambda^{a\dagger} \leq 0 \qquad \forall a \in [K]\backslash\{a^*\}.$$

Here, equation 6 implies $\lambda^{a\dagger} < 0$ for some $a \in [K]\backslash\{a^*\}$. This is because if $\lambda^{a\dagger} = 0$ for all $a \in [K]\backslash\{a^*\}$, $1 + 0 = 1 \neq 0$.

With $\lambda^{a\dagger} < 0$, since $-\lambda^{a\dagger} R^\dagger \frac{(\sigma^a)^2}{(w^\dagger(a))^2} > 0$ for all $a \in [K]$, it follows that $\gamma^\dagger > 0$. This also implies that $\sum_{c\in[K]} w^{c\dagger} - 1 = 0$.

Then, equation 9 implies that

$$R^\dagger \left(\frac{\left(\sigma^{a^*}\right)^2}{w^\dagger(a^*)} + \frac{\left(\sigma^a\right)^2}{w^\dagger(a)} \right) = 1 \qquad \forall a \in [K]\backslash\{a^*\}.$$

Therefore, we have

$$\frac{\left(\sigma^a\right)^2}{w^\dagger(a)} = \frac{\left(\sigma^b(P_0)\right)^2}{w^\dagger(b)} \qquad \forall a, b \in [K]\backslash\{a^*\}. \tag{10}$$

Let $\frac{(\sigma^a)^2}{w^\dagger(a)} = \frac{\left(\sigma^b(P_0)\right)^2}{w^\dagger(b)} = \frac{1}{R^\dagger} - \frac{\left(\sigma^{a^*}\right)^2}{w^\dagger(a^*)} = U$. From equation 10 and equation 6,

$$\sum_{b\in[K]\backslash\{a^*\}} \lambda^{b\dagger} = -\frac{1}{\frac{\left(\sigma^{a^*}\right)^2}{w^\dagger(a^*)} + U} \tag{11}$$

From equation 7 and equation 8, we have

$$\frac{\left(\sigma^{a^*}\right)^2}{(w^\dagger(a^*))^2} \sum_{b\in[K]\backslash\{a^*\}} \lambda^{b\dagger} = \lambda^{a\dagger} \frac{\left(\sigma^a\right)^2}{(w^\dagger(a))^2} \qquad \forall a \in [K]\backslash\{a^*\}. \tag{12}$$

From equation 11 and equation 12, we have

$$-\frac{\left(\sigma^{a^*}\right)^2}{(w^\dagger(a^*))^2} = \lambda^{a\dagger} \frac{\left(\sigma^a\right)^2}{(w^\dagger(a))^2} \left(\frac{\left(\sigma^{a^*}\right)^2}{w^\dagger(a^*)} + U \right) \qquad \forall a \in [K]\backslash\{a^*\}. \tag{13}$$

From equation 6 and equation 13, we have

$$w^\dagger(a^*) = \sqrt{\left(\sigma^{a^*}\right)^2 \sum_{a\in[K]\backslash\{a^*\}} \frac{(w^\dagger(a))^2}{\left(\sigma^a\right)^2}}.$$

In summary, we have the following KKT conditions:

$$\begin{aligned}
&w^\dagger(a^*) = \sqrt{\left(\sigma^{a^*}\right)^2 \sum_{a\in[K]\backslash\{a^*\}} \frac{(w^\dagger(a)^2}{\left(\sigma^a\right)^2}} \\
&\frac{\left(\sigma^{a^*}\right)^2}{(w^\dagger(a^*))^2} = -\lambda^{a\dagger} \frac{\left(\sigma^a\right)^2}{(w^\dagger(a))^2} \left(\left(\frac{\left(\sigma^{a^*}\right)^2}{w^\dagger(a^*)} + \frac{\left(\sigma^a\right)^2}{w^\dagger(a)} \right) \right) \qquad \forall a \in [K]\backslash\{a^*\} \\
&-\lambda^{a\dagger} \frac{\left(\sigma^a\right)^2}{(w^\dagger(a))^2} = \widetilde{\gamma}^\dagger \qquad \forall a \in [K]\backslash\{a^*\} \\
&\frac{\left(\sigma^a\right)^2}{w^\dagger(a)} = \frac{1}{R^\dagger} - \frac{\left(\sigma^{a^*}\right)^2}{w^\dagger(a^*)} \qquad \forall a \in [K]\backslash\{a^*\} \\
&\sum_{a\in[K]} w^\dagger(a) = 1 \\
&\lambda^{a\dagger} \leq 0 \qquad \forall a \in [K]\backslash\{a^*\},
\end{aligned}$$

where $\widetilde{\gamma}^\dagger = \gamma^\dagger / 2R^\dagger$. From $w^\dagger(a^*) = \sqrt{\left(\sigma^{a^*}\right)^2 \sum_{a\in[K]\backslash\{a^*\}} \frac{(w^\dagger(a))^2}{\left(\sigma^a\right)^2}}$ and $-\lambda^{a\dagger} \frac{(\sigma^a)^2}{(w^\dagger(a))^2} = \widetilde{\gamma}^\dagger$, we have

$$w^\dagger(a^*) = \sigma^{a^*} \sqrt{\sum_{a\in[K]\backslash\{a^*\}} -\lambda^{a\dagger}} / \sqrt{\widetilde{\gamma}^\dagger}$$

$$w^\dagger(a) = \sqrt{-\lambda^{a\dagger}/\widetilde{\gamma}^\dagger}\sigma^a.$$

From $\sum_{a\in[K]} w^\dagger(a) = 1$, we have

$$\sigma^{a^*}\sqrt{\sum_{a\in[K]\setminus\{a^*\}} -\lambda^{a\dagger}}/\sqrt{\widetilde{\gamma}^\dagger} + \sum_{a\in[K]\setminus\{a^*\}} \sqrt{-\lambda^{a\dagger}/\widetilde{\gamma}^\dagger}\sigma^a = 1.$$

Therefore, the following holds:

$$\sqrt{\widetilde{\gamma}^\dagger} = \sigma^{a^*}\sqrt{\sum_{a\in[K]\setminus\{a^*\}} -\lambda^{a\dagger}} + \sum_{a\in[K]\setminus\{a^*\}} \sqrt{-\lambda^{a\dagger}}\sigma^a.$$

Hence, the target allocation ratio is computed as

$$\begin{aligned}
w^\dagger(a^*) &= \frac{\sigma^{a^*}\sqrt{\sum_{a\in[K]\setminus\{a^*\}} -\lambda^{a\dagger}}}{\sigma^{a^*}\sqrt{\sum_{a\in[K]\setminus\{a^*\}} -\lambda^{a\dagger}} + \sum_{a\in[K]\setminus\{a^*\}} \sqrt{-\lambda^{a\dagger}}\sigma^a} \\
w^\dagger(a) &= \frac{\sqrt{-\lambda^{a\dagger}}\sigma^a}{\sigma^{a^*}\sqrt{\sum_{a\in[K]\setminus\{a^*\}} -\lambda^{a\dagger}} + \sum_{a\in[K]\setminus\{a^*\}} \sqrt{-\lambda^{a\dagger}}\sigma^a},
\end{aligned}$$

where from $\frac{\left(\sigma^{a^*}\right)^2}{(w^\dagger(a^*))^2} = -\lambda^{a\dagger}\frac{(\sigma^a)^2}{(w^\dagger(a))^2}\left(\frac{\left(\sigma^{a^*}\right)^2}{w^\dagger(a^*)} + \frac{(\sigma^a)^2}{w^\dagger(a)}\right)$, $(\lambda^{a^*})_{a\in[K]\setminus\{a^*\}}$ satisfies,

$$\begin{aligned}
&\frac{1}{\sum_{a\in[K]\setminus\{a^*\}} -\lambda^{a\dagger}} \\
&= \left(\frac{\sigma^{a^*}}{\sqrt{\sum_{a\in[K]\setminus\{a^*\}} -\lambda^{a\dagger}}} + \frac{\sigma^a}{\sqrt{-\lambda^{a\dagger}}}\right)\left(\sigma^{a^*}\sqrt{\sum_{c\in[K]\setminus\{a^*\}} -\lambda^{c\dagger}} + \sum_{c\in[K]\setminus\{a^*\}} \sqrt{-\lambda^{c\dagger}}\sigma_0^c\right) \\
&= \left(\sigma^{a^*} + \frac{\sigma^a}{\sqrt{-\lambda^{a\dagger}}}\sqrt{\sum_{c\in[K]\setminus\{a^*\}} -\lambda^{c\dagger}}\right)\left(\sigma^{a^*} + \frac{\sum_{c\in[K]\setminus\{a^*\}} \sqrt{-\lambda^{c\dagger}}\sigma_0^c}{\sum_{c\in[K]\setminus\{a^*\}} -\lambda^{c\dagger}}\sqrt{\sum_{c\in[K]\setminus\{a^*\}} -\lambda^{c\dagger}}\right).
\end{aligned}$$

Then, the following solutions satisfy the above KKT conditions:

$$\begin{aligned}
R^\dagger\left(\sigma^{a^*} + \sqrt{\sum_{b\in[K]\setminus\{a^*\}} \left(\sigma^b\right)^2}\right)^2 &= 1 \\
w^\dagger(a^*) &= \frac{\sigma^{a^*}\sqrt{\sum_{b\in[K]\setminus\{a^*\}} \left(\sigma^b\right)^2}}{\sigma^{a^*}\sqrt{\sum_{b\in[K]\setminus\{a^*\}} \left(\sigma^b\right)^2} + \sum_{b\in[K]\setminus\{a^*\}} \left(\sigma^b\right)^2} \\
w^\dagger(a) &= \frac{\left(\sigma^a\right)^2}{\sigma^{a^*}\sqrt{\sum_{b\in[K]\setminus\{a^*\}} \left(\sigma^b\right)^2} + \sum_{b\in[K]\setminus\{a^*\}} \left(\sigma^b\right)^2} \\
\lambda^{a\dagger} &= -\left(\sigma^a\right)^2 \\
\gamma^\dagger &= 2\left(\sigma^a\right)^2.
\end{aligned}$$

□

Note that a target allocation ratio w in the maximum corresponds to a limit of an expectation of allocation rule $\frac{1}{T}\sum_{t=1}^T \mathbb{1}[A_t = a]$ from the definition of asymptotically invariant strategies.

D Proof of Theorem 4.1

Proof of Theorem 4.1. Note that the probability of misidentification can be written as

$$\mathbb{P}_{P_0}\left(\widehat{\mu}_t^{a^*} \leq \widehat{\mu}_t^a\right) = \mathbb{P}_{P_0}\left(\sum_{t=1}^{T}\left\{\Psi_t^{a^*}(P_0) + \Psi_t^a(P_0)\right\} \leq -T\Delta^a(P_0)\right),$$

where

$$\Psi_t^{a^*}(P_0) = \frac{\mathbb{1}[A_t = a^*]\left\{Y_t^{a^*} - \mu^{a^*}(P_0)\right\}}{\widetilde{w}(a^*)},$$
$$\Psi_t^{a}(P_0) = -\frac{\mathbb{1}[A_t = a]\left\{Y_t^{a} - \mu^{a}(P_0)\right\}}{\widetilde{w}(a)},$$

and $\widetilde{w}(a) = \frac{1}{T}\sum_{t=1}^{T}\mathbb{1}[A_t = a]$. Note that $\widetilde{w}(a)$ is a non-random variable by the definition of the strategy and converges to $w^{\mathrm{GNA}}(a)$ as $T \to \infty$.

By applying the Chernoff bound, for any $v < 0$ and any $\lambda < 0$, we have

$$\begin{aligned}
&\mathbb{P}_{P_0}\left(\sum_{t=1}^{T}\left\{\Psi_t^{a^*}(P_0) + \Psi_t^a(P_0)\right\} \leq v\right) \\
&\leq \mathbb{E}_{P_0}\left[\exp\left(\lambda\sum_{t=1}^{T}\left\{\Psi_t^{a^*}(P_0) + \Psi_t^a(P_0)\right\}\right)\right]\exp\left(-\lambda v\right) \\
&= \mathbb{E}_{P_0}\left[\exp\left(\lambda\sum_{t=1}^{T}\Psi_t^{a^*}(P_0)\right)\right]\mathbb{E}_{P_0}\left[\exp\left(\lambda\sum_{t=1}^{T}\Psi_t^{a}(P_0)\right)\right]\exp\left(-\lambda v\right). \qquad (14)
\end{aligned}$$

First, we consider $\mathbb{E}_{P_0}\left[\exp\left(\lambda\sum_{t=1}^{T}\Psi_t^{a^*}(P_0)\right)\right]$. Because $\Psi_1^{a^*}, \Psi_2^{a^*}, \ldots, \Psi_T^{a^*}$ are i.i.d., we have

$$\mathbb{E}_{P_0}\left[\exp\left(\lambda\sum_{t=1}^{T}\Psi_t^{a^*}(P_0)\right)\right] = \prod_{t=1}^{T}\mathbb{E}_{P_0}\left[\exp\left(\lambda\Psi_t^{a^*}(P_0)\right)\right] = \exp\left(\sum_{t=1}^{T}\log\mathbb{E}_{P_0}\left[\exp\left(\lambda\Psi_t^{a^*}(P_0)\right)\right]\right).$$

By applying the Taylor expansion around $\lambda = 0$, we have

$$\mathbb{E}_{P_0}\left[\exp\left(\lambda\Psi_t^{a^*}(P_0)\right)\right] = 1 + \sum_{k=1}^{\infty}\lambda^k\mathbb{E}_{P_0}\left[(\Psi_t^{a^*}(P_0))^k/k!\right].$$

holds.

Here, for $t \in [T]$ such that $A_t = a^*$,

$$\mathbb{E}_{P_0}\left[\Psi_t^{a^*}(P_0)\right] = \mathbb{E}_{P_0}\left[\frac{1}{\widetilde{w}(a^*)}\left\{Y_t^{a^*} - \mu^{a^*}(P_0)\right\}\right] = 0,$$

and

$$\mathbb{E}_{P_0}\left[\left(\Psi_t^{a^*}(P_0)\right)^2\right] = \mathbb{E}_{P_0}\left[\frac{1}{(\widetilde{w}(a^*))^2}\left\{Y_t^{a^*} - \mu^{a^*}(P_0)\right\}^2\right] = \frac{\left(\sigma^{a^*}\right)^2}{(\widetilde{w}(a^*))^2}$$

hold. Additionally, $\mathbb{E}_{P_0}\left[\left(\Psi_t^{a^*}(P_0)\right)^k\right] = 0$ holds for $k \geq 3$ because Y^{a^*} follows a Gaussian distribution.

For $t \in [T]$ such that $A_t \neq a^*$, $\mathbb{E}_{P_0}\left[\left(\Psi_t^{a^*}(P_0)\right)^k\right] = 0$ hold for $k \geq 1$.

Note that the Taylor expansion of $\log(1+z)$ around $z=0$ is given as $\log(1+z) = z - z^2/2 + z^3/3 - \cdots$. Therefore, we have

$$\begin{aligned}
&\log \mathbb{E}_{P_0}\left[\exp\left(\lambda \Psi_t^{a^*}(P_0)\right)\right] \\
&= \left\{\lambda \mathbb{E}_{P_0}\left[\Psi_t^{a^*}(P_0)\right] + \lambda^2 \mathbb{E}_{P_0}\left[(\Psi_t^{a^*}(P_0))^2/2!\right] + O\left(\lambda^3\right)\right\} - \frac{1}{2}\left\{\lambda \mathbb{E}_{P_0}\left[\Psi_t^{a^*}(P_0)\right]\right\}^2 \\
&= \lambda^2 \frac{\left(\sigma^{a^*}(P_0)\right)^2}{(\widetilde{w}(a^*))^2}.
\end{aligned}$$

Thus,

$$\sum_{t=1}^{T} \log \mathbb{E}_{P_0}\left[\exp\left(\lambda \Psi_t^{a^*}(P_0)\right)\right] = T\frac{\lambda^2}{2}\frac{\left(\sigma^{a^*}\right)^2}{\widetilde{w}(a^*)}$$

holds.

Similarly,

$$\sum_{t=1}^{T} \log \mathbb{E}_{P_0}\left[\exp\left(\lambda \Psi_t^{a}(P_0)\right)\right] = T\frac{\lambda^2}{2}\frac{\left(\sigma^{a}\right)^2}{\widetilde{w}(a)}$$

holds.

Therefore, from equation 14, we have

$$\mathbb{P}_{P_0}\left(\sum_{t=1}^{T}\left\{\Psi_t^{a^*}(P_0) + \Psi_t^{a}(P_0)\right\} \leq v\right) \leq \exp\left(T\frac{\lambda^2}{2}\frac{\left(\sigma^{a^*}\right)^2}{\widetilde{w}(a^*)} + T\frac{\lambda^2}{2}\frac{\left(\sigma^{a}\right)^2}{\widetilde{w}(a)} - \lambda v\right).$$

Let $v = T\lambda\Omega^{a^*,a}(w^{\mathrm{GNA}})$ and $\lambda = -\frac{\Delta^a(P_0)}{\Omega^{a^*,a}(\widetilde{w})}$. Then, we have

$$\mathbb{P}_{P_0}\left(\sum_{t=1}^{T}\left\{\Psi_t^{a^*}(P_0) + \Psi_t^{a}(P_0)\right\} \leq -T\Delta^a(P_0)\right) \leq \exp\left(-\frac{T\left(\Delta^a(P_0)\right)^2}{2\Omega^{a^*,a}(\widetilde{w})}\right).$$

Finally, we have

$$-\frac{1}{T}\log \mathbb{P}_{P_0}\left(\widehat{\mu}_T^{a^*} \leq \widehat{\mu}_T^{a}\right) \geq \frac{(\Delta^a(P_0))^2}{2\Omega^{a^*,a}(\widetilde{w})}.$$

Because $\widetilde{w} \to w^{\mathrm{GNA}}(a)$ as $T \to \infty$, we have

$$\begin{aligned}
&\liminf_{T\to\infty} -\frac{1}{T}\log \mathbb{P}_{P_0}(\widehat{a}_T^{\mathrm{EBA}} \neq a^*) \geq \liminf_{T\to\infty} -\frac{1}{T}\log \sum_{a\neq a^*} \mathbb{P}_{P_0}(\widehat{\mu}_T^{a^*} \leq \widehat{\mu}_T^{a}) \\
&\geq \liminf_{T\to\infty} -\frac{1}{T}\log\left\{(K-1)\max_{a\neq a^*}\mathbb{P}_{P_0}(\widehat{\mu}_T^{a^*} \leq \widehat{\mu}_T^{a})\right\} \geq \min_{a\neq a^*}\frac{(\Delta^a(P_0))^2}{2\Omega^{a^*,a}(w^{\mathrm{GNA}})} \geq \min_{a\neq a^*}\frac{\underline{\Delta}^2}{2\Omega^{a^*,a}(w^{\mathrm{GNA}})}.
\end{aligned}$$

Thus, the proof of Theorem 4.1 is complete. □