Locally Optimal Fixed-Budget Best Arm Identification in Two-Armed Gaussian Bandits with Unknown Variances

Anonymous authors
Paper under double-blind review

Abstract

We address the problem of best arm identification (BAI) with a fixed budget for two-armed Gaussian bandits. In BAI, given multiple arms, we aim to find the best arm, an arm with the highest expected reward, through an adaptive experiment. Kaufmann et al. (2016) develops a lower bound for the probability of misidentifying the best arm. They also propose a strategy, assuming that the variances of rewards are known, and show that it is asymptotically optimal in the sense that its probability of misidentification matches the lower bound as the budget approaches infinity. However, an asymptotically optimal strategy is unknown when the variances are unknown. For this open issue, we propose a strategy that estimates variances during an adaptive experiment and draws arms with a ratio of the estimated standard deviations. We refer to this strategy as the *Neyman Allocation (NA)-Augmented Inverse Probability weighting (AIPW)* strategy. We then demonstrate that this strategy is asymptotically optimal by showing that its probability of misidentification matches the lower bound when the budget approaches infinity, and the gap between the expected rewards of two arms approaches zero (*small-gap regime*). Our results suggest that under the worst-case scenario characterized by the small-gap regime, our strategy, which employs estimated variance, is asymptotically optimal even when the variances are unknown.

1 Introduction

This study investigates the problem of *best arm identification (BAI) with a fixed budget* in stochastic two-armed Gaussian bandits. In this problem, we consider an adaptive experiment with a fixed number of rounds, called a *budget*. At each round, we can draw an arm and observe the reward. The goal of the problem is to identify the best arm with the highest expected reward at the end of the experiment (Bubeck et al., 2009; Audibert et al., 2010).

Formally, we consider the following adaptive experiment with two arms and Gaussian rewards. There are two arms 1 and 2, and an arm $a \in \{1, 2\}$ has an $\mathbb{R}$-valued Gaussian reward $Y_a \sim \mathcal{N}(\mu_a, \sigma_a^2)$ with the mean $\mu_a \in \mathbb{R}$ and the variance $\sigma_a^2 > 0$. Let $P = P(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2) = (\mathcal{N}(\mu_1, \sigma_1^2), \mathcal{N}(\mu_2, \sigma_2^2))$ be a pair of distributions that generate (Y_1, Y_2), where $\mathcal{N}(\mu, \sigma^2)$ is a Gaussian distribution with a mean μ and a variance σ^2. Let $\mathcal{P}^{\mathrm{G}} = \{P(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2) : \mu_1 \neq \mu_2\}$ be a set of the distributions, which is referred to as the *Gaussian bandit models*. For an instance P, the best arm $a^\star(P) \in \{1, 2\}$ is defined as $a^\star(P) = \arg\max_{a \in \{1,2\}} \mu_a$, which is assumed to exist uniquely.

In the adaptive experiment, we consider a strategy to identify the best arm. A fixed budget T is given. For each round $t \in [T] := \{1, 2, \ldots, T\}$, let $(Y_{1,t}, Y_{2,t})$ be an independent and identically distributed (i.i.d.) copy of (Y_1, Y_2). At each round t, we draw arm $A_t \in \{1, 2\}$ and observe a reward $Y_t = \sum_{a \in \{1,2\}} \mathbb{1}[A_t = a] Y_{a,t}$. At the end of the experiment (after round T), we recommend an estimated best arm $\widehat{a}_T \in \{1, 2\}$. During an experiment, we follow a *strategy* that determines which arm to draw and which arm to recommend as the best arm. The performance of strategies is evaluated by a minimal probability of misidentification $\mathbb{P}_P(\widehat{a}_T \neq a^\star(P))$, where $\mathbb{P}_P$ is the probability law under P.

Background. In fixed-budget BAI, it has been an important question of interest to investigate the probability of misidentification $\mathbb{P}_P(\widehat{a}_T \neq a^\star(P))$ in the limit $T \to \infty$. For the interest, a typical approach is to derive an upper and lower bound of the probability separately and specify its value.

For a lower bound of the probability of misidentification, Kaufmann et al. (2016) develops a general theory for deriving lower bounds of the probability. Their theory applies the change-of-measure argument, which has been employed in various problems (van der Vaart, 1998), including studies for regret minimization (Lai & Robbins, 1985). Their lower bound is general and can be applied to a wide range of settings, such as the fixed confidence setting (Garivier & Kaufmann, 2016) as well as the fixed budget setting.

In contrast, an upper bound of the misidentification probability has not been fully clarified. A typical way to derive upper bounds is to construct a specific strategy and evaluate its misidentification probability. Kaufmann et al. (2016) develops a strategy under a setting in which the variance (σ_1^2, σ_2^2) of the reward is known and shows its misidentification probability corresponds to the lower bound. However, this strategy is not available under the usual setting with unknown variance. Based on these situations, the current results are insufficient to establish an upper bound for the misidentification probability when the variances are unknown.

Based on the situation above, our interest is in strategies for identifying misidentification probabilities in the adaptive experimental setting described above. Specifically, we need a strategy such that an upper bound on its misidentification probability is aligns with the lower bound proposed in Kaufmann et al. (2016). Further, this strategy must be valid when the variance is unknown.

Our approach and contribution. In this study, we develop a strategy whose probability of misidentification aligns with the lower bound under an additional setting. To accomplish this, we develop the *Neyman allocation-augmented inverse probability weight* (NA-AIPW) strategy. Then, we show that the probability of misidentification aligns with the lower bound under a *small-gap regime*. The details of each are described below.

The NA-AIPW strategy consists of a sampling rule using the Neyman allocation (NA) and a recommendation rule using the augmented inverse probability weighting (AIPW) estimator. NA is a method of sampling arms using a ratio of the root of the variance of rewards, as utilized in Neyman (1934); Kaufmann et al. (2016). The NA-AIPW strategy samples the arms by estimating this variance during the adaptive experiment. At the end of the experiment, the NA-AIPW strategy recommends an arm with the highest expected reward estimated by using the AIPW estimator, which is an unbiased estimator with a small asymptotic variance.

The small-gap regime considers a situation $\mu_1 - \mu_2 \to 0$ as $T \to \infty$. Although this additional setting slightly simplifies the problem with BAI, the problem is still sufficiently complicated since the small gap makes it difficult to identify the best arm. This setting has been utilized in BAI with fixed confidence, such as the analysis of lil'UCB (Jamieson et al., 2014). In statistical test, such an evaluation framework is known as the local Bahadur efficiency (Bahadur, 1960; Wieand, 1976; Akritas & Kourouklis, 1988; He & Shao, 1996). . From a technical perspective, the small-gap regime is a situation where we can ignore the estimation error of the variances compared to the difficulty of identifying the best arm. Since the error of the estimation of the variance is relatively negligible in the small-gap setting, we can show that the misidentification probability of the NA-AIPW strategy matches the lower bound.

We summarize the backgrounds and our contributions. In BAI with two-armed Gaussian rewards and a fixed budget, a strategy has been needed in which its misidentification probability achieves the lower bound derived by Kaufmann et al. (2016). Although Kaufmann et al. (2016) demonstrates an asymptotically optimal strategy that satisfies the requirement with known variances, it remains an unresolved issue to find a strategy whose upper bound matches their derived lower bound when variances are unknown. For this issue, this study proposes the NA-AIPW strategy whose probability of misidentification matches the lower bound under the small-gap regime.

We note some relation to the existing studies on the BAI problem. Because we restrict our analysis to the small-gap setting, our result does not contradict the analysis on another lower bound of the misidentification probability by Ariu et al. (2021) and Degenne (2023). We discuss this problem in Section 4.

Organization. This study is organized as follows. First, in Section 2, we review the lower bound of Kaufmann et al. (2016). Then, in Section 3, we propose our NA-AIPW strategy. In Section 4, we show that the misidentification probability of the strategy asymptotically corresponds to the lower bound by Kaufmann et al. (2016) under the small-gap setting. We show the proof in Section 5, where we also provide a novel concentration inequality based on the Chernoff bound. In Section 6, we discuss related work and remaining problems, which includes an extension of our small-gap setting to a setting with multi-armed bandits and non-Gaussian rewards.

Notation. Let $\mathcal{F}_t$ be the sigma-algebra generated by all observations up to round t. We define a truncation operator: for a variable $v \in \mathbb{R}$ and a constant $c \geq 1$, $\mathrm{thre}(v; c) := \max\{\min\{v, c\}, 1/c\}$.

2 Lower Bound of Probability of Misidentification

As a preparation, we introduce a lower bound for the probability of misidentification in BAI with a fixed budget. We call a strategy is *consistent*, if for any $P \in \mathcal{P}^{\mathrm{G}}$, $\mathbb{P}_P(\widehat{a}_T \neq a^\star(P)) \to 0$ as $T \to \infty$. To evaluate the performance of strategies for each $P \in \mathcal{P}^{\mathrm{G}}$, we focus on the following metric for $\mathbb{P}_P(\widehat{a}_T \neq a^\star(P))$ used in many studies, such as Kaufmann et al. (2016):

$$-\frac{1}{T}\log \mathbb{P}_P(\widehat{a}_T \neq a^\star(P)).$$

Note that the upper bound (resp. lower bound) of this term works as a lower bound (resp. upper bound) of the probability of misidentification $\mathbb{P}_P(\widehat{a}_T \neq a^\star(P))$ since $x \mapsto -\log x$ is a strictly decreasing function.

For two-armed Gaussian bandits, Kaufmann et al. (2016) presents the following lower bounds. Let $P^* \in \mathcal{P}^{\mathrm{G}}$ be a distribution that generates the rewards $\{(Y_{1,t}, Y_{2,t})\}_{t\in[T]}$.

Proposition 2.1 (Theorem 12 in Kaufmann et al. (2016))**.** *For each* $P^* = (\mathcal{N}(\mu_1^*, \sigma_1^2), \mathcal{N}(\mu_2^*, \sigma_2^2) \in \mathcal{P}^{\mathrm{G}}$ *and* $\Delta = \mu_1^* - \mu_2^*$, *any consistent strategy satisfies*

$$\limsup_{T\to\infty} -\frac{1}{T}\log \mathbb{P}_{P^*}(\widehat{a}_T \neq a^\star(P^*)) \leq \frac{\Delta^2}{2\left(\sqrt{\sigma_1^2} + \sqrt{\sigma_2^2}\right)^2}.$$

From the statement, there are some important aspects of this lower bound: (i) The term $\Delta = \mu_1^* - \mu_2^*$, which referred to a *gap*, appears in the numerator and the magnitude of the error is described by the gap. (ii) The variances (σ_1^2, σ_2^2) appear in the denominator, which plays an important role.

It has been discussed to find a strategy in which the upper bound of its probability of misidentification coincides with this lower bound in Proposition 2.1. Although Kaufmann et al. (2016) develops a strategy that satisfies the requirement, it needs to sample arms with some probability depending on the known variances (σ_1^2, σ_2^2). To the best of our knowledge, if the variances are unknown and need to be estimated during adaptive experiments, no one has found the desired strategy. For this open problem, also see Kaufmann (2020), Ariu et al. (2021), Qin (2022), and Degenne (2023).

3 The NA-AIPW Strategy

In this section, we define our strategy. Formally, a strategy gives a pair $((A_t)_{t\in[T]}, \widehat{a}_T)$, where (i) $(A_t)_{t\in[T]} \in \{1,2\}^T$ is a sequence of arms generated by a sampling rule that determines which arm A_t is chosen in each t based on $\mathcal{F}_{t-1}$, and (ii) $\widehat{a}_T \in \{1,2\}$ is a recommended arm by a recommendation rule based on $\mathcal{F}_T$. Our proposed NA-AIPW strategy consists of (i) a sampling rule with the Neyman Allocation (NA) (Neyman, 1923), and (ii) a recommendation rule using the Augmented Inverse Probability Weighting (AIPW) estimator (Robins et al., 1994; Bang & Robins, 2005). Based on these rules, we refer to this strategy as the NA-AIPW strategy[1].

---

[1] Similar strategies are often used in the context of the average treatment effect estimation by an adaptive experiment (van der Laan, 2008; Kato et al., 2020).

3.1 Target Allocation Ratio

As preparation, we introduce the notion of a target allocation ratio, which will be used for the sampling rule. We define target allocation ratios $w_1^*, w_2^* \in (0, 1)$ as

$$w_1^* = \frac{\sigma_1}{\sigma_1^2 + \sigma_2}, \quad \text{and} \quad w_2^* = 1 - w_1^*.$$

A sampling rule following this target allocation ratio is known as the Neyman allocation rule (Neyman, 1934). Glynn & Juneja (2004) and Kaufmann et al. (2016) also propose this allocation. This target allocation ratio is characterized by the variances (standard deviations); therefore, the target allocation ratio is unknown when the variances are unknown. Therefore, to use this ratio, we need to estimate it from observations.

3.2 Sampling Rule with Neyman Allocation (NA)

We present the sampling rule with the NA. At each round $t \in [T]$, our sampling rule randomly draws an arm $a \in \{1, 2\}$ with a probability identical to an estimated version of the target allocation ratio w_a^*. To estimate the target allocation ratio w_a^*, we estimate the variances during the adaptive experiment. For $a \in \{1, 2\}$, let $\{\widehat{\sigma}_a\}_{t\in[T]}$ and $\{\widehat{w}_{a,t}\}$ be sequences of estimators of σ_a, μ_a and w_a^*, that will be defined bellow.

We use the rounds $t = 1$ and $t = 2$ for initialization. Specifically, we draw the arm 1 at round $t = 1$ and the arm 2 at round $t = 2$, and also set $\widehat{w}_{a,1} = \widehat{w}_{a,2} = 1/2$ for $a \in \{1, 2\}$.

At the round $t \geq 3$, we estimate the target allocation ratio (variances) w_a^* for $a \in \{1, 2\}$ using past observations $\mathcal{F}_{t-1}$. For each $t \geq 3$, we first define an estimator of the expected reward μ_a as

$$\widetilde{\mu}_{a,t} = \frac{1}{\sum_{s=1}^{t-1} \mathbb{1}[A_s = a]} \sum_{s=1}^{t-1} \mathbb{1}[A_s = a] Y_{a,s}.$$

Also, we define a second moment estimator $\widetilde{\zeta}_{a,t} = (\sum_{s=1}^{t-1} \mathbb{1}[A_s = a])^{-1} \sum_{s=1}^{t-1} \mathbb{1}[A_s = a] Y_{a,s}^2$, and a root of variance estimator $\widetilde{\sigma}_{a,t} = \{\widetilde{\zeta}_{a,t} - (\widetilde{\mu}_{a,t})^2\}^{1/2}$. Then, we define the estinator $\widehat{\sigma}_{a,t} = \mathrm{thre}(\widetilde{\sigma}_{a,t}; C_{\sigma^2}^{1/2})$ with some predetermined constant $C_{\sigma^2} > 0$. Also, we define the estimator $\widehat{w}_{1,t}$ and $\widehat{w}_{2,t}$ as

$$\widehat{w}_{1,t} = \frac{\widehat{\sigma}_{1,t}}{\widehat{\sigma}_{1,t} + \widehat{\sigma}_{2,t}}, \quad \text{and} \quad \widehat{w}_{2,t} = 1 - \widehat{w}_{1,t}. \tag{1}$$

At the end of the round $t \geq 3$, we sample the arm A_t; we generate γ_t from the uniform distribution on $[0, 1]$ and set

$$A_t = \begin{cases} 1 & \text{if } \gamma_t \leq \widehat{w}_{1,t} \\ 2 & \text{otherwise.} \end{cases}$$

We note that it is possible to increase the number of rounds for initialization, while we present that only the first two rounds are used for initialization. The more rounds for initialization have the role of stabilizing the sampling rule in practice. This idea is similar to the forced-sampling (Garivier & Kaufmann, 2016).

3.3 Recommendation Rule with the Augmented Inverse Probability Weighting (AIPW) Estimator

We present our recommendation rule. In the recommendation phase after round T, we estimate μ_a^* for each $a \in \{1, 2\}$ and recommend an arm with the bigger estimated expected reward. With a truncated version of the estimated expected reward $\widehat{\mu}_{a,t} = \mathrm{thre}(\widetilde{\mu}_{a,t}, C_\mu)$ with some predetermined constant $C_\mu > 0$, we define the *augmented inverse probability weighting* (AIPW) estimator of μ_a^* for each $a \in \{1, 2\}$ as

$$\widehat{\mu}_{a,T}^{\mathrm{AIPW}} = \frac{1}{T} \sum_{t=1}^{T} \psi_{a,t}, \qquad \text{where } \psi_{a,t} = \frac{\mathbb{1}[A_t = a]\big(Y_{a,t} - \widehat{\mu}_{a,t}\big)}{\widehat{w}_{a,t}} + \widehat{\mu}_{a,t}. \tag{2}$$

Algorithm 1 NA-AIPW Strategy

```
**Parameter:** Positive constants $C_\mu$ and $C_{\sigma^2}$.
**Initialization:**
At $t = 1$, sample $A_t = 1$; at $t = 2$, sample $A_t = 2$. For $a \in \{1, 2\}$, set $\widehat{w}_{a,1} = \widehat{w}_{a,2} = 0.5$.
for $t = 3$ to $T$ do
    Construct $\widehat{w}_{a,t}$ following equation 1.
    Sample $\gamma_t$ from the uniform distribution on $[0, 1]$.
    $A_t = 1$ if $\gamma_t \leq \widehat{w}_{1,t}$; $A_t = 2$ if $\gamma_t > \widehat{w}_{1,t}$.
    Observe $Y_t$.
end for
Construct $\widehat{\mu}^{\mathrm{AIPW}}_{a,T}$ for $a \in \{1, 2\}$. following equation 2.
Recommend $\widehat{a}_T$ following equation 3.
```

At the end of the experiment (after the round $t = T$), we recommend $\widehat{a}_T$ as

$$\widehat{a}_T = \begin{cases} 1 & \text{if} \quad \widehat{\mu}^{\mathrm{AIPW}}_{1,T} \geq \widehat{\mu}^{\mathrm{AIPW}}_{2,T}, \\ 2 & \text{otherwise.} \end{cases} \tag{3}$$

We adopt the AIPW estimator for our strategy because it has several advantages. First, the AIPW estimator has the property of semiparametric efficiency, which indicates that it has the smallest asymptotic variance among a certain class (Hahn, 1998). The property is necessary to prove that the strategy using the AIPW estimator is optimal, which means the misidentification probability is small enough to achieve its lower bound. The second reason is more technical; the AIPW estimator simplifies a proof for theoretical analysis. Specifically, we can decomposed an error by the AIPW estimator into a sum of random variables with martingale properties, making it suitable for analysis using the central limit theorem. This property is unique to the AIPW estimator, but not to naive estimators such as an empirical average. Details will be given in Section 5.

We provide the pseudo-code for our proposed strategy in Algorithm 1. Note that we introduce C_μ and C_{σ^2} for technical purposes to bound the estimators and any large positive value can be used.

4 Misidentification Probability and Asymptotic Optimality

In this section, we show the following upper bound of the misspecification probability of the NA-AIPW strategy, which also implies that the strategy is asymptotically optimal.

Theorem 4.1 (Upper bound of the NA-AIPW strategy)**.** *For each $P^* \in \mathcal{P}^{\mathrm{G}}$, the following holds as $\Delta \to 0$:*

$$\liminf_{T\to\infty} -\frac{1}{T} \log \mathbb{P}_{P^*}\left(\widehat{a}^{\mathrm{AIPW}}_T \neq a^\star(P^*)\right) \geq \frac{\Delta^2}{2(\sigma_1 + \sigma_2)^2} + O\left(\Delta^3\right).$$

Note that the lower bound of $-\log \mathbb{P}_{P^*}\left(\widehat{a}^{\mathrm{AIPW}}_T \neq a^\star(P^*)\right)$ implies the upper bound of $\mathbb{P}_{P^*}\left(\widehat{a}^{\mathrm{AIPW}}_T \neq a^\star(P^*)\right)$. This theorem implies us to evaluate the probability of misidentification up to the constant term, even when it is exponentially small, as $\Delta \to 0$.

This result directly implies the asymptotic optimality of the NA-AIPW strategy. As $\Delta \to 0$, the upper bound matches the lower bound in Proposition 2.1. This asymptotic optimality result suggests that the estimation error of the target allocation ratio (variances) w^* is negligible when $\Delta \to 0$. This is because the estimation error is insignificant compared to the challenges of identifying the best arm due to the small gap.

Although studies, such as Ariu et al. (2021), Qin (2022), and Degenne (2023), point out the non-existence of the optimal strategies in fixed-budget BAI against the lower bound shown by Kaufmann et al. (2016), our result does not yield a contradiction. Existing impossibility results discuss the existence of a strategy that violates the lower bound. Note that the lower bounds in Kaufmann et al. (2016) are applicable to any instances in the bandit models (with some regularity conditions). In other words, if we consider the lower bound in Kaufmann et al. (2016) for all instances, there exists an instance under which there exists a strategy whose lower bound is larger than the lower bound derived by Kaufmann et al. (2016). In contrast, we only consider bandit models where $\Delta \to 0$. Our result implies that if we restrict bandit models, the upper bounds of our strategy within the restricted bandit models match the lower bound. Because our optimality is limited to a case where $\Delta \to 0$, we refer to our optimality as asymptotic optimality under the small-gap regime or local asymptotic optimality.

We conjecture that even if we replace the AIPW estimator with the sample average estimator, defined as $\widetilde{\mu}_{a,t} = (\sum_{s=1}^{t-1} \mathbb{1}[A_s = a])^{-1} \sum_{s=1}^{t-1} \mathbb{1}[A_s = a]Y_{a,s}$ in Section 3.2, the upper bound of the strategy still matches the lower bound. However, the proof is an open issue. Hirano et al. (2003) and Hahn et al. (2011) show that the sample average estimator $\widetilde{\mu}_{a,t}$ and the AIPW estimator have the same asymptotic variance (or asymptotic distribution). To show the result, we need to employ empirical process arguments. One of the problems in extending the result to analysis for BAI is that their result focuses on the asymptotic distribution, not the tail probability. Therefore, to show the asymptotic optimality of the strategy with the sample average in the sense of the probability of misidentification, we need to modify the result in Hirano et al. (2003) and Hahn et al. (2011) to analyze the tail probability.

5 Proof of Theorem 4.1

To show Theorem 4.1, we derive the upper bound of $\mathbb{P}_{P^*}(\widehat{\mu}^{\mathrm{AIPW}}_{a^\star(P^*),T} \leq \widehat{\mu}^{\mathrm{AIPW}}_{b,T})$ for $b \in \{1, 2\}\backslash\{a^\star(P^*)\}$, which is equivalent to $\mathbb{P}_{P^*}\big(\widehat{a}^{\mathrm{AIPW}}_T \neq a^\star(P^*)\big)$. Without loss of generality, we assume that $a^\star(P^*) = 1$ and $b = 2$. Let us define $V = \frac{\sigma_1^2}{w_1^*} + \frac{\sigma_2^2}{w_2^*} = (\sigma_1 + \sigma_2)^2$ and

$$\Psi_t = \frac{\psi_{1,t} - \psi_{2,t} - \Delta}{\sqrt{V}}.$$

Therefore, in the following parts, we aim to derive the upper bound of $\mathbb{P}_{P^*}\left(\widehat{\mu}^{\mathrm{AIPW}}_{a^\star(P^*),T} \leq \widehat{\mu}^{\mathrm{AIPW}}_{b,T}\right) = \mathbb{P}_{P^*}\left(\widehat{\mu}^{\mathrm{AIPW}}_{1,T} \leq \widehat{\mu}^{\mathrm{AIPW}}_{2,T}\right) = \mathbb{P}_{P^*}\left(\sum_{t=1}^{T} \Psi_t \leq -\frac{T\Delta}{\sqrt{V}}\right)$. Let $\mathbb{E}_P$ be the expectation under $P \in \mathcal{P}^{\mathrm{G}}$. We derive the upper bound using the Chernoff bound. This proof is partially inspired by techniques in Hadad et al. (2021), and Kato et al. (2020).

First, because there exists a constant $C > 0$ independent of T such that $\widehat{w}_{a,t} > C$ by construction, the following lemma holds.

Lemma 5.1. *For any $P^* \in \mathcal{P}^{\mathrm{G}}$ and all $a \in \{1, 2\}$, $\widehat{\mu}_{a,t} \xrightarrow{\mathrm{a.s}} \mu^*_a$ and $\widehat{\sigma}^2_a \xrightarrow{\mathrm{a.s}} \sigma^2_a$.*

Furthermore, from $\widehat{\sigma}^2_a \xrightarrow{\mathrm{a.s}} \sigma^2_a$ and continuous mapping theorem, for any $P^* \in \mathcal{P}^{\mathrm{G}}$ and all $a \in \{1, 2\}$, $\widehat{w}_{a,t} \xrightarrow{\mathrm{a.s}} w^*_{a,t}$.

Step 1: Sequence $\{\Psi_t\}_{t=1}^T$ is an MDS

We prove that $\{\Psi_t\}_{t=1}^T$ is an MDS; that is, $\mathbb{E}_{P^*}[\Psi_t|\mathcal{F}_{t-1}] = 0$. Although this fact is well-known in the literature of causal inference (van der Laan, 2008; Hadad et al., 2021; Kato et al., 2020), we show the proof for the sake of completeness.

Lemma 5.2. *For any $P^* \in \mathcal{P}^{\mathrm{G}}$, $\{\Psi_t\}_{t=1}^T$ is an MDS.*

Proof. For each $t \in [T]$, it holds that

$$\begin{aligned}
&\sqrt{V}\mathbb{E}_{P^*}[\Psi_t|\mathcal{F}_{t-1}] \\
&= \mathbb{E}_{P^*}\left[\frac{\mathbb{1}[A_t = 1]\big(Y_{1,t} - \widehat{\mu}_{1,t}\big)}{\widehat{w}_{1,t}} + \widehat{\mu}_{1,t}|\mathcal{F}_{t-1}\right] - \mathbb{E}_{P^*}\left[\frac{\mathbb{1}[A_t = 2]\big(Y_{2,t} - \widehat{\mu}_{2,t}\big)}{\widehat{w}_{2,t}} + \widehat{\mu}_{2,t}|\mathcal{F}_{t-1}\right] - \Delta
\end{aligned}$$

$$= \frac{\widehat{w}_{1,t}\big(\mu^*_1 - \widehat{\mu}_{1,t}\big)}{\widehat{w}_{1,t}} + \widehat{\mu}_{1,t} - \frac{\widehat{w}_{2,t}\big(\mu^*_2 - \widehat{\mu}_{2,t}\big)}{\widehat{w}_{2,t}} + \widehat{\mu}_{2,t} - \Delta = \{(\mu^*_1 - \mu^*_2) - (\mu^*_1 - \mu^*_2)\} = 0$$

□

Step 2: Evaluation by using the Chernoff Bound with Martingales

By applying the Chernoff bound, for any $v < 0$ and any $\lambda < 0$, it holds that

$$\mathbb{P}_{P^*}\left(\frac{1}{T}\sum^T_{t=1}\Psi_t \leq v\right) \leq \mathbb{E}_{P^*}\left[\exp\left(\lambda\sum^T_{t=1}\Psi_t\right)\right]\exp\left(-T\lambda v\right).$$

From the Chernoff bound and a property of an MDS, we have

$$\mathbb{E}_{P^*}\left[\exp\left(\lambda\sum^T_{t=1}\Psi_t\right)\right] = \mathbb{E}_{P^*}\left[\prod^T_{t=1}\mathbb{E}_{P^*}\left[\exp\left(\lambda\Psi_t\right)|\mathcal{F}_{t-1}\right]\right] = \mathbb{E}_{P^*}\left[\exp\left(\sum^T_{t=1}\log\mathbb{E}_{P^*}\left[\exp\left(\lambda\Psi_t\right)|\mathcal{F}_{t-1}\right]\right)\right]. \tag{4}$$

Then, the Taylor expansion around $\lambda = 0$ yields

$$\log\mathbb{E}_{P^*}\left[\exp\left(\lambda\Psi_t\right)|\mathcal{F}_{t-1}\right] = \frac{\lambda^2}{2}\mathbb{E}_{P^*}\left[\Psi^2_t|\mathcal{F}_{t-1}\right] + O\left(\lambda^3\right). \tag{5}$$

Here, $\mathbb{E}_{P^*}\left[\exp\left(\lambda\Psi_t\right)|\mathcal{F}_{t-1}\right] = 1 + \sum^\infty_{k=1}\lambda^k\mathbb{E}_{P^*}\left[\Psi^k_t/k!|\mathcal{F}_{t-1}\right]$. Because $\mathbb{E}_{P^*}\left[\Psi^k_t/k!|\mathcal{F}_{t-1}\right]$ is bounded by a constant that is independent of T for all $k \geq 1$,

$$\mathbb{E}_{P^*}\left[\exp\left(\lambda\Psi_t\right)|\mathcal{F}_{t-1}\right] = 1 + \sum^2_{k=1}\lambda^k\mathbb{E}_{P^*}\left[\Psi^k_t/k!|\mathcal{F}_{t-1}\right] + O\left(\lambda^3\right).$$

Note that the Taylor series 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

$$\begin{aligned}
&\log\mathbb{E}_{P^*}\left[\exp\left(\lambda\Psi_t\right)|\mathcal{F}_{t-1}\right]\\
&= \left\{\lambda\mathbb{E}_{P^*}\left[\Psi_t|\mathcal{F}_{t-1}\right] + \frac{\lambda^2}{T}\mathbb{E}_{P^*}\left[\Psi^2_t/2!|\mathcal{F}_{t-1}\right] + O\left(\lambda^3\right)\right\} - \frac{1}{2}\left\{\lambda\mathbb{E}_{P^*}\left[\Psi_t|\mathcal{F}_{t-1}\right] + O\left(\lambda^2\right)\right\}^2\\
&= \frac{\lambda^2}{T}\mathbb{E}_{P^*}\left[\Psi^2_t/2!|\mathcal{F}_{t-1}\right] + O\left(\lambda^3\right).
\end{aligned}$$

Here, we used $\mathbb{E}_{P^*}\left[\Psi_t|\mathcal{F}_{t-1}\right] = 0$. Thus, the equation 5 holds.

Step 3: Convergence of the Second Moment

We next show $\mathbb{E}_{P^*}\left[\Psi^2_t|\mathcal{F}_{t-1}\right] - 1 \xrightarrow{\text{a.s}} 0$. This result is a direct consequence of Lemma 5.1.

Lemma 5.3. *For any $P^* \in \mathcal{P}^{\mathrm{G}}$,*

$$\mathbb{E}_{P^*}\left[\Psi^2_t|\mathcal{F}_{t-1}\right] - 1 \xrightarrow{\text{a.s}} 0 \qquad \text{as} \quad t \to \infty.$$

Proof. We have

$$\begin{aligned}
&V\mathbb{E}_{P^*}\left[\Psi^2_t|\mathcal{F}_{t-1}\right] = \mathbb{E}_{P^*}\left[\left(\psi_{1,t} - \psi_{2,t} - \Delta\right)^2\Big|\mathcal{F}_{t-1}\right]\\
&= \mathbb{E}_{P^*}\Bigg[\left(\frac{\mathbb{1}[A_t = 1]\big(Y_{1,t} - \widehat{\mu}_{1,t}\big)}{\widehat{w}_{1,t}} - \frac{\mathbb{1}[A_t = 2]\big(Y_{2,t} - \widehat{\mu}_{2,t}\big)}{\widehat{w}_{2,t}}\right)^2\\
&\qquad + 2\left(\frac{\mathbb{1}[A_t = a^\star(P^*)]\big(Y_{1,t} - \widehat{\mu}_{1,t}\big)}{\widehat{w}_{1,t}} - \frac{\mathbb{1}[A_t = a]\big(Y_{2,t} - \widehat{\mu}_{2,t}\big)}{\widehat{w}_{2,t}}\right) \times \left(\widehat{\mu}_{1,t} - \widehat{\mu}_{2,t} - (\mu^*_1 - \mu^*_2)\right)
\end{aligned}$$

$$
\begin{aligned}
&\qquad\qquad + \left(\widehat{\mu}_{1,t} - \widehat{\mu}_{2,t} - (\mu^*_1 - \mu^*_2)\right)^2 |\mathcal{F}_{t-1}\Bigg] \\
&= \mathbb{E}_{P^*}\Bigg[\frac{\mathbb{1}[A_t = 1]\big(Y_{1,t} - \widehat{\mu}_{1,t}\big)^2}{\big(\widehat{w}_{1,t}\big)^2} + \frac{\mathbb{1}[A_t = 2]\big(Y_{2,t} - \widehat{\mu}_{2,t}\big)^2}{\big(\widehat{w}_{2,t}\big)^2} \\
&\qquad\qquad + 2\left(\frac{\mathbb{1}[A_t = 1]\big(Y_{1,t} - \widehat{\mu}_{1,t}\big)}{\widehat{w}_{1,t}} - \frac{\mathbb{1}[A_t = a]\big(Y_{2,t} - \widehat{\mu}_{2,t}\big)}{\widehat{w}_{2,t}}\right)\left(\widehat{\mu}_{1,t} - \widehat{\mu}_{2,t} - (\mu^*_1 - \mu^*_2)\right) \\
&\qquad\qquad + \left(\big(\widehat{\mu}_{1,t} - \widehat{\mu}_{2,t}\big) - (\mu^*_1 - \mu^*_2)\right)^2 |\mathcal{F}_{t-1}\Bigg] \\
&= \sum_{a\in\{1,2\}} \mathbb{E}_{P^*}\left[\frac{\big(Y_{a,t} - \widehat{\mu}_{a,t}\big)^2}{\big(\widehat{w}_{a,t}\big)^2} |\mathcal{F}_{t-1}\right] - \mathbb{E}_{P^*}\left[\left(\big(\widehat{\mu}_{1,t} - \widehat{\mu}_{2,t}\big) - \big(\mu^*_1 - \mu^*_2\big)\right)^2 |\mathcal{F}_{t-1}\right].
\end{aligned}
$$

Here, for $a \in \{1, 2\}$, the followings hold:

$$
\begin{aligned}
&\mathbb{E}_{P^*}\left[\frac{\mathbb{1}[A_t = a]\big(Y_{a,t} - \widehat{\mu}_{a,t}\big)^2}{\big(\widehat{w}_{a,t}\big)^2} |\mathcal{F}_{t-1}\right] = \mathbb{E}_{P^*}\left[\frac{\big(Y_{a,t} - \widehat{\mu}_{a,t}\big)^2}{\widehat{w}_{a,t}} |\mathcal{F}_{t-1}\right] = \frac{\mathbb{E}_{P^*}[(Y_{a,t})^2] - 2\mu^*_a\widehat{\mu}_{a,t} + (\widehat{\mu}_{a,t})^2}{\widehat{w}_{a,t}} \\
&= \frac{\mathbb{E}_{P^*}[(Y_{a,t})^2] - (\mu^*_a)^2 + (\mu^*_a - \widehat{\mu}_{a,t})^2}{\widehat{w}_{a,t}} = \frac{\sigma^2_a + (\mu^*_a - \widehat{\mu}_{a,t})^2}{\widehat{w}_{a,t}},
\end{aligned}
$$

and

$$
\mathbb{E}_{P^*}\left[\frac{\mathbb{1}[A_t = a]\big(Y_{a,t} - \widehat{\mu}_{2,t}\big)^2}{\big(\widehat{w}_{1,t}\big)^2}\frac{\mathbb{1}[A_t = a]\big(Y_{a,t} - \widehat{\mu}_{2,t}\big)^2}{\big(\widehat{w}_{2,t}\big)^2} |\mathcal{F}_{t-1}\right] = 0,
$$

where we used $\mathbb{E}_{P^*}[(Y_{a,t})^2|x] - (\mu^*_a)^2 = \sigma^2_a$. Therefore, the following holds:

$$
\begin{aligned}
&\mathbb{E}_{P^*}\left[\frac{\big(Y_{1,t} - \widehat{\mu}_{1,t}\big)^2}{\widehat{w}_{1,t}} |\mathcal{F}_{t-1}\right] + \mathbb{E}_{P^*}\left[\frac{\big(Y_{2,t} - \widehat{\mu}_{2,t}\big)^2}{\widehat{w}_{2,t}} |\mathcal{F}_{t-1}\right] - \mathbb{E}_{P^*}\left[\left(\big(\widehat{\mu}_{1,t} - \widehat{\mu}_{2,t}\big) - (\mu^*_1 - \mu^*_2)\right)^2 |\mathcal{F}_{t-1}\right] \\
&= \mathbb{E}_{P^*}\left[\frac{\sigma^2_1 + (\mu^*_1 - \widehat{\mu}_{1,t})^2}{\widehat{w}_{1,t}}\right] + \mathbb{E}_{P^*}\left[\frac{\sigma^2_2 + (\mu^*_2 - \widehat{\mu}_{2,t})^2}{\widehat{w}_{2,t}}\right] - \mathbb{E}_{P^*}\left[\left(\big(\widehat{\mu}_{1,t} - \widehat{\mu}_{2,t}\big) - (\mu^*_1 - \mu^*_2)\right)^2\right].
\end{aligned}
$$

Because $\widehat{\mu}_{a,t} \xrightarrow{\text{a.s.}} \mu^*_a$ and $\widehat{w}_{a,t} \xrightarrow{\text{a.s.}} w^*_a$, we have

$$
\begin{aligned}
&\lim_{t\to\infty}\Bigg|\left(\frac{\sigma^2_1 + (\mu^*_1 - \widehat{\mu}_{1,t})^2}{\widehat{w}_{1,t}}\right) + \left(\frac{\sigma^2_2 + (\mu^*_2 - \widehat{\mu}_{2,t})^2}{\widehat{w}_{2,t}}\right) - \left(\big(\widehat{\mu}_{1,t} - \widehat{\mu}_{2,t}\big) - \big(\mu^*_1 - \mu^*_2\big)\right)^2 \\
&\qquad\qquad\qquad\qquad\qquad\qquad - \left(\frac{\sigma^2_1}{w^*_1} + \frac{\sigma^2_a}{w^*_2} + \left(\big(\mu^*_1 - \mu^*_2\big) - (\mu^*_1 - \mu^*_2)\right)^2\right)\Bigg| \\
&\leq \lim_{t\to\infty}\left|\frac{\sigma^2_1}{\widehat{w}_{1,t}} - \frac{\sigma^2_1}{w^*_1}\right| + \lim_{t\to\infty}\left|\frac{\sigma^2_2}{\widehat{w}_{2,t}} - \frac{\sigma^2_2}{w^*_2}\right| + \lim_{t\to\infty}\frac{(\mu^*_1 - \widehat{\mu}_{1,t})^2}{\widehat{w}_{1,t}} + \lim_{t\to\infty}\frac{(\mu^*_2 - \widehat{\mu}_{2,t})^2}{\widehat{w}_{2,t}} \\
&\qquad\qquad + \lim_{t\to\infty}\left|\left(\big(\widehat{\mu}_{1,t} - \widehat{\mu}_{2,t}\big) - \big(\mu^*_1 - \mu^*_2\big)\right)^2 - \left(\big(\mu^*_1 - \mu^*_2\big) - \big(\mu^*_1 - \mu^*_2\big)\right)^2\right| = 0,
\end{aligned}
$$

with probability 1. Therefore, from Lebesgue's dominated convergence theorem, we obtain

$$
\begin{aligned}
&V\mathbb{E}_{P^*}\left[\Psi^2_t|\mathcal{F}_{t-1}\right] - V \\
&= \mathbb{E}_{P^*}\left[\frac{\sigma^2_1 + (\mu^*_1 - \widehat{\mu}_{1,t})^2}{\widehat{w}_{1,t}} |\mathcal{F}_{t-1}\right] + \mathbb{E}_{P^*}\left[\frac{\sigma^2_a + (\mu^*_2 - \widehat{\mu}_{2,t})^2}{\widehat{w}_{2,t}} |\mathcal{F}_{t-1}\right] \\
&\quad - \mathbb{E}_{P^*}\left[\left(\big(\widehat{\mu}_{1,t} - \widehat{\mu}_{2,t}\big) - \big(\mu^*_1 - \mu^*_2\big)\right)^2 |\mathcal{F}_{t-1}\right] - \mathbb{E}_{P^*}\left[\frac{\sigma^2_1}{w^*_1} + \frac{\sigma^2_2}{w^*_2} + \left(\big(\mu^*_1 - \mu^*_2\big) - \big(\mu^*_1 - \mu^*_2\big)\right)^2 |\mathcal{F}_{t-1}\right]
\end{aligned}
$$

$$\xrightarrow{\text{a.s.}} 0.$$

□

This lemma immediately yields the following lemma.

Lemma 5.4. *For any $P^* \in \mathcal{P}^{\mathrm{G}}$, any $\epsilon > 0$, there exists $t(\epsilon) > 0$ such that for all $T > t(\epsilon)$, we have*

$$\frac{1}{T}\sum_{t=1}^{T}\left|\mathbb{E}_{P^*}\left[\Psi_t^2|\mathcal{F}_{t-1}\right] - 1\right| < \epsilon$$

with probability one.

Our proof refers to the proof of Lemma 10 in Hadad et al. (2021).

Proof. Let u_t be $u_t = \mathbb{E}_{P^*}\left[\Psi_t^2|\mathcal{F}_{t-1}\right] - 1$. Note that $\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{P^*}\left[\Psi_t^2|\mathcal{F}_{t-1}\right] - 1 = \frac{1}{T}\sum_{t=1}^{T} u_t$.

From the proof of Lemma 5.3, we can find that u_t is a bounded random variable. Recall that

$$\begin{aligned}
&V\mathbb{E}_{P^*}\left[\Psi_t^2|\mathcal{F}_{t-1}\right]\\
&= \mathbb{E}_{P^*}\left[\frac{\sigma_1^2 + (\mu_1 - \widehat{\mu}_{1,t})^2}{\widehat{w}_{1,t}}|\mathcal{F}_{t-1}\right] + \mathbb{E}_{P^*}\left[\frac{\sigma_2^2 + (\mu_2 - \widehat{\mu}_{2,t})^2}{\widehat{w}_{2,t}}|\mathcal{F}_{t-1}\right] - \mathbb{E}_{P^*}\left[\left(\left(\widehat{\mu}_{1,t} - \widehat{\mu}_{2,t}\right) - \left(\mu_1 - \mu_2\right)\right)^2|\mathcal{F}_{t-1}\right].
\end{aligned}$$

We assumed that $(\mu_1, \mu_2, \widehat{\mu}_{1,t}, \widehat{\mu}_{2.t}, \widehat{w}_{1,t}, \widehat{w}_{2,t})$ are all bounded random variables. Let C be a constant independent of T such that $|u_t| < C$ for all $t \in \mathbb{N}$.

Almost-sure convergence of u_t to zero as $t \to \infty$ implies that for any $\epsilon' > 0$, there exists $t(\epsilon)$ such that $|u_t| < \epsilon'$ for all $t \geq t(\epsilon')$ with probability one. Let $\mathcal{E}(\epsilon')$ denote the event in which this happens; that is, $\mathcal{E}(\epsilon') = \{|u_t| < \epsilon' \quad \forall\, t \geq t(\epsilon')\}$. Under this event, for $T > t(\epsilon')$, the following holds:

$$\frac{1}{T}\sum_{t=1}^{T}|u_t| \leq \frac{1}{T}\sum_{t=1}^{t(\epsilon')} C + \frac{1}{T}\sum_{t=t(\epsilon')+1}^{T}\epsilon = \frac{1}{T}t(\epsilon')C + \epsilon',$$

where $\frac{1}{T}t(\epsilon')C \to 0$ as $T \to \infty$.

Therefore, for any $\epsilon > 0$, there exists $t(\epsilon)$ such that for all $T > t(\epsilon)$, $\frac{1}{T}\sum_{t=1}^{T}|u_t| < \epsilon$ holds with probability one. □

Step 4: Tail Bound with the Approximated Second Moment

Let $v = \lambda$. Then, we have

$$\mathbb{P}_{P^*}\left(\sum_{t=1}^{T}\Psi_t \leq Tv\right) \leq \mathbb{E}_{P^*}\left[\exp\left(-\frac{T\lambda^2}{2} + \frac{\lambda^2}{2}\left\{\sum_{t=1}^{T}\mathbb{E}_{P^*}\left[\Psi_t^2|\mathcal{F}_{t-1}\right] - 1\right\} + TO\left(\lambda^3\right)\right)\right].$$

From Lemma 5.4, for any $\epsilon > 0$, there exists $t(\epsilon) > 0$ such that for all $T > t(\epsilon)$, we have

$$\begin{aligned}
&\mathbb{E}_{P^*}\left[\exp\left(-\frac{T\lambda^2}{2} + \frac{\lambda^2}{2}\left\{\sum_{t=1}^{T}\mathbb{E}_{P^*}\left[\Psi_t^2|\mathcal{F}_{t-1}\right] - 1\right\} + TO\left(\lambda^3\right)\right)\right]\\
&= \exp\left(-\frac{T\lambda^2}{2} + TO\left(\lambda^3\right)\right)\mathbb{E}_{P^*}\left[\exp\left(\frac{\lambda^2}{2}\left\{\sum_{t=1}^{T}\mathbb{E}_{P^*}\left[\Psi_t^2|\mathcal{F}_{t-1}\right] - 1\right\}\right)\right]\\
&\leq \exp\left(-\frac{T\lambda^2}{2} + TO\left(\lambda^3\right)\right)\exp\left(\frac{\lambda^2}{2}T\epsilon\right) = \exp\left(-\frac{T\lambda^2}{2} + TO\left(\lambda^3\right) + \frac{\lambda^2}{2}T\epsilon\right).
\end{aligned}$$

Step 5: Final Step of the Proof of Theorem 4.1

For any $\epsilon > 0$, there exists $t(\epsilon) > 0$ such that for all $T > t(\epsilon)$, we obtain

$$-\frac{1}{T}\log\mathbb{P}_{P^*}\left(\widehat{\mu}^{\mathrm{AIPW}}_{1,T} \leq \widehat{\mu}^{\mathrm{AIPW}}_{2,T}\right) \geq -\left\{-\frac{\lambda^2}{2} + O\left(\frac{\lambda^3}{\sqrt{T}}\right) + \frac{\lambda^2}{2}\epsilon\right\} = \frac{\lambda^2}{2} - O\left(\lambda^3\right) - \frac{\lambda^2}{2}\epsilon,$$

as $\lambda \to 0$.

Let $\lambda = -\frac{\Delta}{\sqrt{V}}$. Then, we have

$$-\frac{1}{T}\log\mathbb{P}_{P^*}\left(\widehat{\mu}^{\mathrm{AIPW},a^*(P)}_{T} \leq \widehat{\mu}^{\mathrm{AIPW},a}_{T}\right) \geq \frac{\Delta^2}{2V} - O\left(\frac{\Delta^3}{V^{3/2}}\right) - \frac{\epsilon\Delta^2}{2V},$$

as $\Delta \to 0$. Letting $\Delta \to 0$ and $T \to \infty$, and then letting $\epsilon \to 0$ independently of T and Δ, then $T \to \infty$, we have

$$-\frac{1}{T}\log\mathbb{P}_{P^*}\left(\widehat{\mu}^{\mathrm{AIPW}}_{1,T} \leq \widehat{\mu}^{\mathrm{AIPW}}_{2,T}\right) \geq \frac{\Delta^2}{2V} - o\left(\Delta^2\right).$$

Thus, the proof is complete.

6 Discussion and Related Work

This section presents discussions and related works.

6.1 On the Asymptotic Optimality in Fixed-Budget BAI

There is a long debate on the optimal strategies for fixed-budget BAI. Glynn & Juneja (2004) develops their strategies by using the large deviation principles. However, while they justify their strategies using the large deviation principles, they do not provide lower bounds for strategies. Therefore, there remains a question about whether their strategies are truly asymptotically optimal.

Kaufmann et al. (2016) establishes distribution-dependent lower bounds for BAI with fixed confidence and budget, utilizing change-of-measure arguments. According to their results, we can confirm that for two-armed Gaussian bandits, the strategy of Glynn & Juneja (2004) is optimal.

However, Kaufmann et al. (2016) leaves lower bounds for multi-armed fixed-budget BAI as an open issue. Based on the arguments of Glynn & Juneja (2004) and Russo (2020), Kasy & Sautmann (2021) attempts to derive an asymptotically optimal strategy, but their attempt does not succeed. As pointed out by Ariu et al. (2021), without additional assumptions, there exists an instance P^* whose lower bound is larger than that of Kaufmann et al. (2016). This result is based on another lower bound discovered by Carpentier & Locatelli (2016). These arguments are summarized by Qin (2022).

To address this issue, Kato et al. (2023b) and Degenne (2023) consider a restriction such that sampling rules do not depend on P^*. Under this restriction, we can show the asymptotic optimality of the strategy provided by Glynn & Juneja (2004), which requires full knowledge about P^* and is practically infeasible.

Komiyama et al. (2022) and Atsidakou et al. (2023) discuss asymptotically optimal strategies from minimax and Bayesian perspectives, respectively, where the leading factor ignoring some constant terms of lower and upper bounds match, unlike our optimality up to constant terms. This open issue is further explored by Komiyama et al. (2022), Wang et al. (2023a), Wang et al. (2023b), and Kato (2023).

Note that in the fixed confidence BAI setting, Garivier & Kaufmann (2016) proposes a strategy with an upper bound matching the derived lower bound. However, in the fixed-budget BAI, it remains unclear whether a strategy with an upper bound matching Kaufmann et al. (2016)'s lower bound exists.

Alternative lower bounds have been proposed by Audibert et al. (2010), Bubeck et al. (2011), Komiyama et al. (2023) and Kato et al. (2023a) for the expected simple regret minimization, which is another performance measure different from the probability of misidentification.

Some research employs local asymptotics to examine the asymptotic optimality of the Neyman allocation rule in this context, such as Armstrong (2022) and Adusumilli (2022).

Ordinal optimization in the operations research community is another related field (Ahn et al., 2021; Chen et al., 2000).

6.2 Neyman Allocation with Unknown Variances

For two-armed Gaussian bandits with known variances, Chen et al. (2000), Glynn & Juneja (2004), and Kaufmann et al. (2016) conclude that sampling each arm with a proportion of the standard deviation is optimal, which corresponds to the Neyman allocation Neyman (1934).

The Neyman allocation with unknown variances has been long studied in various fields. van der Laan (2008) and Hahn et al. (2011) develop algorithms for estimating the gap parameter Δ itself in an adaptive experiment with the Neyman allocation. They estimate the variances and show their algorithms' optimalities under the framework of semiparametric efficiency, which closely connects to the Gaussian approximation of estimators using the central limit theorem. Although they show their optimality under the framework, they do not investigate the asymptotic optimality in the large-deviation framework. Tabord-Meehan (2022), Kato et al. (2020), and Zhao (2023) also attempt to adrress related problems.

Jourdan et al. (2023) examines BAI with unknown variances in a fixed-confidence setting. Beyond the difference in settings (we focus on fixed-budget BAI), the methods of deriving lower bounds differ between our approach and theirs. They determine the lower bound while incorporating the assumption that the variances are unknown. Moreover, under a large-gap regime (Δ is fixed), they confirm a discrepancy between the lower bounds when variances are known versus unknown. Specifically, they consider alternative hypotheses related to both variances and means. In contrast, the lower bounds presented by Kaufmann et al. (2016) and ourselves are based on alternative hypotheses with fixed variances. While Jourdan et al. (2023) suggests that the upper bounds of strategies with unknown variances cannot align with the lower bound when variances are known, our findings indicate a match under the small-gap regime.

6.3 The AIPW, IPW, and Sample Average Estimators

A key component of our analysis is the AIPW estimator, which comprises an MDS and boasts minimum asymptotic variance. By using the properties of an MDS, we tackle the dependence among observations. The upper bound can also be applied to the IPW estimator, but in this case, the upper bound may not coincide with the lower bound. This discrepancy occurs because the AIPW estimator's asymptotic variance is smaller than that of the IPW estimator. The minimum variance property of the AIPW estimator stems from the efficient influence function (Hahn, 1998; Tsiatis, 2007).

We conjecture that the asymptotic optimality of strategies employing the naive sample average estimator in the recommendation rule can be demonstrated, although we do not prove it in this study. This is because Hahn et al. (2011) shows that, using the CLT, the AIPW and sample average estimators have the same asymptotic distribution. However, due to the inability to utilize MDS properties and the presence of sample dependency, the analysis becomes challenging when we derive a corresponding result for a large deviation (exponential rate of the probability of misidentification). The findings of Hirano et al. (2003) and Hahn et al. (2011) may aid in resolving this issue.

6.4 Extension to BAI in Multi-Armed Bandit (MAB) Problems

In contrast to two-armed bandit problems and BAI with fixed confidence, lower bounds for MAB problems remain unknown. One primary reason is the reversal of KL divergence. Kato et al. (2023b), Degenne (2023), Kato (2023) consider strategies that use sampling rules that are (asymptotically) invariant for any $P^* \in \mathcal{P}^{\mathrm{G}}$. Such a class of strategies is sometimes called *static* in the sense that it cannot estimate parameters during an adaptive experiment to avoid the dependency on P^*. However, if we consider Gaussian bandit models, sampling strategies that are invariant for P^* do not imply non-adaptive (static) strategies because we can still adaptively estimate the variances during an adaptive experiment (the variances are assumed to be the same for any P^*).

7 Conclusion

This study investigated fixed-budget BAI for two-armed Gaussian bandits with unknown variances. We first reviewed the lower bound shown by Kaufmann et al. (2016). Then, we proposed the NA-AIPW strategy and found that its probability of misidentification matches the lower bound when the budget approaches infinity and the gap between the expected rewards of the two arms approaches zero. We referred to this setting as the small-gap regime and the optimality as the local asymptotic optimality. Although there are several remaining open questions, our result provides insight into long-standing open problems in BAI.

References

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

Dohyun Ahn, Dongwook Shin, and Assaf Zeevi. Online ordinal optimization under model misspecification, 2021. URL `https://api.semanticscholar.org/CorpusID:235389954`. SSRN.

Michael G. Akritas and Stavros Kourouklis. Local bahadur efficiency of score tests. *Journal of Statistical Planning and Inference*, 19(2):187–199, 1988.

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.

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.

Heejung Bang and James M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

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.

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.

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.

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

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

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.

Xuming He and Qi-man Shao. Bahadur efficiency and robustness of studentized score tests. *Annals of the Institute of Statistical Mathematics*, 48(2):295–314, 1996.

Keisuke Hirano, Guido Imbens, and Geert Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 2003.

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

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.

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

Masahiro Kato. Worst-case optimal multi-armed gaussian best arm identification with a fixed budget, 2023. arXiv:2310.19788.

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

Junpei Komiyama, Kaito Ariu, Masahiro Kato, and Chao Qin. Rate-optimal bayesian simple regret in best arm identification. *Mathematics of Operations Research*, 2023.

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

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.

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

James M. Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866, 1994.

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

Max Tabord-Meehan. Stratification Trees for Adaptive Randomisation in Randomised Controlled Trials. *The Review of Economic Studies*, 90(5):2646–2673, 2022.

Anastasios Tsiatis. *Semiparametric Theory and Missing Data.* Springer Series in Statistics. Springer New York, 2007.

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. *Asymptotic Statistics.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

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

Po-An Wang, Ruo-Chun Tzeng, and Alexandre Proutiere. Best arm identification with fixed budget: A large deviation perspective. In *Advances in Neural Information Processing Systems*, 2023b.

Harry S. Wieand. A Condition Under Which the Pitman and Bahadur Approaches to Efficiency Coincide. *The Annals of Statistics*, 4(5):1003 – 1011, 1976.

Jinglong Zhao. Adaptive neyman allocation, 2023. arXiv:2309.08808.