OPTIMAL FIXED-BUDGET BEST ARM IDENTIFICATION USING THE AUGMENTED INVERSE PROBABILITY WEIGHTING ESTIMATOR IN TWO-ARMED GAUSSIAN BANDITS WITH UNKNOWN VARIANCES

A PREPRINT

Masahiro Kato*[1], Kaito Ariu*[1,2], Masaaki Imaizumi[3], Masatoshi Uehara[4], Masahiro Nomura[1], and Chao Qin[5]

[1]AI Lab, CyberAgent, Inc.
[2]School of Electrical Engineering and Computer Science, KTH
[3]Department of Basic Science / Komaba Institute for Science, the University of Tokyo
[4]Department of Computer Science, Cornell University
[5]Columbia Business School, Columbia University

January 24, 2022

ABSTRACT

We consider the fixed-budget best arm identification problem in two-armed Gaussian bandits with unknown variances. The tightest lower bound on the complexity and an algorithm whose performance guarantee matches the lower bound have long been open problems when the variances are unknown and when the algorithm is agnostic to the optimal proportion of the arm draws. In this paper, we propose a strategy comprising a sampling rule with randomized sampling (RS) following the estimated target allocation probabilities of arm draws and a recommendation rule using the augmented inverse probability weighting (AIPW) estimator, which is often used in the causal inference literature. We refer to our strategy as the RS-AIPW strategy. In the theoretical analysis, we first derive a large deviation principle for martingales, which can be used when the second moment converges in mean, and apply it to our proposed strategy. Then, we show that the proposed strategy is asymptotically optimal in the sense that the probability of misidentification achieves the lower bound by Kaufmann et al. (2016) when the sample size becomes infinitely large and the gap between the two arms goes to zero.

1 Introduction

This paper studies the best arm identification (BAI) with a fixed budget in stochastic two-armed bandit problems—also known as A/B testing (Kaufmann et al., 2014). The goal is to identify an arm that has the highest expected reward with the smallest failure probability under a fixed time horizon (Audibert et al., 2010; Bubeck et al., 2011; Carpentier and Locatelli, 2016).

Formally, we consider the following problem setting. We are given a fixed budget T. At each time $t \in [T] = \{1, 2, \dots, T\}$, an agent selects an arm $A_t \in \{1, 2\}$. Then, the agent immediately receives a reward (or outcome) X_t, which is linked to arm A_t. This observation process is called the bandit feedback. Following the Rubin causal model (Rubin, 1974), we relate the reward in round t to a potential reward as $X_t = \mathbb{1}[A_t = 1]X_{1,t} + \mathbb{1}[A_t = 2]X_{2,t}$, where $X_{1,t}, X_{2,t} \in \mathbb{R}$ are potential independent (random) rewards. Throughout of this paper, the potential rewards follow Gaussian distributions with unknown means and variances. We denote the reward distributions of the potential outcomes (bandit model) by $\nu = (\boldsymbol{\mu}, \boldsymbol{\sigma})$; here, $\boldsymbol{\mu} = (\mu_1, \mu_2)$ and $\boldsymbol{\sigma} = (\sigma_1, \sigma_2)$ denote the mean and standard deviation, respectively. We also suppose that $\sigma_1, \sigma_2 > 0$. Let $\mathbb{P}_\nu$ and $\mathbb{E}_\nu$ be the probability and expectation under the model ν,

*Equal contributions. Masahiro Kato: masahiro_kato@cyberagent.co.jp. Kaito Ariu: ariu@kth.se.

respectively. We assume that ν belongs to a class $\Omega = \{(\boldsymbol{\mu}, \boldsymbol{\sigma}) : \max_{a\in\{1,2\}} \mu_a > \min_{a'\in\{1,2\}} \mu_{a'}\}$; that is, the best arm is uniquely defined as $a^*(\nu) = \arg\max_{a\in\{1,2\}} \mu_a$. For model class Ω, we make a mild assumption on the values of means and the variances; that is, there exist positive constants C_μ and C_{σ^2} such that, for all $\nu = ((\mu_1, \mu_2), (\sigma_1, \sigma_2)) \in \Omega$, for each $a \in \{1, 2\}$, $|\mu_a| \leq C_\mu$ and $\max\{\sigma^2_a, 1/\sigma^2_a\} \leq C_{\sigma^2}$. We assume that Ω satisfies some regularity conditions (see Section 2). We also denote the gap $\Delta = mu_1 - \mu_2$. In the BAI with a fixed budget, we recommend an arm, $\hat{a}_T$, after pulling arms for the T period. The goal is to minimize the probability of misidentificaton defined as

$$\mathbb{P}_\nu(\hat{a}_T \neq a^*(\nu)).$$

Now, we define the recommendation rule $\hat{a}_T$ and associated rules in more detail. Let $\mathcal{F}_{t-1} = \sigma(A_1, X_1, \dots, A_{t-1}, X_{t-1})$ be the sigma algebra generated by all observations up to time t. In general, the strategy of the BAI with a fixed budget is characterized by a pair $\pi = ((A_t), \hat{a}_T)$, where

- the sampling rule selects an arm A_t for each t based on past observations $\mathcal{F}_{t-1}$ (here, A_t is $\mathcal{F}_{t-1}$-measurable).
- the recommendation rule estimates the best arm $\hat{a}_T$ based on observation up to time T (here, $\hat{a}_T$ is $\mathcal{F}_T$-measurable).

A strategy π is called *consistent* if, for every choice of $\nu \in \Omega$, $\mathbb{P}_\nu(\hat{a}_T \neq a^*(\nu))$ tends to zero as T increases to infinity. We define a family of strategies $\mathcal{P}$ as the set of all consistent strategies. Note that $\mathcal{P}$ is non-empty, as a strategy with a uniform sampling rule and recommending rule by the highest empirical average is consistent.

Glynn and Juneja (2004) discusses target allocation ratio, including non-Gaussian bandit models. They also note that the variances characterize the target allocation ratio when the distributions are Gaussian. As the variances are given in Glynn and Juneja (2004), they do not consider the estimation; that is, the problem is static. For the BAI with a fixed budget, Kaufmann et al. (2014, 2016) prove a problem-dependent lower bound for two-armed BAI with a fixed budget; when the rewards follow Gaussian distributions, the problem-dependent lower bound of the probability of misidentification is characterized as

$$\limsup_{T\to\infty} -\frac{1}{T}\log \mathbb{P}_\nu(\hat{a}_T \neq a^*(\nu)) \leq \frac{\Delta^2}{2(\sigma_1+\sigma_2)^2}. \tag{1}$$

However, Kaufmann et al. (2014, 2016) only propose static algorithms considering the optimal proportion of arm draws, similar to Glynn and Juneja (2004). If we include the estimation of the target allocation ratio, which potentially depends on the parameters of bandit models, the probability that the estimation error of the the parameters is non-negligible may affect the probability of misidentification.

To the best of our knowledge, although it appears simple at first glance, the optimal algorithm for a two-armed Gaussian BAI with a fixed budget, when the algorithm is agnostic to the optimal proportion of the arm draws, is surprisingly an open problem.

Contribution. The study contribution is that it proposes a consistent strategy that asymptotically achieves the lower bound. The proposed strategy consists of a sampling rule with random sampling (RS), following an estimated target allocation ratio of arm draws and in a recommendation rule using the augmented inverse probability weighting (AIPW) estimator, which is often used in the causal inference literature to make the asymptotic variance small. We call our strategy the RS-AIPW strategy, which we prove to be asymptotically optimal in that the probability of misidentification attains the lower bound by Kaufmann et al. (2016) when the sample size becomes infinitely large and the gap between the two arms goes to zero. In Theorem 4.1, we present the performance guarantee of the RS-AIPW strategy, as

$$\liminf_{\Delta\to 0}\liminf_{T\to\infty} -\frac{1}{T\Delta^2}\log \mathbb{P}_\nu\left(\hat{a}_T \neq a^*(\nu)\right) \geq \frac{1}{2(\sigma_1+\sigma_2)^2}.$$

This implies that, when the gap decays to zero, the strategy is optimal in that its performance guarantee matches the lower bound (1).

We use the AIPW estimator to perform a theoretical analysis with large deviations for martingales. In the AIPW estimator, the elements of the sample average constitute a martingale difference sequence. For the sequence, we apply a large deviation principle for martingales, which is derived in Section 3 and Appendix E.

Our theorem is optimal when gap Δ goes to zero. This implies that the lower bound is tight in the case when the problem becomes the most difficult. An intuitive explanation of this result is that we consider a challenging problem wherein the estimator error of the target allocation ratio is relatively insensitive to the probability of misidentification. To the best of our knowledge, our study is the first to consider the BAI with a fixed budget. Conversely, in the BAI with fixed confidence—another formulation of the BAI, Jamieson et al. (2014) proposes the lil' UCB, which is optimal scaling of the complexity is guaranteed when the gap goes to zero.

Here, the key point is to usage of the AIPW estimator. For instance, the large deviation expansions can also be applied to the inverse probability weighting (IPW) estimator; however, in this case, the upper bound does not match the lower bound. This is because the fact that the asymptotic variance of the AIPW estimator is smaller than that of the IPW, which is a well-known property in causal inference. Additionally, by focusing on the martingale difference sequence, we simplify the dependency among the observations. Controlling dependency is a challenging task in this problem.

Organization. The remainder of this paper is organized as follows. In Section 2, we describe the lower bound and target allocation ratio, following Kaufmann et al. (2014, 2016). In Section 3, we define the RS-AIPW strategy. In Section 4, we present our main theorem on the asymptotic optimality of our proposed RS-AIPW strategy with its proof. In Section 5, we discuss the related work. In Section 6, we provide the empirical results of the numerical experiments to demonstrate the efficiency of our strategy.

2 Problem-dependent Lower Bound

This section characterizes the lower bound and target allocation ratio in the BAI with a fixed budget. The problem-dependent lower bounds in the regret minimization problem have been popularized by Lai and Robbins (1985), where we use the changes in distributions to derive them. Kaufmann et al. (2014, 2016) derive the problem-dependent lower bounds specific to each reward distribution in the BAI with fixed confidence and fixed budget. In particular, when the rewards follow the Gaussian distributions with the target allocation, Kaufmann et al. (2014, 2016) show the following problem-dependent lower bound.

Proposition 2.1 (From Theorem 12 in Kaufmann et al. (2016))**.** *Let us consider the fixed budget setting with Gaussian rewards with unknown variances. For each $\nu \in \Omega$, any consistent strategy satisfies*

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

3 Proposed Algorithm: The RS-AIPW Strategy

In this section, we define our BAI strategy, which consists of sampling and recommendation rules. First, based on the results of Glynn and Juneja (2004) and ,Kaufmann et al. (2014, 2016), we define the target allocation ratio as

$$w_1^* = \frac{\sigma_1}{\sigma_1+\sigma_2}, \qquad w_2^* = 1 - w_1^*.$$

This target allocation ratio is unknown when the variances are unknown; therefore, to use this ratio, we need to estimate it from observations during the trials. In each $t \in [T]$, our sampling rule randomly draws an arm, following an estimated target allocation ratio; that is, we pull an arm with the probability identical to the estimated target allocation ratio (target allocation probability). Unlike the algorithm of Garivier and Kaufmann (2016), we do not attempt to track the optimal proportion. In final round T, we recommend an arm with the highest estimated mean reward. Here, we estimated the mean reward using the AIPW estimator, which is known to reduce the sensitivity of the estimation error of the target allocation ratio to the asymptotic variance. Based on these rules, we refer to this as the RS-AIPW strategy.

We assume that the means and variances of the arms are bounded by certain constants.

Assumption 3.1. *There exist known positive constants C_μ and C_{σ^2} such that, for any $\nu = ((\mu_1,\mu_2),(\sigma_1,\sigma_2)) \in \Omega$, for all $a \in \{1,2\}$,*

$$|\mu_a| \leq C_\mu \quad \textit{and} \quad 1/C_{\sigma^2} \leq \sigma_a^2 \leq C_{\sigma^2}.$$

Note these positive constants C_μ and C_{σ^2} are introduced for technical purposes. We can use a sufficiently large positive number for C_μ and C_{σ^2}.

3.1 Sampling Rule with Estimated Targeting Probability

In round t, for all $a \in \{1,2\}$, we estimate the target allocation ratio w_a^* using past observations and denote the $\mathcal{F}_{t-1}$ measurable estimator as $w_{a,t}$ such that $\forall a \in \{1,2\},\ w_{a,t} > 0$ and $\sum_{a\in\{1,2\}} w_{a,t} = 1$. Then, in round t, we select arm

a with probability $w_{a,t}$. To construct $w_{a,t}$, we estimate σ^2_a for $a \in \{1,2\}$ in round t as

$$\hat{\sigma}^2_{a,t} = \begin{cases} C_{\sigma^2} & \text{if } C_{\sigma^2} < \tilde{\sigma}^2_{a,t} \\ \tilde{\sigma}^2_{a,t} & \text{if } 1/C_{\sigma^2} \leq \tilde{\sigma}^2_{a,t} \leq C_{\sigma^2} \\ 1/C_{\sigma^2} & \text{if } \tilde{\sigma}^2_{a,t} < 1/C_{\sigma^2}, \end{cases} \tag{2}$$

where for each $t \in \{1, 2 \dots, T\}$,

$$\tilde{\mu}_{a,t} = \frac{1}{\sum^{t-1}_{s=1} \mathbb{1}[A_s = a]} \sum^{t-1}_{s=1} \mathbb{1}[A_s = a] X_{a,s}, \tag{3}$$

$$\tilde{\zeta}_{a,t} = \frac{1}{\sum^{t-1}_{s=1} \mathbb{1}[A_s = a]} \sum^{t-1}_{s=1} \mathbb{1}[A_s = a] X^2_{a,s}, \quad \text{and}$$

$$\tilde{\sigma}^2_{a,t} = \tilde{\zeta}_{a,t} - \left(\tilde{\mu}_{a,t}\right)^2.$$

When $\sum^{t-1}_{s=1} \mathbb{1}[A_s = a] = 0$, we define $\tilde{\mu}_{a,t} = \tilde{\zeta}_{a,t} = 0$. Then, we construct $w_{a,t}$ as

$$w_{1,t} = \frac{\sqrt{\hat{\sigma}^2_{1,t}}}{\sqrt{\hat{\sigma}^2_{1,t}} + \sqrt{\hat{\sigma}^2_{2,t}}}, \qquad w_{2,t} = 1 - w_{1,t}.$$

We refer to this sampling strategy as an RS strategy. We use this strategy to apply the large deviation expansion for martingales to the estimator of the expected reward, which is the core of our theoretical analysis in Section 4.

3.2 Recommendation Rule with the AIPW Estimator

In the recommendation phase in round T, for each $a \in \{1, 2\}$, we construct an estimator of μ_a, as

$$\hat{\mu}^{\mathrm{AIPW}}_{a,T} = \frac{1}{T} \sum^{T}_{t=1} \hat{X}_{a,t}, \tag{4}$$

where

$$\hat{X}_{a,t} = \frac{\mathbb{1}[A_t = a]\big(X_{a,t} - \hat{\mu}_{a,t}\big)}{w_{a,t}} + \hat{\mu}_{a,t}, \qquad \text{and}$$

$$\hat{\mu}_{a,t} = \begin{cases} C_\mu & \text{if } C_\mu < \tilde{\mu}_{a,t} \\ \tilde{\mu}_{a,t} & \text{if } -C_\mu \leq \tilde{\mu}_{a,t} \leq C_\mu \\ -C_\mu & \text{if } \tilde{\mu}_{a,t} < -C_\mu. \end{cases} \tag{5}$$

In the end of round $t = T$, we recommend $\hat{a}_T \in \arg\max_{a \in \{1,2\}} \hat{\mu}^{\mathrm{AIPW}}_{a,T}$. In other words, our recommendation rule is

$$\hat{a}_T = \left\{ \begin{array}{ll} 1 & \text{if} \quad \hat{\mu}^{\mathrm{AIPW}}_{1,T} \geq \hat{\mu}^{\mathrm{AIPW}}_{2,T}, \\ 2 & \text{otherwise}. \end{array} \right. \tag{6}$$

Both $\hat{\mu}^{\mathrm{AIPW}}_{a,T}$ and $\{\hat{\mu}_{a,t}\}^T_{t=1}$ are estimators of μ_a. As explained, the AIPW estimator $\hat{\mu}^{\mathrm{AIPW}}_{a,T}$ has the following advantages. (i) Its components satisfy the martingale property, and hence, allow us to use the large deviation principle shown in Theorem 4.2. Moreover, (ii) it has the smallest asymptotic variance among the possible estimators, which is needed for achieving the lower bound. On the other hand, we cannot use the martingale property for $\hat{\mu}_{a,t}$, which makes the theoretical analysis is difficult. Besides, we can construct other estimators with a martingale property, such as the inverse probability weighting (IPW) estimator, but their asymptotic variance is not necessarily the same as the AIPW estimator. For instance, as we explain in Section 6, the variance of the IPW estimator can be larger than that of the AIPW estimator. If the asymptotic variance becomes large, the upper bound does not match the lower bound.

3.3 Implementation of the RS-AIPW Strategy

We show the pseudo-code in Algorithm 1. We note again that C_μ and C_{σ^2} are introduced for technical purposes in order for the estimators to be bounded . Therefore, any large positive value can be assumed. In the pseudo-code, for brevity, only the first two rounds are used for initialization. In practice, the number of initialization rounds can be adjusted for each application.

Algorithm 1 RS-AIPW strategy

```
Parameter: Positive constants $C_\mu$ and $C_{\sigma^2}$.
Initialization:
At $t = 1, 2$, select $A_t = t$ and set $w_{a,t} = 0.5$ for $a \in \{1, 2\}$.
for $t = 3$ to $T$ do
    Construct $\hat{\mu}_{a,t}$ and $w_{a,t}$ following (2) and (5).
    Draw $\xi_t$ from the uniform distribution on $[0, 1]$.
    $A_t = 1$ if $\xi_t \leq w_{1,t}$; $A_t = 2$ if $\xi_t > w_{1,t}$.
end for
Construct $\hat{\mu}_{a,T}$ and $w_{a,T}$ for $a \in \{1, 2\}$.
Construct $\hat{\mu}^{\mathrm{AIPW}}_{a,T}$ for $a \in \{1, 2\}$. following (4).
Recommend $\hat{a}_T$ following (6).
```

4 Asymptotic Optimality of the RS-AIPW Strategy

The following theorem provides the asymptotic optimality of the proposed RS-AIPW estimator when $T \to \infty$ and $\Delta \to 0$.

Theorem 4.1 (Asymptotic Optimality of the RS-AIPW Strategy)**.** *Suppose that Assumption 3.1 holds. Fix any standard deviations σ_1, σ_2 in the model class Ω. Under the RS-AIPW strategy,*

$$\liminf_{\Delta \to 0} \liminf_{T \to \infty} -\frac{1}{T\Delta^2} \log \mathbb{P}_\nu \left(\hat{a}_T \neq a^*(\nu)\right) \geq \frac{1}{2(\sigma_1 + \sigma_2)^2},$$

where the limit $\liminf_{\Delta \to 0}$ is taken by the models in Ω.

In BAI, we are interested in the evaluation of the exponentially small probability of misidentification. For this purpose, the central limit theorem cannot provide an answer because it gives an approximation around the mean, not the tail. On the other hand, owing to the non-stationary adaptive process in BAI, it is also difficult to apply the large deviation principle (Dembo and Zeitouni, 2009) to the simple sample average. For example, although Gärtner-Ellis theorem (Gärtner, 1977; Ellis, 1984) provides a large deviation principle for dependent samples, it requires the existence of the logarithmic moment generating function, which potentially holds in our problem but is not easy to be proved. Besides, even if we can use the result, the upper bound of the logarithmic probability (rate function) of misidentification has a complex form and is not easy to be connected with the lower bound. For these problems, we first propose constructing the AIPW estimator, whose elements consist of a martingale difference sequence. Then, we can apply a large deviation principle for a martingale difference sequences shown in Theorem 4.2. We derive this large deviation principle from the Cramér large deviation expansion for martingales by Grama and Haeusler (2000) and Fan et al. (2013, 2014). Then, letting the gap Δ go to zero, we show that the upper bound matches the lower bound. Our result can be regarded as a Gaussian approximation of the probability of misidentification, but it evaluates the probability of a region that cannot be approximated by the central limit theorem.

Thus, the proof is based on Cramér large deviation expansions for martingales shown in Fan et al. (2013, 2014). By extending these results and extending with the property of the AIPW estimator, we can show Theorem 4.1. The remaining part of this section provides the proof of Theorem 4.1.

4.1 Cramér Large Deviation Expansions for the AIPW Estimator

First, we present the large deviation principle for martingales. Our large deviation principle is inspired by (Grama and Haeusler, 2000) and Fan et al. (2013, 2014). Note that their original large deviation principle is only applicable to martingales whose conditional second moment is bounded by any accuracy if the fixed budget is larger than a constant given deterministically in advance. However, in BAI, we can only bound the conditional second moment by any accuracy if the fixed budget is larger than a constant given in a random path. This randomness prevents us from applying the original results of (Grama and Haeusler, 2000) and Fan et al. (2013, 2014). In this paper, we show that under the mean convergence of the unconditional second moment, we can derive the large deviation principle for martingales given by the BAI strategy.

Let us define $\xi_{1,2,t}$ as

$$\xi_{1,2,t} = \frac{\hat{X}_{1,t} - \hat{X}_{2,t} - \Delta}{\sqrt{T}(\sigma_1 + \sigma_2)}.$$

Let us also define

$$Z_0 = 0, \qquad Z_s = \sum_{t=1}^{s} \xi_{1,2,t}, \quad \text{for } s = 1, \dots, T$$

and

$$W_0 = 0, \qquad W_s = \sum_{t=1}^{s} \mathbb{E}_\nu\left[\xi^2_{1,2,t} | \mathcal{F}_{t-1}\right], \quad \text{for } s = 1, \dots, T.$$

Here, $\{(\xi_{1,2,t}, \mathcal{F}_t)\}_{t=1}^T$ is a martingale difference sequence because

$$\begin{aligned}
\mathbb{E}_\nu\left[\xi_{1,2,t}|\mathcal{F}_{t-1}\right] &= \frac{1}{\sqrt{T}(\sigma_1+\sigma_2)}\mathbb{E}_\nu\left[\hat{X}_{1,t} - \hat{X}_{2,t} - \Delta \big| \mathcal{F}_{t-1}\right]\\
&= \frac{1}{\sqrt{T}(\sigma_1+\sigma_2)}\mathbb{E}_\nu\left[\frac{\mathbb{1}[A_t = 1]\big(X_{1,t} - \hat{\mu}_{1,t}\big)}{w_{1,t}} + \hat{\mu}_{1,t} - \frac{\mathbb{1}[A_t = 2]\big(X_{2,t} - \hat{\mu}_{2,t}\big)}{w_{2,t}} - \hat{\mu}_{2,t} - \Delta\big|\mathcal{F}_{t-1}\right]\\
&= \frac{1}{\sqrt{T}(\sigma_1+\sigma_2)}\mathbb{E}_\nu\left[\frac{w_{1,t}\big(X_{1,t} - \hat{\mu}_{1,t}\big)}{w_{1,t}} + \hat{\mu}_{1,t} - \frac{w_{2,t}\big(X_{2,t} - \hat{\mu}_{2,t}\big)}{w_{2,t}} - \hat{\mu}_{2,t} - \Delta\big|\mathcal{F}_{t-1}\right]\\
&= \frac{1}{\sqrt{T}(\sigma_1+\sigma_2)}(\mu_1 - \mu_2 - \Delta) = 0.
\end{aligned}$$

We used the fact that $\hat{\mu}_{a,t}$ and $w_{a,t}$ are $\mathcal{F}_{t-1}$-measurable random variables.

Let us also define

$$V_T = \mathbb{E}_\nu\left[\left|\frac{1}{T\left(\sigma_1+\sigma_2\right)^2}\sum_{t=1}^{T}\mathbb{E}_\nu\left[\left(\hat{X}_{1,t} - \hat{X}_{2,t} - \Delta\right)^2 \Big| \mathcal{F}_{t-1}\right] - 1\right|\right]$$

and the cumulative distribution function of the standard normal distribution:

$$\Phi(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}\exp\left(-\frac{t^2}{2}\right)dt.$$

Theorem 4.2. *Suppose that the following condition hold:*

Condition A: $\sup_{1\leq t\leq T}\mathbb{E}_\nu\left[\exp\left(C_0\sqrt{T}\left|\xi_{1,2,t}\right|\right) \Big| \mathcal{F}_{t-1}\right] \leq C_1$ *for some positive constants* C_0 *and* C_1.

Then, there exist constants $c_1, c_2 > 0$ *such that, for all* $1 \leq u \leq \sqrt{T}\min\left\{\frac{1}{4}C_0, \sqrt{\frac{3C_0^2}{8C_1}}\right\}$, *we have*

$$\frac{\mathbb{P}_\nu\left(Z_T \leq -u\right)}{\Phi(-u)} \leq c_1 u \exp\left(c_2\left(\frac{u^3}{\sqrt{T}} + \frac{u^4}{T} + u^2 V_T\right)\right),$$

where constants c_1, c_2 *depends on* C_0 *and* C_1, *but does not depend on* $\{(\xi_t, \mathcal{F}_t)\}_{t=1}^T$, u, *and the bandit model* ν.

From Theorem 4.2, we can directly obtain the following result. Suppose that $\lim_{T\to\infty} V_T = 0$ and $\frac{\Delta}{(\sigma_1+\sigma_2)^2} \leq \min\left\{\frac{1}{4}C_0, \sqrt{\frac{3C_0^2}{8C_1}}\right\}$ (without loss of generality as $\Delta \to 0$). Then, if we let $u = \frac{\sqrt{T}\Delta}{\sigma_1+\sigma_2}$, we have the following equalities:

$$\begin{aligned}
\mathbb{P}_\nu\left(Z_T \leq -\frac{\sqrt{T}\Delta}{\sigma_1+\sigma_2}\right) &= \mathbb{P}_\nu\left(\sum_{t=1}^{T}\frac{\hat{X}_{1,t} - \hat{X}_{2,t} - \Delta}{\sqrt{T}(\sigma_1+\sigma_2)} \leq -\frac{\sqrt{T}\Delta}{(\sigma_1+\sigma_2)}\right)\\
&= \mathbb{P}_\nu\left(\hat{\mu}^{\mathrm{AIPW}}_{1,T} \leq \hat{\mu}^{\mathrm{AIPW}}_{2,T}\right).
\end{aligned}$$

Therefore, we have

$$\frac{\mathbb{P}_\nu\left(\hat{\mu}^{\mathrm{AIPW}}_{1,T} \leq \hat{\mu}^{\mathrm{AIPW}}_{2,T}\right)}{\Phi(-\frac{\sqrt{T}\Delta}{\sigma_1+\sigma_2})} \leq c_1\frac{\sqrt{T}\Delta}{\sigma_1+\sigma_2}\exp\left(c_2\left(\frac{\Delta^3 T}{(\sigma_1+\sigma_2)^3} + \frac{\Delta^4 T}{(\sigma_1+\sigma_2)^4} + \frac{\Delta^2 T}{(\sigma_1+\sigma_2)^2}V_T\right)\right). \tag{7}$$

4.2 Gaussian Approximation under Small Gap

Finally, we consider an approximation of the large deviation principle. First, we have the following upper bound and lower bound on $\Phi(-x)$ (see (Section 2.2., Fan et al., 2013))

$$\frac{1}{\sqrt{2\pi}(1+u)}\exp\left(-\frac{u^2}{2}\right) \leq \Phi(-u) \leq \frac{1}{\sqrt{\pi}(1+u)}\exp\left(-\frac{u^2}{2}\right),\ u \geq 0. \tag{8}$$

By combining this bound with Theorem 4.2 and (7), we have the following corollary.

Corollary 4.3. *In addition to Condition A in Theorem 4.2, suppose that the following conditions also hold:*

Condition B: $\frac{\Delta}{(\sigma_1+\sigma_2)^2} \leq \min\left\{\frac{1}{4}C_0, \sqrt{\frac{3C_0^2}{8C_1}}\right\}$.

Condition C: $\lim_{T\to\infty} V_T = 0$.

Then, with some constant $c > 0$, we have

$$\liminf_{T\to\infty} -\frac{1}{T}\log \mathbb{P}_\nu\left(\hat{\mu}^{\mathrm{AIPW}}_{1,T} \leq \hat{\mu}^{\mathrm{AIPW}}_{2,T}\right) \geq \frac{\Delta^2}{2(\sigma_1+\sigma_2)^2} - c\left(\frac{\Delta^3}{(\sigma_1+\sigma_2)^3} + \frac{\Delta^4}{(\sigma_1+\sigma_2)^4}\right).$$

This approximation can be thought of as a Gaussian approximation because the probability $\mathbb{P}_\nu\left(\hat{\mu}^{\mathrm{AIPW}}_{1,T} \leq \hat{\mu}^{\mathrm{AIPW}}_{2,T}\right)$ is represented by $\exp\left(-\frac{\Delta^2}{2(\sigma_1+\sigma_2)^2}T\right)$.

To use Corollary 4.3, we need to show that Conditions A and C hold. First, the following lemma state that Condition A holds with the constants C_0 and C_1 that are universal to the problems in Ω.

Lemma 4.4. *For each positive constant C_0, there exists positive constant $C_1(C_0, C_{\sigma^2}, C_\mu)$ which depends only on C_0, C_{σ^2}, C_μ such that for any $\mu_1, \mu_2 \in [0,1]$, we have*

$$\sup_{t\in[T]} \mathbb{E}_\nu[\exp(C_0\sqrt{T}|\xi_{1,2,t}|) \mid \mathcal{F}_{t-1}] \leq C_1(C_0, C_{\sigma^2}, C_\mu).$$

Next, regarding Condition C, we introduce a lemma for the convergence of V_T.

Lemma 4.5. *Suppose that Assumption 3.1 hold. Under the RS-AIPW strategy, $\lim_{T\to\infty} V_T = 0$; that is, for any $\delta > 0$, there exists T_0 such that for all $T > T_0$,*

$$\mathbb{E}\left[\left|\frac{1}{T(\sigma_1+\sigma_2)^2}\sum_{t=1}^T \mathbb{E}_\nu\left[\left(\hat{X}_{1,t} - \hat{X}_{2,t} - \Delta\right)^2 \middle| \mathcal{F}_{t-1}\right] - 1\right|\right] \leq \delta,$$

The proofs of Lemma 4.4 and Lemma 4.5 are shown in Appendix C and D, respectively. By combining Corollary 4.3 and Lemmas 4.4–4.5, we conclude the proof of Theorem 4.1.

5 Related Work

The stochastic multi-armed bandit (MAB) problem is a classical abstraction of the sequential decision-making problem (Thompson, 1933; Robbins, 1952; Lai and Robbins, 1985). BAI is a paradigm of the MAB problem, where we consider pure exploration to find the best arm (Even-Dar et al., 2006; Audibert et al., 2010; Bubeck et al., 2011). Though the problem of BAI itself goes back decades, variants go as far back as the 1950s in the context of ranking and selection problems that were studied in statistics. Some of the earliest advances on this topic are summarized in Bechhofer et al. (1968).

In the BAI literature, there is another setting called BAI with fixed confidence (Jennison et al., 1982; Mannor and Tsitsiklis, 2004; Kalyanakrishnan et al., 2012). For the fixed confidence setting, Garivier and Kaufmann (2016) provides an algorithm whose upper bound matches the problem-dependent lower bound. In contrast, in the fixed-budget setting, It was not clear whether there was an algorithm that matches the lower bound proposed by Kaufmann et al. (2016). In some studies, the lower bounds are associated with the gap Δ^d Audibert et al. (2010); Bubeck et al. (2011); Carpentier and Locatelli (2016), unlike the problem-dependent lower bounds Kaufmann et al. (2016). In addition to these studies, Russo (2016) proposes the TTTS strategy with another direction of asymptotic optimality. There are other Bayesian algorithms for BAI, such as Qin et al. (2017). Shang et al. (2020) shows the asymptotic optimality of the TTTS in the fixed confidence setting. Komiyama et al. (2021) discusses the optimality of Bayesian simple regret minimization, which is closely related to BAI in a Bayesian setting. They showed that parameters with a small gap make a significant contribution to Bayesian simple regret.

The asymptotically optimal strategy proposed by Garivier and Kaufmann (2016) is called the Track-Stop strategy because it tracks the estimated target allocation ratio until it stops by the predetermined stopping rule. This strategy is further developed by Degenne et al. (2019). Unlike the Track-Stop strategy, where $w_{a,t}$ can be 0, our proposed RS-AIPW strategy randomly pulls arms following an estimated target allocation ratio for $w_{a,t}$ to be bounded away from zero to construct the AIPW estimator, which uses the inverse of the probability.

There is less research on BAI with a fixed budget compared to BAI with fixed confidence. Following Audibert et al. (2010) and Bubeck et al. (2011), Gabillon et al. (2012) and Karnin et al. (2013) discuss the link between the fixed confidence and fixed budget settings. Carpentier and Locatelli (2016) discusses the optimality of the method proposed by Audibert et al. (2010) up to constant for a limited number of bandit problem instances. Several studies consider fixed budget BAI with linear models (Hoffman et al., 2014; Katz-Samuels et al., 2020).

Another literature on ordinal optimization has been studied in the operation research community (Ahn et al., 2021), and a modern formulation was established in the 2000s (Chen et al., 2000; Glynn and Juneja, 2004). Most of those studies have considered the estimation of the target allocation ratio separately from the error rate under known target allocations.

In the economics literature, Hirano and Porter (2009, 2020) consider how we conduct decision-making on arm selection given observations. Although their interests are a bit different from BAI, their results are established based on local asymptotic normality, which also considers a case where the gap goes to zero to derive the optimality. Liang et al. (2019) considers a problem similar to BAI with continuous time setting. Kasy and Sautmann (2021) and Ariu et al. (2021) apply a BAI strategy proposed by Russo (2016) (Top-Two Thompson Sampling; TTTS) in economic application. van der Laan (2008), Hahn et al. (2011), Tabord-Meehan (2018), Kato et al. (2020) and Gupta et al. (2021) also consider related problem in efficient treatment effect estimation with adaptive experiments.

The AIPW estimator, which is also referred to as a doubly robust estimator when the allocation probability is unknown, has been used in causal inference literature (Hahn, 1998; Hirano et al., 2003; van der Laan, 2008; Dudík et al., 2011; Hahn et al., 2011; van der Laan and Lendle, 2014; Luedtke and van der Laan, 2016; Wang et al., 2017; Chernozhukov et al., 2018; Narita et al., 2019; Bibaut et al., 2019; Oberst and Sontag, 2019; Hadad et al., 2021; Kato et al., 2020, 2021; Bibaut et al., 2021; Zhan et al., 2021). Note that when constructing AIPW estimator with samples obtained from adaptive experiments including BAI algorithm, a typical construction is to use sample splitting and martingales (van der Laan, 2008; Hadad et al., 2021; Kato et al., 2021). The AIPW estimator is also used in the recent bandit literature, mainly in regret minimization (Dimakopoulou et al., 2021; Kim et al., 2021).

6 Experiments

In this section, we show the soundness of the proposed RS-AIPW strategy through experiments. We compare our proposed RS-AIPW (RA) strategy with the alpha-elimination (Alpha, Kaufmann et al., 2014, 2016) and uniform sampling strategies (Uniform). The alpha-elimination strategy is an oracle strategy, which assumes that the variances are known. See Kaufmann et al. (2014, 2016) for more details of the alpha-elimination strategy. In the uniform sampling strategy, we select arm 1 if the round is odd and select 2 otherwise; finally, we recommend an arm with the highest average reward defined as $\tilde{\mu}_{a,T+1}$ in Section 3. Besides, we also investigate the performances of the following two additional strategy:

RS-DR (RD) strategy: We replace the AIPW estimator in the RS-AIPW strategy with the DR estimator defined as

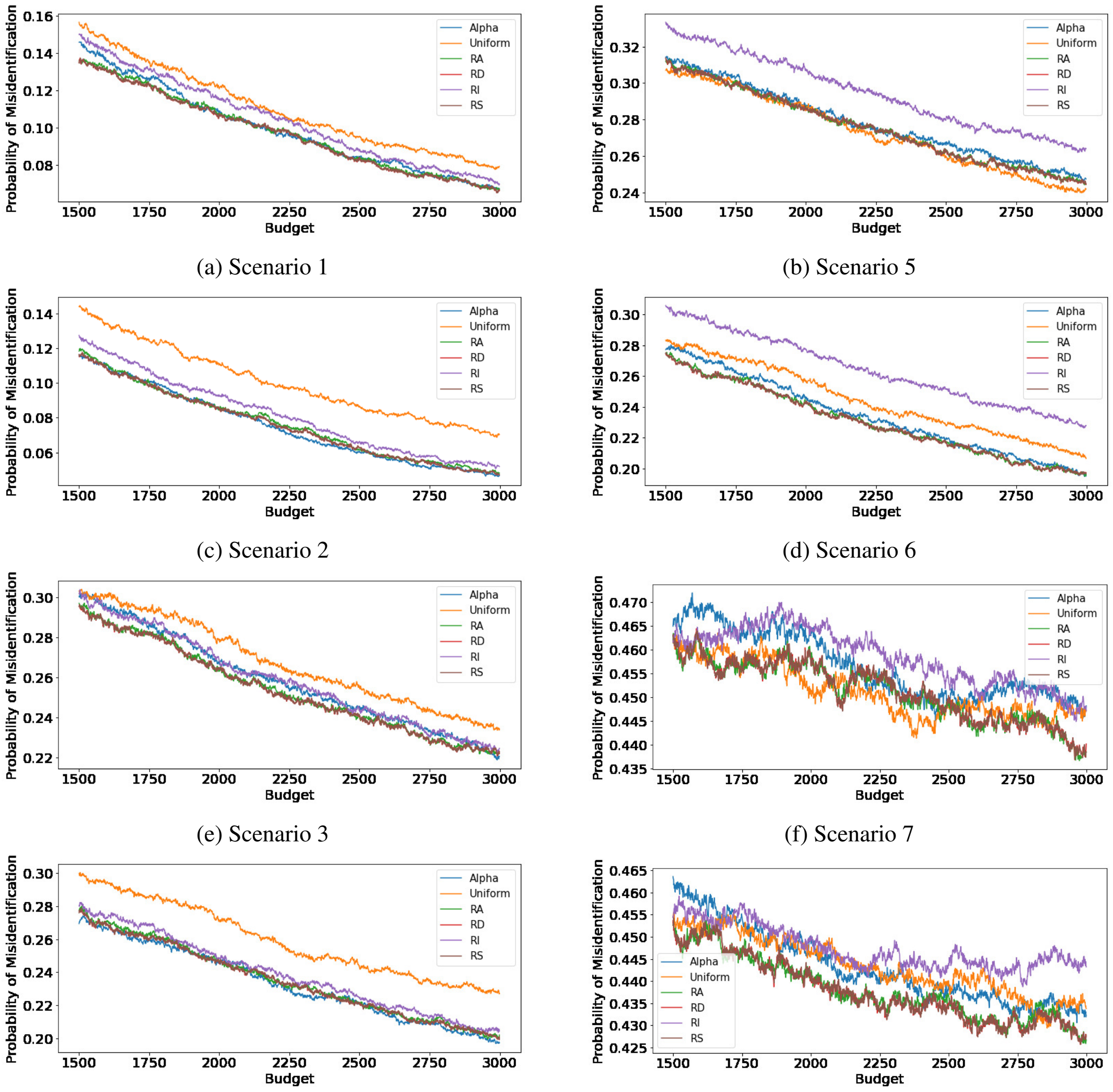
$$\hat{\mu}^{\mathrm{DR}}_{a,T} = \frac{1}{T}\sum_{t=1}^{T}\hat{X}^{\dagger}_{a,t}, \qquad \hat{X}^{\dagger}_{a,t} = \frac{\mathbb{1}[A_t = a]\big(X_{a,t} - \hat{\mu}_{a,t}\big)}{\frac{1}{t-1}\sum_{s=1}^{t-1}\mathbb{1}[A_s = a]} + \hat{\mu}_{a,t}.$$

RS-IPW (RI) strategy: We replace the AIPW estimator in the RS-AIPW strategy with the IPW estimator defined as

$$\hat{\mu}^{\mathrm{IPW}}_{a,T} = \frac{1}{T}\sum_{t=1}^{T}\hat{X}^{\diamond}_{a,t}, \qquad \hat{X}^{\diamond}_{a,t} = \frac{\mathbb{1}[A_t = a]X_{a,t}}{w_{a,t}}.$$

RS-SA (RS) strategy: We replace the AIPW estimator in the RS-AIPW strategy with the simple sample average defined as

$$\hat{\mu}^{\mathrm{SA}}_{a,T} = \tilde{\mu}_{a,T+1},$$



(a) Scenario 1

(b) Scenario 5

(c) Scenario 2

(d) Scenario 6

(e) Scenario 3

(f) Scenario 7

(g) Scenario 4

(h) Scenario 8

Figure 1: Results of the simulation studies. We compute the empirical probability of misidentification.

where $\tilde{\mu}_{a,T+1}$ is defined in (3).

The DR estimator replaces the allocation probability $w_{a,t}$ with its estimator. The IPW estimator does not use the adjustment term $\hat{\mu}_{a,t}$. The sample average corresponds to the IPW estimator whose allocation probability is replaced with its estimator $\frac{1}{T}\sum_{t=1}^{T}\mathbb{1}[A_t = a]$ because

$$\tilde{\mu}_{a,T+1} = \frac{1}{\sum_{t=1}^{T}\mathbb{1}[A_t = a]}\sum_{t=1}^{T}\mathbb{1}[A_t = a]X_{a,t} = \frac{1}{T}\sum_{t=1}^{T}\frac{\mathbb{1}[A_t = a]X_{a,t}}{\frac{1}{T}\sum_{t=1}^{T}\mathbb{1}[A_t = a]}.$$

Besides, to stabilize the allocation probability $w_{a,t}$, instead of directly using $w_{a,t}$, we use

$$w^{\gamma}_{a,t} = \gamma_t\frac{1}{2} + (1-\gamma_t)w_{a,t}.$$

for γ_t such that $\gamma_t \to 0$ as $t \to \infty$. This prevents $w_{a,t}$ from being some extreme value. Note that $w^{\gamma}_{a,t} \xrightarrow{\text{a.s.}} w^*_a$ if $w_{a,t} \xrightarrow{\text{a.s.}} w^*_a$. In this experiments, we use $w^{\gamma}_{a,t}$ with $\gamma_t = \frac{1}{\sqrt{t}}$ and 100 initialization rounds for the RA, RD, RI, and RS strategies.

We consider the following ten sample scenarios:

Scenario 1 $\mu_1 = 0.05$, $\mu_2 = 0.01$, $\sigma_1^2 = 1$, and $\sigma_2^2 = 0.2$;
Scenario 2 $\mu_1 = 0.05$, $\mu_2 = 0.01$, $\sigma_1^2 = 1$, and $\sigma_2^2 = 0.1$;
Scenario 3 $\mu_1 = 0.05$, $\mu_2 = 0.03$, $\sigma_1^2 = 1$, and $\sigma_2^2 = 0.2$;
Scenario 4 $\mu_1 = 0.05$, $\mu_2 = 0.01$, $\sigma_1^2 = 1$, and $\sigma_2^2 = 0.1$;
Scenario 5 $\mu_1 = 0.8$, $\mu_2 = 0.75$, $\sigma_1^2 = 5$, and $\sigma_2^2 = 3$;
Scenario 6 $\mu_1 = 0.8$, $\mu_2 = 0.75$, $\sigma_1^2 = 5$, and $\sigma_2^2 = 1$;
Scenario 7 $\mu_1 = 0.8$, $\mu_2 = 0.79$, $\sigma_1^2 = 5$, and $\sigma_2^2 = 3$;
Scenario 8 $\mu_1 = 0.8$, $\mu_2 = 0.79$, $\sigma_1^2 = 5$, and $\sigma_2^2 = 1$.

We set $T = 1,000$ and ran $10,000$ independent trials for each setting. Although we set $T = 1,000$, we save the recommended arm for each $t \in \{1, 2, \dots, 10000\}$. Then, by taking the average over $10,000$ independent trials, we compute the empirical probability of identification for $t \in \{1, 2, \dots, 3000\}$. The results of the experiments are shown in Figure 1.

We can confirm that the RS-AIPW, RS-DR, and RS-SA strategies perform as well as the alpha-elimination. On the other hand, the uniform sampling and RS-IPW strategies show sub-optimal results. The sub-optimal performance of the RS-IPW strategy can be considered that the IPW estimator has a larger variance than that of the AIPW estimator; therefore, the upper bound of the RS-IPW strategy does not match the lower bound.

As well as Lemma 4.5, we can show the following corollary.

Corollary 6.1. *Suppose that Assumption 3.1 hold. Under the RS-IPW strategy, for any $\delta > 0$, there exists T_0 such that for all $T > T_0$,*

$$\mathbb{E}\left[\left|\frac{1}{T\kappa}\sum_{t=1}^{T}\mathbb{E}_\nu\left[\left(\hat{X}_{1,t} - \hat{X}_{2,t} - \Delta\right)^2 \middle| \mathcal{F}_{t-1}\right] - 1\right|\right] \leq \delta,$$

where

$$\kappa = \frac{\mathbb{E}_\nu[X_{1,t}^2]}{w_1^*} + \frac{\mathbb{E}_\nu[X_{2,t}^2]}{w_2^*}.$$

Thus, if we use the RS-IPW strategy, we cannot achieve the lower bound owing to the variance κ, which is different from that of Lemma 4.5.

We conjecture that the reason why the RS-DR strategy performs well is that the re-estimated allocation probability mitigates the fluctuation of the allocation probability $w_{a,t}$. A similar phenomenon is reported by Kato et al. (2021) in the context of off-policy evaluation as a paradox because even if we know the true allocation probability, replacing it with an estimator improves the performance.

We conjecture that the RS-SA strategy also has the same optimal asymptotically optimal properties as our proposed RS-AIPW strategy. However, since we cannot use large deviation expansion for martingales to investigate the asymptotic properties of the RS-SA strategy, the analysis becomes difficult owing to the dependency. The results of Hirano et al. (2003) and Hahn et al. (2011) may be helpful for solving this problem, which discusses a similar problem when showing the asymptotic normality. Investigating the asymptotic properties of the RS-SA strategy is a future task, but it does not improve on our proposed RS-AIPW strategy, which is shown to be optimal.

7 Conclusion

We provided an asymptotically optimal strategy for two-armed Gaussian BAI in the fixed budget setting with unknown variances for a case where the gap between the arms goes to zero. To show the optimality, we used the martingale and variance-reduction properties of our proposed RS-AIPW strategy. Our result provides an insight into long-standing open problems in the bandit community.

References

Ahn, D., Shin, D., and Zeevi, A. (2021), "Online Ordinal Optimization under Model Misspecification," . 8

Ariu, K., Kato, M., Komiyama, J., McAlinn, K., and Qin, C. (2021), "Policy Choice and Best Arm Identification: Asymptotic Analysis of Exploration Sampling," . 8

Audibert, J.-Y., Bubeck, S., and Munos, R. (2010), "Best Arm Identification in Multi-Armed Bandits," in *The 23rd Conference on Learning Theory*, pp. 41–53. 1, 7, 8

Bechhofer, R., Kiefer, J., and Sobel, M. (1968), *Sequential Identification and Ranking Procedures: With Special Reference to Koopman-Darmois Populations*, University of Chicago Press. 7

Bibaut, A., Chambaz, A., Dimakopoulou, M., Kallus, N., and van der Laan, M. (2021), "Post-Contextual-Bandit Inference," . 8

Bibaut, A., Malenica, I., Vlassis, N., and Van Der Laan, M. (2019), "More Efficient Off-Policy Evaluation through Regularized Targeted Learning," in *ICML*. 8

Bubeck, S., Munos, R., and Stoltz, G. (2011), "Pure exploration in finitely-armed and continuous-armed bandits," *Theoretical Computer Science*. 1, 7, 8

Carpentier, A. and Locatelli, A. (2016), "Tight (Lower) Bounds for the Fixed Budget Best Arm Identification Bandit Problem," in *COLT*. 1, 7, 8

Chen, C.-H., Lin, J., Yücesan, E., and Chick, S. E. (2000), "Simulation Budget Allocation for Further Enhancing TheEfficiency of Ordinal Optimization," *Discrete Event Dynamic Systems*, 10, 251–270. 8

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018), "Double/debiased machine learning for treatment and structural parameters," *The Econometrics Journal*. 8

Degenne, R., Koolen, W. M., and Ménard, P. (2019), "Non-Asymptotic Pure Exploration by Solving Games," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., vol. 32. 8

Dembo, A. and Zeitouni, O. (2009), *Large Deviations Techniques and Applications*, Stochastic Modelling and Applied Probability, Springer Berlin Heidelberg. 5

Dimakopoulou, M., Ren, Z., and Zhou, Z. (2021), "Online Multi-Armed Bandits with Adaptive Inference," in *Thirty-Fifth Conference on Neural Information Processing Systems*. 8

Dudík, M., Langford, J., and Li, L. (2011), "Doubly Robust Policy Evaluation and Learning," in *ICML*. 8

Ellis, R. S. (1984), "Large Deviations for a General Class of Random Vectors," *The Annals of Probability*, 12, 1 – 12. 5

Even-Dar, E., Mannor, S., Mansour, Y., and Mahadevan, S. (2006), "Action Elimination and Stopping Conditions for the Multi-Armed Bandit and Reinforcement Learning Problems." *Journal of machine learning research*. 7

Fan, X., Grama, I., and Liu, Q. (2013), "Cramér large deviation expansions for martingales under Bernstein's condition," *Stochastic Processes and their Applications*, 123, 3919–3942. 5, 7, 22, 23

— (2014), "A generalization of Cramér large deviations for martingales," *Comptes Rendus Mathematique*, 352, 853–858. 5

Gabillon, V., Ghavamzadeh, M., and Lazaric, A. (2012), "Best Arm Identification: A Unified Approach to Fixed Budget and Fixed Confidence," in *NeurIPS*. 8

Garivier, A. and Kaufmann, E. (2016), "Optimal Best Arm Identification with Fixed Confidence," in *29th Annual Conference on Learning Theory*, Proceedings of Machine Learning Research. 3, 7, 8

Glynn, P. and Juneja, S. (2004), "A large deviations perspective on ordinal optimization," in *Proceedings of the 2004 Winter Simulation Conference, 2004.*, IEEE, vol. 1. 2, 3, 8

Grama, I. and Haeusler, E. (2000), "Large deviations for martingales via Cramér's method," *Stochastic Processes and their Applications*, 85, 279–293. 5

Gupta, S., Lipton, Z. C., and Childers, D. (2021), "Efficient Online Estimation of Causal Effects by Deciding What to Observe," in *Advances in Neural Information Processing Systems*, eds. Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. 8

Gärtner, J. (1977), "On Large Deviations from the Invariant Measure," *Theory of Probability & Its Applications*, 22, 24–39. 5

Hadad, V., Hirshberg, D. A., Zhan, R., Wager, S., and Athey, S. (2021), "Confidence intervals for policy evaluation in adaptive experiments," *Proceedings of the National Academy of Sciences*, 118. 8

Hahn, J. (1998), "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica*, 66, 315–331. 8

Hahn, J., Hirano, K., and Karlan, D. (2011), "Adaptive experimental design using the propensity score," *Journal of Business and Economic Statistics*. 8, 10

Hamilton, J. (1994), *Time series analysis*, Princeton, NJ: Princeton Univ. Press. 14

Hirano, K., Imbens, G., and Ridder, G. (2003), "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica*. 8, 10

Hirano, K. and Porter, J. R. (2009), "Asymptotics for Statistical Treatment Rules," *Econometrica*, 77, 1683–1701. 8

— (2020), "Chapter 4 - Asymptotic analysis of statistical decision rules in econometrics," in *Handbook of Econometrics, Volume 7A*, Elsevier, vol. 7 of *Handbook of Econometrics*, pp. 283–354. 8

Hoffman, M., Shahriari, B., and Freitas, N. (2014), "On correlation and budget constraints in model-based bandit optimization with application to automatic machine learning," in *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, Reykjavik, Iceland: PMLR, vol. 33 of *Proceedings of Machine Learning Research*, pp. 365–374. 8

Jamieson, K., Malloy, M., Nowak, R., and Bubeck, S. (2014), "lil' UCB : An Optimal Exploration Algorithm for Multi-Armed Bandits," in *Proceedings of The 27th Conference on Learning Theory*. 3

Jennison, C., Johnstone, I. M., and Turnbull, B. W. (1982), "Asymptotically Optimal Procedures for Sequential Adaptive Selection of the Best of Several Normal Means," in *Statistical Decision Theory and Related Topics III*, Academic Press, pp. 55–86. 7

Kalyanakrishnan, S., Tewari, A., Auer, P., and Stone, P. (2012), "PAC Subset Selection in Stochastic Multi-Armed Bandits," in *International Conference on Machine Learning*, Omnipress, ICML'12, p. 227–234. 7

Karnin, Z., Koren, T., and Somekh, O. (2013), "Almost optimal exploration in multi-armed bandits," in *International Conference on Machine Learning*. 8

Kasy, M. and Sautmann, A. (2021), "Adaptive Treatment Assignment in Experiments for Policy Choice," *Econometrica*, 89, 113–132. 8

Kato, M., Ishihara, T., Honda, J., and Narita, Y. (2020), "Adaptive Experimental Design for Efficient Treatment Effect Estimation: Randomized Allocation via Contextual Bandit Algorithm," *arXiv:2002.05308*. 8

Kato, M., McAlinn, K., and Yasui, S. (2021), "The Adaptive Doubly Robust Estimator and a Paradox Concerning Logging Policy," in *Thirty-Fifth Conference on Neural Information Processing Systems*. 8, 10

Katz-Samuels, J., Jain, L., Karnin, Z., and Jamieson, K. (2020), "An Empirical Process Approach to the Union Bound: Practical Algorithms for Combinatorial and Linear Bandits," . 8

Kaufmann, E., Cappé, O., and Garivier, A. (2016), "On the Complexity of Best-Arm Identification in Multi-Armed Bandit Models," *Journal of Machine Learning Research*, 17, 1–42. 1, 2, 3, 7, 8, 14

Kaufmann, E., Cappé, O., and Garivier, A. (2014), "On the Complexity of A/B Testing," in *Proceedings of The 27th Conference on Learning Theory*, eds. Balcan, M. F., Feldman, V., and Szepesvári, C., Barcelona, Spain: PMLR, vol. 35 of *Proceedings of Machine Learning Research*, pp. 461–481. 1, 2, 3, 8

Kim, W., Kim, G.-S., and Paik, M. C. (2021), "Doubly Robust Thompson Sampling with Linear Payoffs," in *Thirty-Fifth Conference on Neural Information Processing Systems*. 8

Komiyama, J., Ariu, K., Kato, M., and Qin, C. (2021), "Optimal Simple Regret in Bayesian Best Arm Identification," . 8

Lai, T. and Robbins, H. (1985), "Asymptotically efficient adaptive allocation rules," *Advances in Applied Mathematics*. 3, 7

Liang, A., Mu, X., and Syrgkanis, V. (2019), "Dynamically Aggregating Diverse Information," . 8

Loeve, M. (1977), *Probability Theory*, Graduate Texts in Mathematics, Springer. 14

Luedtke, A. R. and van der Laan, M. J. (2016), "Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy," *Annals of Statistics*. 8

Mannor, S. and Tsitsiklis, J. N. (2004), "The sample complexity of exploration in the multi-armed bandit problem," *Journal of Machine Learning Research*. 7

Narita, Y., Yasui, S., and Yata, K. (2019), "Efficient Counterfactual Learning from Bandit Feedback," in *AAAI*. 8

Oberst, M. and Sontag, D. (2019), "Counterfactual Off-Policy Evaluation with Gumbel-Max Structural Causal Models," in *ICML*. 8

Qin, C., Klabjan, D., and Russo, D. (2017), "Improving the Expected Improvement Algorithm," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., vol. 30. 8

Robbins, H. (1952), "Some aspects of the sequential design of experiments," *Bulletin of the American Mathematical Society*. 7

Rubin, D. B. (1974), "Estimating causal effects of treatments in randomized and nonrandomized studies," *Journal of Educational Psychology*. 1

Russo, D. (2016), "Simple Bayesian Algorithms for Best Arm Identification," . 7, 8

Shang, X., de Heide, R., Menard, P., Kaufmann, E., and Valko, M. (2020), "Fixed-confidence guarantees for Bayesian best-arm identification," in *International Conference on Artificial Intelligence and Statistics*, vol. 108 of *Proceedings of Machine Learning Research*, pp. 1823–1832. 8

Tabord-Meehan, M. (2018), "Stratification Trees for Adaptive Randomization in Randomized Controlled Trials," . 8

Thompson, W. R. (1933), "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples," *Biometrika*. 7

van der Laan, M. J. (2008), "The Construction and Analysis of Adaptive Group Sequential Designs," . 8

van der Laan, M. J. and Lendle, S. D. (2014), "Online Targeted Learning," . 8

Wang, Y.-X., Agarwal, A., and Dudik, M. (2017), "Optimal and adaptive off-policy evaluation in contextual bandits," in *ICML*. 8

Zhan, R., Hadad, V., Hirshberg, D. A., and Athey, S. (2021), "Off-Policy Evaluation via Adaptive Weighting with Data from Contextual Bandits," . 8

A Preliminaries

A.1 Mathematical Tools

Definition A.1. *[Uniformly Integrable, Hamilton (1994), p. 191] A sequence $\{A_t\}$ is said to be uniformly integrable if for every $\epsilon > 0$ there exists a number $c > 0$ such that*

$$\mathbb{E}_\nu[|A_t| \cdot I[|A_t \geq c|]] < \epsilon$$

for all t.

The following proposition is from Hamilton (1994), Proposition 7.7, p. 191.

Proposition A.2 (Sufficient Conditions for Uniform Integrability)**.** *(a) Suppose there exist $r > 1$ and $M < \infty$ such that $\mathbb{E}_\nu[|A_t|^r] < M$ for all t. Then $\{A_t\}$ is uniformly integrable. (b) Suppose there exist $r > 1$ and $M < \infty$ such that $\mathbb{E}_\nu[|b_t|^r] < M$ for all t. If $A_t = \sum_{j=-\infty}^{\infty} h_j b_{t-j}$ with $\sum_{j=-\infty}^{\infty} |h_j| < \infty$, then $\{A_t\}$ is uniformly integrable.*

Proposition A.3 (L^r Convergence Theorem, p 165, Loeve (1977))**.** *Let $0 < r < \infty$, suppose that $\mathbb{E}_\nu\left[|a_n|^r\right] < \infty$ for all n and that $a_n \xrightarrow{\mathrm{p}} a$ as $n \to \infty$. The following are equivalent:*

(i) $a_n \to a$ in L^r as $n \to \infty$;

(ii) $\mathbb{E}_\nu\left[|a_n|^r\right] \to \mathbb{E}_\nu\left[|a|^r\right] < \infty$ as $n \to \infty$;

(iii) $\left\{|a_n|^r, n \geq 1\right\}$ is uniformly integrable.

B Proof of Proposition 2.1

Let us define $N_{a,\tau} = \sum_{t=1}^{\tau} \mathbb{1}[A_t = a]$ where τ is the stopping time with respect to $(\mathcal{F}_t)_{t\geq 1}$. We follow a proof procedure similar to Theorem 12 in Kaufmann et al. (2016). Without loss of generality, we assume that, for model $\nu = ((\mu_1, \mu_2), (\sigma_1, \sigma_2))$, $\mu_1 > \mu_2$. Consider a perturbed bandit model $\nu' = (\boldsymbol{\mu}', \boldsymbol{\sigma}) = ((\mu'_1, \mu'_2), (\sigma_1, \sigma_2))$, where $\mu'_2 > \mu'_1$. We use the following lemma from Kaufmann et al. (2016).

Lemma B.1 (Lemma 1 in Kaufmann et al. (2016))**.** *Let ν and ν' be two bandit models with K arms such that for all $a \in [K]$, ν_a (distribution of reward of arm a for model ν), and ν'_a (distribution of reward of arm a for the model ν') are mutually absolutely continuous. Define the Kullback-Leibler divergence of the probability distributions p and q as*

$$\mathrm{KL}(p, q) = \begin{cases} \int \log\left(\frac{dp(x)}{dq(x)}\right) dp(x) & \text{if} \quad p \ll q, \\ \infty & \text{otherwise}. \end{cases}$$

For an almost surely finite stopping time τ with respect to $(\mathcal{F}_t)_{t\geq 1}$,

$$\sum_{a=1}^{K} \mathbb{E}_\nu[N_{a,\tau}]\mathrm{KL}(\nu_a, \nu'_a) \geq \sup_{\mathcal{E}\in\mathcal{F}_\tau} d(\mathbb{P}_\nu(\mathcal{E}), \mathbb{P}_{\nu'}(\mathcal{E})), \tag{9}$$

where $d(x, y) := x\log(x/y) + (1 - x)\log((1 - x)/(1 - y))$ is the binary relative entropy with the convention that $d(0, 0) = d(1, 1) = 0$.

Event $\mathcal{E}$ on the right hand side of (9) is measurable (by $\mathcal{F}_\tau$) event. Note that the inequality (9) becomes tight when selecting an event that is discriminative for the two models, ν and ν', such as $\mathbb{P}_\nu(\mathcal{E}) \to 1$ and $\mathbb{P}_\nu(\mathcal{E}) \to 0$. Subsequently, we select $\mathcal{E} = \{\hat{a}_T = 1\}$ as a discriminative event.

By applying Lemma B.1, we have

$$\mathbb{E}_{\nu'}[N_{1,T}]\mathrm{KL}(\nu'_1, \nu_1) + \mathbb{E}_{\nu'}[N_{2,T}]\mathrm{KL}(\nu'_2, \nu_2) \geq \sup_{\mathcal{E}\in\mathcal{F}_T} d(\mathbb{P}_{\nu'}(\mathcal{E}), \mathbb{P}_\nu(\mathcal{E})).$$

Here, for each $a \in \{1, 2\}$, we compute

$$\mathrm{KL}(\nu'_a, \nu_a) = \frac{(\mu'_a - \mu_a)^2}{2\sigma_a^2}.$$

Let $\mathcal{E} = \{\hat{a}_T = 1\}$. Because we assume that the algorithm is consistent for both models, for each $\varepsilon \in (0, 1)$, there exists $t_0(\varepsilon)$ such that for all $T \geq t_0(\varepsilon)$,

$$\mathbb{P}_{\nu'}(\mathcal{E}) \leq \varepsilon \leq \mathbb{P}_\nu(\mathcal{E}).$$

Then, for all $T \geq t_0(\varepsilon)$,

$$\begin{aligned}\mathbb{E}_{\nu'}[N_{1,T}]\mathrm{KL}(\nu'_1, \nu_1) + \mathbb{E}_{\nu'}[N_{2,T}]\mathrm{KL}(\nu'_2, \nu_2) &= \mathbb{E}_{\nu'}[N_{1,T}]\frac{(\mu'_1 - \mu_1)^2}{2\sigma_1^2} + \mathbb{E}_{\nu'}[N_{2,T}]\frac{(\mu'_2 - \mu_2)^2}{2\sigma_2^2}\\ &\geq d(\mathbb{P}_{\nu'}(\mathcal{E}), \mathbb{P}_{\nu}(\mathcal{E}))\\ &\geq d(\varepsilon, 1 - \mathbb{P}_{\nu}(\hat{a}_T \neq 1))\\ &\geq \varepsilon \log \varepsilon + (1-\varepsilon)\log\frac{1-\varepsilon}{\mathbb{P}_{\nu}(\hat{a}_T \neq 1)}.\end{aligned}$$

Considering $\limsup_{T\to\infty}$ and letting $\varepsilon \to 0$, we obtain

$$\begin{aligned}\limsup_{T\to\infty}\frac{1}{T}\log\frac{1}{\mathbb{P}_{\nu}(\hat{a}_T \neq 1)} &\leq \limsup_{T\to\infty}\frac{\mathbb{E}[N_{1,T}]}{T}\frac{(\mu'_1 - \mu_1)^2}{2\sigma_1^2} + \frac{\mathbb{E}[N_{2,T}]}{T}\frac{(\mu'_2 - \mu_2)^2}{2\sigma_2^2}\\ &\leq \max_{a=1,2}\frac{(\mu'_a - \mu_a)^2}{2\sigma_a^2}.\end{aligned}$$

We have

$$\begin{aligned}\limsup_{T\to\infty}\frac{1}{T}\log\frac{1}{\mathbb{P}_{\nu}(\hat{a}_T \neq 1)} &\leq \inf_{(\mu'_1,\mu'_2):\mu'_2>\mu'_1}\max_{a=1,2}\frac{(\mu'_a - \mu_a)^2}{2\sigma_a^2}\\ &= \min_{\lambda\in\mathbb{R}}\max_{a=1,2}\frac{(\lambda - \mu_a)^2}{2\sigma_a^2}.\end{aligned}$$

When the minimum over $\lambda \in \mathbb{R}$ is attained,

$$\begin{aligned}\frac{(\lambda - \mu_1)^2}{2\sigma_1^2} &= \frac{(\lambda - \mu_2)^2}{2\sigma_2^2}\\ \lambda &= \frac{\mu_1\sigma_2 + \mu_2\sigma_1}{\sqrt{2}\sigma_1\sigma_2}.\end{aligned}$$

Thus, we have

$$\limsup_{T\to\infty}\frac{1}{T}\log\frac{1}{\mathbb{P}_{\nu}(\hat{a}_T \neq 1)} \leq \frac{(\mu_1 - \mu_2)^2}{2(\sigma_1 + \sigma_2)^2}$$

This concludes the proof.

C Proof of Lemma 4.4

Proof. For the simplicity, let us denote $\mathbb{E}_{\nu}$ and $\sigma_1 + \sigma_2$ by $\mathbb{E}$ and $\tilde{\sigma}$, respectively. Recall that $\hat{X}_{a,t}$ is constructed as

$$\hat{X}_{a,t} = \frac{\mathbb{1}[A_t = a]\big(X_{a,t} - \hat{\mu}_{a,t}\big)}{w_{a,t}} + \hat{\mu}_{a,t},$$

where

$$\hat{\mu}_{a,t} = \begin{cases} C_\mu & \text{if } C_\mu < \tilde{\mu}_{a,t}\\ \tilde{\mu}_{a,t} & \text{if } -C_\mu \leq \tilde{\mu}_{a,t} \leq C_\mu\\ -C_\mu & \text{if } \tilde{\mu}_{a,t} < -C_\mu.\end{cases}$$

For each $t = 1, \dots, T$, we have

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(C_0\sqrt{T}|\xi_{1,2,t}|\right)\middle|\mathcal{F}_{t-1}\right]\\
&= \mathbb{E}\left[\exp\left(C_0\left|\frac{\hat{X}_{1,t} - \hat{X}_{2,t} - \Delta}{\tilde{\sigma}}\right|\right)\middle|\mathcal{F}_{t-1}\right]\\
&\leq \mathbb{E}\left[\exp\left(\frac{C_0}{\tilde{\sigma}}\left|\hat{X}_{1,t} - \hat{X}_{2,t}\right| + \frac{C_0\Delta}{\tilde{\sigma}}\right)\middle|\mathcal{F}_{t-1}\right]\\
&\leq \mathbb{E}\left[\exp\left(\frac{C_0}{\tilde{\sigma}}\left|\hat{X}_{1,t} - \hat{X}_{2,t}\right| + \frac{2C_0C_\mu}{\tilde{\sigma}}\right)\middle|\mathcal{F}_{t-1}\right]\\
&\overset{(a)}{=} \tilde{C}_1\mathbb{E}\left[\exp\left(\frac{C_0}{\tilde{\sigma}}|\hat{X}_{1,t} - \hat{X}_{2,t}|\right)\middle|\mathcal{F}_{t-1}, A_t = 1\right]\mathbb{P}(A_t = 1 \mid \mathcal{F}_{t-1})\\
&\quad + \tilde{C}_1\mathbb{E}\left[\exp\left(\frac{C_0}{\tilde{\sigma}}|\hat{X}_{1,t} - \hat{X}_{2,t}|\right)\middle|\mathcal{F}_{t-1}, A_t = 2\right]\mathbb{P}(A_t = 2 \mid \mathcal{F}_{t-1})\\
&= \tilde{C}_1\mathbb{E}\left[\exp\left(\frac{C_0}{\tilde{\sigma}}\left|\frac{(X_{1,t} - \hat{\mu}_{1,t})}{w_{1,t}} + \hat{\mu}_{1,t} - \hat{\mu}_{2,t}\right|\right)\middle|\mathcal{F}_{t-1}, A_t = 1\right]\mathbb{P}(A_t = 1 \mid \mathcal{F}_{t-1})\\
&\quad + \tilde{C}_1\mathbb{E}\left[\exp\left(\frac{C_0}{\tilde{\sigma}}\left|-\frac{(X_{2,t} - \hat{\mu}_{2,t})}{w_{2,t}} + \hat{\mu}_{1,t} - \hat{\mu}_{2,t}\right|\right)\middle|\mathcal{F}_{t-1}, A_t = 2\right]\mathbb{P}(A_t = 2 \mid \mathcal{F}_{t-1})
\end{aligned}$$

where for (a), we denote $\tilde{C}_1 = \exp\left(2C_0C_\mu/\tilde{\sigma}\right)$. When $X_{a,t}$ is $\mathcal{N}(\mu_a, \sigma^2_a)$, for each $\lambda \geq 0$,

$$\mathbb{E}[\exp(\lambda X_{a,t})] = \exp\left(\lambda\mu_a - \frac{\sigma^2_a\lambda^2}{2}\right).$$

Therefore, when $|\mu_{a,t}| \leq C_\mu$, $\max\{\sigma^2_a, 1/\sigma^2_a\} \leq C_{\sigma^2}$, $|\hat{\mu}_{a,t}| \leq C_\mu$, and $|1/w_{a,t}| \leq 2/C_{\sigma^2}$ (for all $a \in \{1, 2\}$ and for all $t \in \{1, \dots, T\}$), there exists a positive constant $C_1(C_0, C_{\sigma^2}, C_\mu)$ such that

$$\mathbb{E}\left[\exp(C_0\sqrt{T}|\xi_{1,2,t}|)\middle|\mathcal{F}_{t-1}\right] \leq C_1(C_0, C_{\sigma^2}, C_\mu).$$

This concludes the proof.

□

D Proof of Lemma 4.5

For each $a \in \{1, 2\}$, we denote $\mathbb{E}_\nu[X^2_a]$ by ζ_a.

Lemma D.1. *Suppose that Assumption 3.1 holds. Under the RS-AIPW strategy, for each* $a \in \{1, 2\}$,

$$\begin{aligned}
\hat{\mu}_{a,t} &\xrightarrow{\text{a.s.}} \mu_a,\\
\tilde{\zeta}_{a,t} &\xrightarrow{\text{a.s.}} \zeta_a.
\end{aligned}$$

Proof. Under the RS-AIPW strategy, we select each arm with a positive probability for all $t \in [T]$. This implies that with probability 1, we select each arm infinitely often as $T \to \infty$. Therefore, from the law of large numbers, $\hat{\mu}_{a,t} \xrightarrow{\text{a.s.}} \mu_a$ and $\tilde{\zeta}_{a,t} \xrightarrow{\text{a.s.}} \zeta_a$. □

This lemma directly implies the following corollary, which states the almost sure convergence of $w_{a,t}$.

Corollary D.2. *Suppose that Assumption 3.1 holds. Under the RS-AIPW strategy, for each* $a \in \{1, 2\}$,

$$w_{a,t} \xrightarrow{\text{a.s.}} w^*_a,$$

Then, we present the following results on the convergence of the second moment.

Lemma D.3. *Suppose that Assumption 3.1 holds. Under the RS-AIPW strategy with probability* 1,

$$\lim_{t\to\infty}\left\{\mathbb{E}_\nu\left[\left(\hat{X}_{1,t} - \hat{X}_{2,t} - \Delta\right)^2\middle|\mathcal{F}_{t-1}\right] - (\sigma_1 + \sigma_2)^2\right\} = 0.$$

Proof.

$$
\begin{aligned}
&\mathbb{E}_\nu\left[\left(\hat{X}_{1,t} - \hat{X}_{2,t} - \Delta\right)^2 \Big| \mathcal{F}_{t-1}\right]\\
&= \mathbb{E}_\nu\left[\left(\frac{\mathbb{1}[A_t = 1]\big(X_{1,t} - \hat{\mu}_{1,t}\big)}{w_{1,t}} - \frac{\mathbb{1}[A_t = 2]\big(X_{2,t} - \hat{\mu}_{2,t}\big)}{w_{2,t}} + \hat{\mu}_{1,t} - \hat{\mu}_{2,t} - \Delta\right)^2 \Big| \mathcal{F}_{t-1}\right]\\
&= \mathbb{E}_\nu\Bigg[\left(\frac{\mathbb{1}[A_t = 1]\big(X_{1,t} - \hat{\mu}_{1,t}\big)}{w_{1,t}} - \frac{\mathbb{1}[A_t = 2]\big(X_{2,t} - \hat{\mu}_{2,t}\big)}{w_{2,t}}\right)^2\\
&\qquad + 2\left(\frac{\mathbb{1}[A_t = 1]\big(X_{a,t} - \hat{\mu}_{1,t}\big)}{w_{1,t}} - \frac{\mathbb{1}[A_t = 2]\big(X_{b,t} - \hat{\mu}_{2,t}\big)}{w_{2,t}}\right)\left(\hat{\mu}_{1,t} - \hat{\mu}_{2,t} - \Delta\right)\\
&\qquad + \left(\hat{\mu}_{1,t} - \hat{\mu}_{2,t} - \Delta\right)^2 | \mathcal{F}_{t-1}\Bigg]\\
&= \mathbb{E}_\nu\Bigg[\frac{\mathbb{1}[A_t = 1]\big(X_{1,t} - \hat{\mu}_{1,t}\big)^2}{w^2_{1,t}} + \frac{\mathbb{1}[A_t = 2]\big(X_{2,t} - \hat{\mu}_{2,t}\big)^2}{w^2_{2,t}}\\
&\qquad + 2\left(\frac{\mathbb{1}[A_t = 1]\big(X_{1,t} - \hat{\mu}_{1,t}\big)}{w_{1,t}} - \frac{\mathbb{1}[A_t = 2]\big(X_{2,t} - \hat{\mu}_{2,t}\big)}{w_{2,t}}\right)\left(\hat{\mu}_{1,t} - \hat{\mu}_{2,t} - \Delta\right)\\
&\qquad + \left(\hat{\mu}_{1,t} - \hat{\mu}_{2,t} - \Delta\right)^2 | \mathcal{F}_{t-1}\Bigg]\\
&= \mathbb{E}_\nu\left[\frac{\big(X_{1,t} - \hat{\mu}_{1,t}\big)^2}{w_{1,t}} | \mathcal{F}_{t-1}\right] + \mathbb{E}_\nu\left[\frac{\big(X_{2,t} - \hat{\mu}_{2,t}\big)^2}{w_{2,t}} | \mathcal{F}_{t-1}\right] - \left(\hat{\mu}_{1,t} + \hat{\mu}_{2,t} - \Delta\right)^2.
\end{aligned}
$$

Here, we used

$$
\begin{aligned}
&\mathbb{E}_\nu\left[\frac{\mathbb{1}[A_t = a]\big(X_{a,t} - \hat{\mu}_{a,t}\big)^2}{w^2_{a,t}} | \mathcal{F}_{t-1}\right] = \mathbb{E}_\nu\left[\frac{w_{a,t}\big(X_{a,t} - \hat{\mu}_{a,t}\big)^2}{w^2_{a,t}} | \mathcal{F}_{t-1}\right] = \mathbb{E}_\nu\left[\frac{\big(X_{a,t} - \hat{\mu}_{a,t}\big)^2}{w_{a,t}} | \mathcal{F}_{t-1}\right],\\
&\mathbb{E}_\nu\left[\frac{\mathbb{1}[A_t = 1]\big(X_{1,t} - \hat{\mu}_{1,t}\big)}{w_{1,t}}\left(\hat{\mu}_{1,t} - \hat{\mu}_{2,t} - \Delta\right) | \mathcal{F}_{t-1}\right] = \left(\hat{\mu}_{1,t} - \hat{\mu}_{2,t} - \Delta\right)\mathbb{E}_\nu\left[\frac{w_{1,t}\big(X_{1,t} - \hat{\mu}_{1,t}\big)}{w_{1,t}} | \mathcal{F}_{t-1}\right].
\end{aligned}
$$

We also have

$$
\mathbb{E}_\nu\left[\frac{\big(X_{a,t} - \hat{\mu}_{a,t}\big)^2}{w_{a,t}} | \mathcal{F}_{t-1}\right] = \frac{\mathbb{E}_\nu[X^2_{a,t}] - 2\mu_a\hat{\mu}_{a,t} + \hat{\mu}^2_{a,t}}{w_{a,t}} = \frac{\mathbb{E}_\nu[X^2_{a,t}] - \mu^2_a + (\mu_a - \hat{\mu}_{a,t})^2}{w_{a,t}}.
$$

Then,

$$
\begin{aligned}
&\mathbb{E}_\nu\left[\frac{\big(X_{1,t} - \hat{\mu}_{1,t}\big)^2}{w_{1,t}} | \mathcal{F}_{t-1}\right] + \mathbb{E}_\nu\left[\frac{\big(X_{2,t} - \hat{\mu}_{2,t}\big)^2}{w_{2,t}} | \mathcal{F}_{t-1}\right] - \left(\hat{\mu}_{1,t} + \hat{\mu}_{2,t} - \Delta\right)^2\\
&= \left(\frac{\mathbb{E}_\nu[X^2_{1,t}] - \mu^2_1 + (\mu_1 - \hat{\mu}_{1,t})^2}{w_{1,t}}\right) + \left(\frac{\mathbb{E}_\nu[X^2_{2,t}] - \mu^2_2 + (\mu_2 - \hat{\mu}_{2,t})^2}{w_{2,t}}\right) - \left(\hat{\mu}_{1,t} + \hat{\mu}_{2,t} - \Delta\right)^2.
\end{aligned}
$$

Because $\hat{\mu}_{a,t} \xrightarrow{\text{a.s.}} \mu_a$ and $w_{a,t} \xrightarrow{\text{a.s.}} w^*_a$, with probability 1,

$$
\begin{aligned}
&\lim_{t\to\infty}\left|\left(\frac{\mathbb{E}_\nu[X^2_{1,t}] - \mu^2_1 + (\mu_1 - \hat{\mu}_{1,t})^2}{w_{1,t}}\right) + \left(\frac{\mathbb{E}_\nu[X^2_{2,t}] - \mu^2_2 + (\mu_2 - \hat{\mu}_{2,t})^2}{w_{2,t}}\right) - \left(\hat{\mu}_{1,t} + \hat{\mu}_{2,t} - \Delta\right)^2 - \frac{\sigma^2_1}{w^*_1} - \frac{\sigma^2_2}{w^*_2}\right|\\
&\leq \lim_{t\to\infty}\left|\frac{\mathbb{E}_\nu[X^2_{1,t}] - \mu^2_1}{w_{1,t}} - \frac{\sigma^2_1}{w^*_1}\right| + \lim_{t\to\infty}\left|\frac{\mathbb{E}_\nu[X^2_{2,t}] - \mu^2_2}{w_{2,t}} - \frac{\sigma^2_2}{w^*_2}\right|
\end{aligned}
$$

$$+\lim_{t\to\infty}\frac{(\mu_1-\hat{\mu}_{1,t})^2}{w_{1,t}}+\lim_{t\to\infty}\frac{(\mu_2-\hat{\mu}_{2,t})^2}{w_{2,t}}+\lim_{t\to\infty}\left(\hat{\mu}_{1,t}+\hat{\mu}_{2,t}-\Delta\right)^2$$
$$=0.$$

Note that $\mathbb{E}_\nu[X^2_{a,t}]-\mu^2_a=\sigma^2_a$.

This directly implies that

$$\frac{1}{T}\sum^T_{t=1}\mathbb{E}_\nu\left[\left(\hat{X}_{1,t}-\hat{X}_{2,t}-\Delta\right)^2\Big|\mathcal{F}_{t-1}\right]-(\sigma_1+\sigma_2)^2\xrightarrow{\text{a.s.}}0,$$
$$\Leftrightarrow\frac{1}{T\left(\sigma_1+\sigma_2\right)^2}\sum^T_{t=1}\mathbb{E}_\nu\left[\left(\hat{X}_{1,t}-\hat{X}_{2,t}-\Delta\right)^2\Big|\mathcal{F}_{t-1}\right]-1\xrightarrow{\text{a.s.}}0,$$

Here, note that $\frac{\sigma^2_1}{w^*_1}+\frac{\sigma^2_2}{w^*_2}=(\sigma_1+\sigma_2)^2$ from $w^*_1=\frac{\sigma_1}{\sigma_1+\sigma_2}$ and $w^*_2=1-w^*_1$.

Because the reward follows the Gaussian distribution and the other variables in $\hat{X}_{1,t}$ and $\hat{X}_{2,t}$ are bounded,

$$\frac{1}{T\left(\sigma_1+\sigma_2\right)^2}\sum^T_{t=1}\mathbb{E}_\nu\left[\left(\hat{X}_{1,t}-\hat{X}_{2,t}-\Delta\right)^2\Big|\mathcal{F}_{t-1}\right]-1$$

is uniformly integrable from Proposition A.2. Then, from Proposition A.3, for any δ, there exists T_0 such that for all $T>T_0$

$$\mathbb{E}_\nu\left[\left|\frac{1}{T\left(\sigma_1+\sigma_2\right)^2}\sum^T_{t=1}\mathbb{E}_\nu\left[\left(\hat{X}_{1,t}-\hat{X}_{2,t}-\Delta\right)^2\Big|\mathcal{F}_{t-1}\right]-1\right|\right]\leq\delta.$$

This concludes the proof. □

E Proof of Theorem 4.2: Large Deviation Principle for Martingales

For brevity, let us denote $\xi_{1,2,t}$, $\mathbb{P}_\nu$, and $\mathbb{E}_\nu$ by ξ_t, $\mathbb{P}$, and $\mathbb{E}$, respectively. For all $t=1,\dots,T$, let us define

$$r_t(\lambda)=\frac{\exp\left(\lambda\xi_t\right)}{\mathbb{E}\left[\exp\left(\lambda\xi_t\right)\right]}$$

and

$$\eta_t(\lambda)=\xi_t-b_t(\lambda),$$

where

$$b_t(\lambda)=\mathbb{E}[r_t(\lambda)\xi_t].$$

Then, we obtain the following decomposition:

$$Z_T=U_T(\lambda)+B_T(\lambda),$$

where

$$U_T(\lambda)=\sum^T_{t=1}\eta_t(\lambda)$$

and

$$B_T(\lambda)=\sum^T_{t=1}b_t(\lambda).$$

Let $\Psi_T(\lambda)=\sum^T_{t=1}\log\mathbb{E}\left[\exp\left(\lambda\xi_t\right)\right]$.

Lemma E.1. *Under Condition A,*

$$\mathbb{E}\left[|\xi_t|^k\mid\mathcal{F}_{t-1}\right]\leq k!\left(C_0T^{1/2}\right)^{-k}C_1,\qquad \textit{for all}\quad k\geq 2.$$

Proof. Applying the elementary inequality $x^k/k! \leq \exp(x), \forall x \geq 0$, to $x = C_0|\sqrt{T}\xi_t|$, we have, for $k \geq 2$,

$$|\xi_t|^k \leq k!(C_0T^{1/2})^{-k}\exp(C_0|\sqrt{T}\xi_t|).$$

Taking expectations on both sides, with Condition A, we obtain the desired inequality. Recall that Condition A is

$$\sup_{1\leq t\leq T}\mathbb{E}_\nu\left[\exp\left(C_0\sqrt{T}\,|\xi_t|\right)\,\middle|\,\mathcal{F}_{t-1}\right] \leq C_1$$

for some positive constants C_0 and C_1. □

Lemma E.2. *Under Condition A, there exists some constant $C > 0$ such that for all $0 \leq \lambda \leq \frac{1}{4}C_0\sqrt{T}$,*

$$|B_T(\lambda) - \lambda| \leq C\left(\lambda V_T + \lambda^2/\sqrt{T}\right).$$

Proof. By definition, for $t = 1, \dots, T$,

$$b_t(\lambda) = \frac{\mathbb{E}\left[\xi_t \exp\left(\lambda\xi_t\right)\right]}{\mathbb{E}\left[\exp\left(\lambda\xi_t\right)\right]}.$$

Jensen's inequality and $\mathbb{E}[\xi_t] = \mathbb{E}[\mathbb{E}[\xi_t|\mathcal{F}_{t-1}]] = 0$ implies that $\mathbb{E}[\exp(\lambda\xi_t)] \geq 1$ and

$$\mathbb{E}\left[\xi_t\exp\left(\lambda\xi_t\right)\right] = \mathbb{E}\left[\xi_t\left(\exp\left(\lambda\xi_t\right) - 1\right)\right] \geq 0, \qquad \text{for } \lambda \geq 0.$$

We find that

$$\begin{aligned} B_T(\lambda) &\leq \sum_{t=1}^T \mathbb{E}[\xi_t\exp(\lambda\xi_t)] \\ &= \lambda\mathbb{E}[W_T] + \sum_{t=1}^T\sum_{k=2}^\infty \mathbb{E}\left[\frac{\xi_t(\lambda\xi_t)^k}{k!}\right], \end{aligned}$$

by the Taylor series expansion for $\exp(x)$. Here, using Lemma E.1 and $\mathbb{E}\left[\xi_t^{k+1}\right] = \mathbb{E}\left[\mathbb{E}\left[\xi_t^{k+1}|\mathcal{F}_{t-1}\right]\right]$, for some constant C_2,

$$\begin{aligned} \sum_{t=1}^T\sum_{k=2}^\infty\left|\mathbb{E}\left[\frac{\xi_t(\lambda\xi_t)^k}{k!}\right]\right| &\leq \sum_{t=1}^T\sum_{k=2}^\infty\left|\mathbb{E}\left[\xi_t^{k+1}\right]\right|\frac{\lambda^k}{k!} \\ &\leq \sum_{t=1}^T\sum_{k=2}^\infty(k+1)!\left(C_0T^{1/2}\right)^{-(k+1)}C_1\frac{\lambda^k}{k!} \\ &\leq C_2\lambda^2/\sqrt{T}. \end{aligned} \tag{10}$$

Therefore,

$$B_T(\lambda) \leq \lambda + \lambda V_T + C_2\lambda^2/\sqrt{T}.$$

Next, we show the lower bound of $B_T(\lambda)$. First, by using Lemma E.1, using some constant $C_3 > 0$, for all $0 \leq \lambda \leq \frac{1}{4}C_0\sqrt{T}$,

$$\begin{aligned} \mathbb{E}\left[\exp(\lambda\xi_t)\right] &\leq 1 + \sum_{k=2}^\infty\left|\mathbb{E}\left[\frac{(\lambda\xi_t)^k}{k!}\right]\right| \\ &\leq 1 + C_1\sum_{k=2}^\infty\lambda^k(C_0\sqrt{T})^{-k} \\ &\leq 1 + C_3\lambda^2T^{-1}. \end{aligned}$$

This inequality together with (10) implies the lower bound of $B_T(\lambda)$: for some positive constant C_4,

$$\begin{aligned}
B_T(\lambda) &= \sum_{t=1}^{T} \frac{\mathbb{E}\left[\xi_t \exp\left(\lambda \xi_t\right)\right]}{\mathbb{E}\left[\exp\left(\lambda \xi_t\right)\right]} \\
&\geq \left(\sum_{t=1}^{T} \mathbb{E}[\xi_t \exp(\lambda \xi_t)]\right)\left(1 + C_3\lambda^2 T^{-1}\right)^{-1} \\
&= \left(\lambda W_T + \sum_{t=1}^{T}\sum_{k=2}^{\infty} \mathbb{E}\left[\frac{\xi_t(\lambda\xi_t)^k}{k!}\right]\right)\left(1 + C_3\lambda^2 T^{-1}\right)^{-1} \\
&\geq \left(\lambda W_T - \sum_{t=1}^{T}\sum_{k=2}^{\infty} \left|\mathbb{E}\left[\frac{\xi_t(\lambda\xi_t)^k}{k!}\right]\right|\right)\left(1 + C_3\lambda^2 T^{-1}\right)^{-1} \\
&\geq \left(\lambda - \lambda V_T - C_2\lambda^2/\sqrt{T}\right)\left(1 + C_3\lambda^2 T^{-1}\right)^{-1} \\
&\geq \lambda - \lambda V_T - C_4\lambda^2/\sqrt{T}.
\end{aligned}$$

This concludes the proof. □

Lemma E.3. *Assume Condition A, there exists some constant $C > 0$ such that for all $0 \leq \lambda \leq \frac{1}{4}C_0\sqrt{T}$,*

$$\left|\Psi_T(\lambda) - \frac{\lambda^2}{2}\right| \leq C\left(\lambda^3/\sqrt{T} + \lambda^2 V_T\right).$$

Proof. First, we have $\mathbb{E}\left[\exp(\lambda\xi_t)\right] \geq 1$ from Jensen's inequality. Using a two-term Taylor's expansion of $\log(1+\varphi)$, $\varphi \geq 0$, there exists $0 \leq \varphi_t^\dagger \leq \mathbb{E}\left[\exp(\lambda\xi_t)\right] - 1$ (for $t = 1, \dots, T$) such that

$$\begin{aligned}
\Psi_T(\lambda) &= \log \prod_{t=1}^{T} \mathbb{E}\left[\exp\left(\lambda\xi_t\right)\right] \\
&= \sum_{t=1}^{T}\left(\left(\mathbb{E}\left[\exp(\lambda\xi_t)\right] - 1\right) - \frac{1}{2\left(1+\varphi_t^\dagger\right)^2}\left(\mathbb{E}\left[\exp(\lambda\xi_t)\right] - 1\right)^2\right).
\end{aligned}$$

Because (ξ_t) is a martingale difference sequence, $\mathbb{E}[\xi_t] = \mathbb{E}[\mathbb{E}[\xi_t|\mathcal{F}_{t-1}]] = 0$. Therefore,

$$\Psi_T(\lambda) - \frac{\lambda^2}{2}\mathbb{E}[W_T] = \sum_{t=1}^{T}\left(\left(\mathbb{E}\left[\exp(\lambda\xi_t)\right] - 1\right) - \frac{1}{2\left(1+\varphi_t^\dagger\right)^2}\left(\mathbb{E}\left[\exp(\lambda\xi_t)\right] - 1\right)^2\right) - \sum_{t=1}^{T}\left(\lambda\mathbb{E}[\xi_t] + \frac{\lambda^2}{2}\mathbb{E}[\xi_t^2]\right)$$

Then, by using $\mathbb{E}\left[\exp(\lambda\xi_t)\right] \geq 1$, we have

$$\begin{aligned}
\left|\Psi_T(\lambda) - \frac{\lambda^2}{2}\mathbb{E}[W_T]\right| &\leq \sum_{t=1}^{T}\left|\mathbb{E}\left[\exp(\lambda\xi_t)\right] - 1 - \lambda\mathbb{E}[\xi_t] - \frac{\lambda^2}{2}\mathbb{E}[\xi_t^2]\right| + \frac{1}{2}\sum_{t=1}^{T}\left(\mathbb{E}\left[\exp(\lambda\xi_t)\right] - 1\right)^2 \\
&\leq \sum_{t=1}^{T}\sum_{k=3}^{+\infty}\frac{\lambda^k}{k!}\left|\mathbb{E}\left[\xi_t^k\right]\right| + \frac{1}{2}\sum_{t=1}^{T}\left(\sum_{k=1}^{+\infty}\frac{\lambda^k}{k!}\left|\mathbb{E}\left[\xi_t^k\right]\right|\right)^2.
\end{aligned}$$

From Lemma E.1, for a constant C_3,

$$\left|\Psi_T(\lambda) - \frac{\lambda^2}{2}\mathbb{E}[W_T]\right| \leq C_3\lambda^3/\sqrt{T}$$

In conclusion, we have

$$\left|\Psi_T(\lambda) - \frac{\lambda^2}{2}\right| \leq C_3\lambda^3/\sqrt{T} + \frac{\lambda^2}{2}\left(\mathbb{E}[W_T - 1]\right) \leq C_3\lambda^3/\sqrt{T} + \frac{\lambda^2}{2}\mathbb{E}[|W_T - 1|].$$

Recall that $V_T = \mathbb{E}[|W_T - 1|]$. Then,

$$\left|\Psi_T(\lambda) - \frac{\lambda^2}{2}\right| \leq C\left(\lambda^3/\sqrt{T} + \lambda^2 V_T\right).$$

□

By using Lemmas E.1–E.3, we show the proof of Theorem 4.2.

Proof of Theorem 4.2. There exists some constant $C > 0$ such that for all $1 \leq u \leq \sqrt{T}\min\left\{\frac{1}{4}C_0, \sqrt{\frac{3C_0^2}{8C_1}}\right\}$,

$$\begin{aligned}
&\mathbb{P}\left(Z_T > u\right)\\
&= \int \left(\prod_{t=1}^{T}\frac{\exp\left(\lambda\xi_t\right)}{\mathbb{E}\left[\exp\left(\lambda\xi_t\right)\right]}\right)\left(\prod_{t=1}^{T}\frac{\exp\left(\lambda\xi_t\right)}{\mathbb{E}\left[\exp\left(\lambda\xi_t\right)\right]}\right)^{-1}\mathbb{1}[Z_T > u]\mathrm{d}\mathbb{P}\\
&= \int \left(\prod_{t=1}^{T}\frac{\exp\left(\lambda\xi_t\right)}{\mathbb{E}\left[\exp\left(\lambda\xi_t\right)\right]}\right)\exp\left(-\lambda\sum_{t=1}^{T}\xi_t + \log\left(\prod_{t=1}^{T}\mathbb{E}[\exp(\lambda\xi_t)]\right)\right)\mathbb{1}[Z_T > u]\mathrm{d}\mathbb{P}\\
&= \int \left(\prod_{t=1}^{T}\frac{\exp\left(\lambda\xi_t\right)}{\mathbb{E}\left[\exp\left(\lambda\xi_t\right)\right]}\right)\exp\left(-\lambda Z_T + \Psi_T(\lambda)\right)\mathbb{1}[Z_T > u]\mathrm{d}\mathbb{P}\\
&= \int \left(\prod_{t=1}^{T}\frac{\exp\left(\lambda\xi_t\right)}{\mathbb{E}\left[\exp\left(\lambda\xi_t\right)\right]}\right)\exp\left(-\lambda U_T(\lambda) - \lambda B_T(\lambda) + \Psi_T(\lambda)\right)\mathbb{1}[U_T(\lambda) + B_T(\lambda) > u]\mathrm{d}\mathbb{P},\\
&\leq \int \left(\prod_{t=1}^{T}\frac{\exp\left(\lambda\xi_t\right)}{\mathbb{E}\left[\exp\left(\lambda\xi_t\right)\right]}\right)\exp\left(-\lambda U_T(\lambda) - \frac{\lambda^2}{2} + C(\lambda^3/\sqrt{T} + \lambda^2 V_T)\right)\mathbb{1}\left[U_T(\lambda) + \lambda + C(\lambda V_T + \lambda^2/\sqrt{T}) > u\right]\mathrm{d}\mathbb{P},
\end{aligned}$$

where for the last inequality, we used Lemma E.2 and Lemma E.3.

Let $\overline{\lambda} = \overline{\lambda}(u)$ be the largest solution of the equation

$$\lambda + C(\lambda V_T + \lambda^2/\sqrt{T}) = u.$$

The definition of $\overline{\lambda}$ implies that there exist $C' > 0$ such that, for all $1 \leq u \leq \sqrt{T}\min\left\{\frac{1}{4}C_0, \sqrt{\frac{3C_0^2}{8C_1}}\right\}$,

$$C'u \leq \overline{\lambda}(u) = \frac{2u}{\sqrt{(1 + CV_T)^2 + 4Cu/\sqrt{T}} + CV_T + 1} \leq u \tag{11}$$

and there exists $\theta \in (0, 1]$ such that

$$\begin{aligned}
\overline{\lambda}(u) &= u - C(\overline{\lambda}V_T + \overline{\lambda}^2/\sqrt{T})\\
&= u - C\theta(uV_T + u^2/\sqrt{T}) \in \left[C', \sqrt{T}\min\left\{\frac{1}{4}C_0, \sqrt{\frac{3C_0^2}{8C_1}}\right\}\right].
\end{aligned} \tag{12}$$

Then, we obtain for all $1 \leq u \leq \sqrt{T}\min\left\{\frac{1}{4}C_0, \sqrt{\frac{3C_0^2}{8C_1}}\right\}$,

$$\mathbb{P}\left(Z_T > u\right) \leq \exp\left(C\left(\overline{\lambda}^3 T^{-1/2} + \overline{\lambda}^2 V_T\right) - \overline{\lambda}^2/2\right)\int \left(\prod_{t=1}^{T}\frac{\exp\left(\overline{\lambda}\xi_t\right)}{\mathbb{E}\left[\exp\left(\overline{\lambda}\xi_t\right)\right]}\right)\exp\left(-\overline{\lambda}U_T(\overline{\lambda})\right)\mathbb{1}[U_T(\overline{\lambda}) > 0]\mathrm{d}\mathbb{P}.$$

Here, we have

$$\begin{aligned}
&\int \left(\prod_{t=1}^{T}\frac{\exp\left(\overline{\lambda}\xi_t\right)}{\mathbb{E}\left[\exp\left(\overline{\lambda}\xi_t\right)\right]}\right)\exp\left(-\overline{\lambda}U_T(\overline{\lambda})\right)\mathbb{1}[U_T(\overline{\lambda}) > 0]\mathrm{d}\mathbb{P}\\
&= \mathbb{E}\left[\prod_{t=1}^{T}\frac{\exp\left(\overline{\lambda}\xi_t\right)}{\mathbb{E}\left[\exp\left(\overline{\lambda}\xi_t\right)\right]}\exp\left(-\overline{\lambda}U_T(\overline{\lambda})\right)\mathbb{1}[U_T(\overline{\lambda}) > 0]\right].
\end{aligned}$$

We also define another measure $\widetilde{\mathbb{P}}_\lambda$ as

$$\mathrm{d}\widetilde{\mathbb{P}}_\lambda = \frac{\prod_{t=1}^T \exp\left(\lambda\xi_t\right)}{\mathbb{E}\left[\exp\left(\lambda\sum_{t=1}^T \xi_t\right)\right]}\mathrm{d}\mathbb{P} = \frac{\exp\left(\lambda\sum_{t=1}^T \xi_t\right)}{\mathbb{E}\left[\exp\left(\lambda\sum_{t=1}^T \xi_t\right)\right]}\mathrm{d}\mathbb{P}.$$

Note that $\widetilde{\mathbb{P}}_\lambda$ is a probability measure, as the following holds

$$\begin{aligned}\int \mathrm{d}\widetilde{\mathbb{P}}_\lambda &= \int \frac{\exp\left(\lambda\sum_{t=1}^T \xi_t\right)}{\mathbb{E}\left[\exp\left(\lambda\sum_{t=1}^T \xi_t\right)\right]}\mathrm{d}\mathbb{P}\\ &= \frac{1}{\mathbb{E}\left[\exp\left(\lambda\sum_{t=1}^T \xi_t\right)\right]}\int \exp\left(\lambda\sum_{t=1}^T \xi_t\right)\mathrm{d}\mathbb{P}\\ &= \frac{1}{\mathbb{E}\left[\exp\left(\lambda\sum_{t=1}^T \xi_t\right)\right]}\cdot\mathbb{E}\left[\exp\left(\lambda\sum_{t=1}^T \xi_t\right)\right]\\ &= 1.\end{aligned}$$

We further denote $\widetilde{\mathbb{E}}_\lambda$ as the expectation under the measure $\widetilde{\mathbb{P}}_\lambda$.

In the same way as (37) and (38) in Fan et al. (2013), it is easy to see that

$$\begin{aligned}&\mathbb{E}\left[\prod_{t=1}^T \frac{\exp\left(\overline{\lambda}\xi_t\right)}{\mathbb{E}\left[\exp\left(\overline{\lambda}\xi_t\right)\right]}\exp\left(-\overline{\lambda}U_T(\overline{\lambda})\right)\mathbb{1}[U_T(\overline{\lambda})>0]\right]\\ &= \frac{\mathbb{E}[\exp(\overline{\lambda}\sum_{t=1}^T\xi_t)]}{\prod_{t=1}^T\mathbb{E}\left[\exp\left(\overline{\lambda}\xi_t\right)\right]}\mathbb{E}\left[\frac{\prod_{t=1}^T\exp\left(\overline{\lambda}\xi_t\right)}{\mathbb{E}[\exp(\overline{\lambda}\sum_{t=1}^T\xi_t)]}\exp\left(-\overline{\lambda}U_T(\overline{\lambda})\right)\mathbb{1}[U_T(\overline{\lambda})>0]\right]\\ &= \frac{\mathbb{E}[\exp(\overline{\lambda}\sum_{t=1}^T\xi_t)]}{\prod_{t=1}^T\mathbb{E}\left[\exp\left(\overline{\lambda}\xi_t\right)\right]}\widetilde{\mathbb{E}}_{\overline{\lambda}}[\exp\left(-\overline{\lambda}U_T(\overline{\lambda})\right)\mathbb{1}[U_T(\overline{\lambda})>0]]\\ &= \frac{\mathbb{E}\left[\exp\left(\lambda\sum_{t=1}^T\xi_t\right)\right]}{\prod_{t=1}^T\mathbb{E}\left[\exp\left(\lambda\xi_t\right)\right]}\int_0^\infty \overline{\lambda}\exp(-\overline{\lambda}y)\widetilde{\mathbb{P}}_{\overline{\lambda}}(0<U_T(\overline{\lambda})<y)\mathrm{d}y, \qquad (13)\end{aligned}$$

and for a standard Gaussian random variable $\mathcal{N}$,

$$\mathbb{E}\left[\exp\left(-\overline{\lambda}\mathcal{N}\right)\mathbb{1}[\mathcal{N}>0]\right] = \int_0^\infty \overline{\lambda}\exp(-\overline{\lambda}y)\mathbb{P}(0<\mathcal{N}<y)\mathrm{d}y. \qquad (14)$$

From (13) and (14), we have,

$$\left|\widetilde{\mathbb{E}}_{\overline{\lambda}}[\exp\left(-\overline{\lambda}U_T(\overline{\lambda})\right)\mathbb{1}[U_T(\overline{\lambda})>0]] - \mathbb{E}\left[\exp\left(-\overline{\lambda}\mathcal{N}\right)\mathbb{1}[\mathcal{N}>0]\right]\right| \leq 2\sup_g\left|\widetilde{\mathbb{P}}_{\overline{\lambda}}\left(U_T(\overline{\lambda})\leq g\right)-\Phi(g)\right|$$

Therefore,

$$\begin{aligned}&\mathbb{P}\left(Z_T>u\right)\\ &\leq \frac{\mathbb{E}\left[\exp\left(\lambda\sum_{t=1}^T\xi_t\right)\right]}{\prod_{t=1}^T\mathbb{E}\left[\exp\left(\lambda\xi_t\right)\right]}\exp\left(C\left(\overline{\lambda}^3/\sqrt{T}+\overline{\lambda}^2V_T\right)-\overline{\lambda}^2/2\right)\widetilde{\mathbb{E}}_{\overline{\lambda}}[\exp\left(-\overline{\lambda}U_T(\overline{\lambda})\right)\mathbb{1}[U_T(\overline{\lambda})>0]]\\ &\leq \frac{\mathbb{E}\left[\exp\left(\lambda\sum_{t=1}^T\xi_t\right)\right]}{\prod_{t=1}^T\mathbb{E}\left[\exp\left(\lambda\xi_t\right)\right]}\exp\left(C\left(\overline{\lambda}^3/\sqrt{T}+\overline{\lambda}^2V_T\right)-\overline{\lambda}^2/2\right)\\ &\qquad\times\left(\mathbb{E}\left[\exp\left(-\overline{\lambda}\mathcal{N}\right)\mathbb{1}[\mathcal{N}>0]\right]+2\sup_g\left|\widetilde{\mathbb{P}}_{\overline{\lambda}}\left(U_T(\overline{\lambda})\leq g\right)-\Phi(g)\right|\right)\\ &\leq \frac{\mathbb{E}\left[\exp\left(\lambda\sum_{t=1}^T\xi_t\right)\right]}{\prod_{t=1}^T\mathbb{E}\left[\exp\left(\lambda\xi_t\right)\right]}\exp\left(C\left(\overline{\lambda}^3/\sqrt{T}+\overline{\lambda}^2V_T\right)-\overline{\lambda}^2/2\right)\left(\mathbb{E}\left[\exp\left(-\overline{\lambda}\mathcal{N}\right)\mathbb{1}[\mathcal{N}>0]\right]+2\right).\end{aligned}$$

Here,

$$\exp\left(-\overline{\lambda}^2/2\right)\mathbb{E}\left[\exp\left(-\overline{\lambda}\mathcal{N}\right)\mathbb{1}[\mathcal{N}>0]\right]=\frac{1}{\sqrt{2\pi}}\int_0^\infty \exp\left(-(y+\overline{\lambda})^2\right)\mathrm{d}y = 1-\Phi(\overline{\lambda}).$$

From (41) of Fan et al. (2013), for all $\overline{\lambda}\geq C'$, we have

$$1-\Phi(\overline{\lambda})\geq \frac{C'}{\sqrt{2\pi}(1+C')}\frac{1}{\overline{\lambda}}\exp\left(-\frac{\overline{\lambda}^2}{2}\right).$$

Therefore, with some constant $\tilde{C}$, for all $1\leq u\leq \sqrt{T}\min\left\{\frac{1}{4}C_0, \sqrt{\frac{3C_0^2}{8C_1}}\right\}$,

$$\begin{aligned}\mathbb{P}\left(Z_T>u\right)&\leq \frac{\mathbb{E}\left[\exp\left(\lambda\sum_{t=1}^T\xi_t\right)\right]}{\prod_{t=1}^T\mathbb{E}\left[\exp\left(\lambda\xi_t\right)\right]}\left\{\left(1-\Phi(\overline{\lambda})\right)+\overline{\lambda}\left(1-\Phi(\overline{\lambda})\right)c\right\})\exp\left(C\left(\overline{\lambda}^3/\sqrt{T}+\overline{\lambda}^2V_T\right)\right)\\ &\leq \frac{\mathbb{E}\left[\exp\left(\lambda\sum_{t=1}^T\xi_t\right)\right]}{\prod_{t=1}^T\mathbb{E}\left[\exp\left(\lambda\xi_t\right)\right]}\tilde{C}\overline{\lambda}\left(1-\Phi(\overline{\lambda})\right)\exp\left(C\left(\overline{\lambda}^3/\sqrt{T}+\overline{\lambda}^2V_T\right)\right),\end{aligned}\tag{15}$$

where $c=\sqrt{2\pi}(1+C')/C'$, and $\tilde{C}$ is chosen to be $\tilde{C}\overline{\lambda}\geq(1+\overline{\lambda}c)$ (Note that $\overline{\lambda}\geq C'$ from (11)).
Now, we consider bounding the term

$$\frac{\mathbb{E}\left[\exp\left(\overline{\lambda}\sum_{t=1}^T\xi_t\right)\right]}{\prod_{t=1}^T\mathbb{E}\left[\exp\left(\overline{\lambda}\xi_t\right)\right]}.$$

Here, we have

$$\mathbb{E}\left[\exp\left(\overline{\lambda}\sum_{t=1}^T\xi_t\right)\right]=\mathbb{E}\left[\prod_{t=1}^T\mathbb{E}\left[\exp\left(\overline{\lambda}\xi_t\right)|\mathcal{F}_{t-1}\right]\right].$$

Then, by using Lemma E.1, for each $t=1,\dots,T$,

$$\begin{aligned}\mathbb{E}\left[\exp\left(\overline{\lambda}\xi_t\right)|\mathcal{F}_{t-1}\right]&\leq 1+\frac{\overline{\lambda}^2}{2}\mathbb{E}\left[\xi_t^2|\mathcal{F}_{t-1}\right]+\sum_{k=3}^\infty\frac{\overline{\lambda}^k\mathbb{E}\left[\xi_t^k|\mathcal{F}_{t-1}\right]}{k!}\\ &\leq 1+\frac{\overline{\lambda}^2}{2}\mathbb{E}\left[\xi_t^2|\mathcal{F}_{t-1}\right]+\sum_{k=3}^\infty\overline{\lambda}^kC_1(C_0\sqrt{T})^{-k}\\ &\leq 1+\frac{\overline{\lambda}^2}{2}\mathbb{E}\left[\xi_t^2|\mathcal{F}_{t-1}\right]+O\left(\overline{\lambda}^3/T^{3/2}\right).\end{aligned}$$

Therefore,

$$\begin{aligned}\mathbb{E}\left[\exp\left(\overline{\lambda}\sum_{t=1}^T\xi_t\right)\right]&\leq \mathbb{E}\left[\prod_{t=1}^T\left(1+\frac{\overline{\lambda}^2}{2}\mathbb{E}\left[\xi_t^2|\mathcal{F}_{t-1}\right]+O\left(\overline{\lambda}^3/T^{3/2}\right)\right)\right]\\ &\leq \mathbb{E}\left[\prod_{t=1}^T\exp\left(\frac{\overline{\lambda}^2}{2}\mathbb{E}\left[\xi_t^2|\mathcal{F}_{t-1}\right]+O\left(\overline{\lambda}^3/T^{3/2}\right)\right)\right].\end{aligned}$$

Similarly, by using Lemma E.1 and constants $c, \tilde{c} > 0$, we have

$$
\begin{aligned}
&\mathbb{E}\left[\exp\left(\overline{\lambda}\xi_t\right)\right]\\
&= \exp\left(\log \mathbb{E}\left[\exp\left(\overline{\lambda}\xi_t\right)\right]\right)\\
&= \exp\left(\log\left(1 + \sum_{k=2}^{\infty}\mathbb{E}\left[\frac{(\overline{\lambda}\xi_t)^k}{k!}\right]\right)\right)\\
&= \exp\left(\frac{\overline{\lambda}^2}{2}\mathbb{E}\left[\xi_t^2\right] + \sum_{k=3}^{\infty}\mathbb{E}\left[\frac{(\overline{\lambda}\xi_t)^k}{k!}\right] - \frac{1}{2}\left(\sum_{k=2}^{\infty}\mathbb{E}\left[\frac{(\overline{\lambda}\xi_t)^k}{k!}\right]\right)^2 + \frac{1}{3}\left(\sum_{k=2}^{\infty}\mathbb{E}\left[\frac{(\overline{\lambda}\xi_t)^k}{k!}\right]\right)^3 + \cdots\right)\\
&\overset{(a)}{\geq} \exp\left(\frac{\overline{\lambda}^2}{2}\mathbb{E}\left[\xi_t^2\right] - \sum_{k=3}^{\infty}\mathbb{E}\left[\frac{|\overline{\lambda}\xi_t|^k}{k!}\right] - \frac{1}{2}\left(\sum_{k=2}^{\infty}\mathbb{E}\left[\frac{|\overline{\lambda}\xi_t|^k}{k!}\right]\right)^2 - \frac{1}{3}\left(\sum_{k=2}^{\infty}\mathbb{E}\left[\frac{|\overline{\lambda}\xi_t|^k}{k!}\right]\right)^3 + \cdots\right)\\
&\overset{(b)}{\geq} \exp\left(\frac{\overline{\lambda}^2}{2}\mathbb{E}\left[\xi_t^2\right] - c\overline{\lambda}^3/T^{3/2} - \frac{1}{2}\left(\frac{4C_1\overline{\lambda}^2}{3C_0^2T}\right)^2 - \frac{1}{3}\left(\frac{4C_1\overline{\lambda}^2}{3C_0^2T}\right)^3 - \frac{1}{4}\left(\frac{4C_1\overline{\lambda}^2}{3C_0^2T}\right)^4 - \cdots\right)\\
&\geq \exp\left(\frac{\overline{\lambda}^2}{2}\mathbb{E}\left[\xi_t^2\right] - c\overline{\lambda}^3/T^{3/2} - \left(\frac{4C_1\overline{\lambda}^2}{3C_0^2T}\right)^2 - \left(\frac{4C_1\overline{\lambda}^2}{3C_0^2T}\right)^3 - \left(\frac{4C_1\overline{\lambda}^2}{3C_0^2T}\right)^4 - \cdots\right)\\
&\geq \exp\left(\frac{\overline{\lambda}^2}{2}\mathbb{E}\left[\xi_t^2\right] - c\overline{\lambda}^3/T^{3/2} - \left(\frac{4C_1\overline{\lambda}^2}{3C_0^2T}\right)^2\frac{1}{1-\frac{1}{2}}\right)\\
&\overset{(c)}{\geq} \exp\left(\frac{\overline{\lambda}^2}{2}\mathbb{E}\left[\xi_t^2\right] - c\left(\overline{\lambda}^3/\sqrt{T}\right)^3 - \tilde{c}\overline{\lambda}^4/T^2\right).
\end{aligned}
$$

For (a), we used Jensen's inequality for $m = 2, 3, \dots$ as

$$
-(-1)^m\frac{1}{m}\left(\sum_{k=2}^{\infty}\mathbb{E}\left[\frac{(\overline{\lambda}\xi_t)^k}{k!}\right]\right)^m \geq -\frac{1}{m}\left(\sum_{k=2}^{\infty}\left|\mathbb{E}\left[\frac{(\overline{\lambda}\xi_t)^k}{k!}\right]\right|\right)^m \geq -\frac{1}{m}\left(\sum_{k=2}^{\infty}\mathbb{E}\left[\frac{|\overline{\lambda}\xi_t|^k}{k!}\right]\right)^m.
$$

For (b), we used the fact there exist a constant $c > 0$ such that

$$
\begin{aligned}
\mathbb{E}\left[\sum_{k=2}^{\infty}\frac{|\overline{\lambda}\xi_t|^k}{k!}\right] &\overset{(c)}{\leq} \sum_{k=2}^{\infty}\frac{\overline{\lambda}^k}{k!}\cdot k!C_1\frac{1}{(C_0\sqrt{T})^k}\\
&= C_1\sum_{k=2}^{\infty}\left(\frac{\overline{\lambda}}{C_0\sqrt{T}}\right)^k\\
&= \frac{C_1\overline{\lambda}^2}{C_0^2T}\frac{1}{1-\frac{\overline{\lambda}}{C_0\sqrt{T}}}\\
&\overset{(d)}{\leq} \frac{C_1\overline{\lambda}^2}{C_0^2T}\frac{1}{1-\frac{1}{4}}\\
&= \frac{4C_1\overline{\lambda}^2}{3C_0^2T}\\
&\overset{(d)}{\leq} \frac{1}{2},
\end{aligned}
$$

and

$$
\begin{aligned}
\mathbb{E}\left[\sum_{k=3}^{\infty}\frac{|\overline{\lambda}\xi_t|^k}{k!}\right] &\overset{(c)}{\leq} \sum_{k=3}^{\infty}\frac{\overline{\lambda}^k}{k!}\cdot k!C_1\frac{1}{(C_0\sqrt{T})^k}\\
&\leq c\left(\frac{\overline{\lambda}}{\sqrt{T}}\right)^3,
\end{aligned}
$$

for (c), we used Lemma E.1, and for (d), we used (11).

Then, by combining the above upper and lower bounds, with some constant $\tilde{C}_0, \tilde{C}_1 > 0$,

$$\begin{aligned}\frac{\mathbb{E}\left[\exp\left(\overline{\lambda}\sum_{t=1}^T \xi_t\right)\right]}{\prod_{t=1}^T \mathbb{E}\left[\exp\left(\overline{\lambda}\xi_t\right)\right]} &\leq \frac{\mathbb{E}\left[\prod_{t=1}^T \exp\left(\frac{\overline{\lambda}^2}{2}\mathbb{E}\left[\xi_t^2|\mathcal{F}_{t-1}\right] + O\left(\left(\overline{\lambda}/\sqrt{T}\right)^3\right)\right)\right]}{\prod_{t=1}^T \exp\left(\frac{\overline{\lambda}^2}{2}\mathbb{E}\left[\xi_t^2\right] - c\left(\overline{\lambda}^3/\sqrt{T}\right)^3 - \tilde{c}\overline{\lambda}^4/T^2\right)}\\ &= \exp\left(\tilde{C}_0\overline{\lambda}^4/T + \tilde{C}_1\overline{\lambda}^3/\sqrt{T}\right)\mathbb{E}\left[\prod_{t=1}^T \exp\left(\overline{\lambda}^2\left(\mathbb{E}[\xi_t^2|\mathcal{F}_{t-1}] - \mathbb{E}[\xi_t^2]\right)/2\right)\right].\end{aligned}$$

Using Hölder's inequality,

$$\begin{aligned}\frac{\mathbb{E}\left[\exp\left(\overline{\lambda}\sum_{t=1}^T \xi_t\right)\right]}{\prod_{t=1}^T \mathbb{E}\left[\exp\left(\overline{\lambda}\xi_t\right)\right]} &\leq \exp\left(\tilde{C}_0\overline{\lambda}^4/T + \tilde{C}_1\overline{\lambda}^3/\sqrt{T}\right)\mathbb{E}\left[\prod_{t=1}^T \exp\left(\overline{\lambda}^2\left(\mathbb{E}[\xi_t^2|\mathcal{F}_{t-1}] - \mathbb{E}[\xi_t^2]\right)/2\right)\right]\\ &\leq \exp\left(\tilde{C}_0\overline{\lambda}^4/T + \tilde{C}_1\overline{\lambda}^3/\sqrt{T}\right)\prod_{t=1}^T\left(\mathbb{E}\left[\exp\left(T\overline{\lambda}^2\left(\mathbb{E}[\xi_t^2|\mathcal{F}_{t-1}] - \mathbb{E}[\xi_t^2]\right)/2\right)\right]\right)^{\frac{1}{T}}\\ &\leq \exp\left(\tilde{C}_0\overline{\lambda}^4/T + \tilde{C}_1\overline{\lambda}^3/\sqrt{T}\right)\prod_{t=1}^T\left(\mathbb{E}\left[\exp\left(\overline{\lambda}^2\left(\mathbb{E}[\xi_t^2|\mathcal{F}_{t-1}] - \mathbb{E}[\xi_t^2]\right)\right)\right]\right)^{\frac{1}{2}}, \qquad (16)\end{aligned}$$

where the last inequality is from Jensen's inequality. Note that the term

$$\overline{\lambda}^2\left(\mathbb{E}[\xi_t^2|\mathcal{F}_{t-1}] - \mathbb{E}[\xi_t^2]\right) = \frac{\overline{\lambda}^2}{T}\left(\mathbb{E}\left[\left(\hat{X}_{1,t} - \hat{X}_{2,t} - \Delta\right)^2\middle|\mathcal{F}_{t-1}\right] - \mathbb{E}\left[\left(\hat{X}_{1,t} - \hat{X}_{2,t} - \Delta\right)^2\right]\right)$$

is bounded by some constant because $w_{a,t}$ and $\hat{\mu}_{a,t}$ are bounded and $\overline{\lambda} \leq \sqrt{T}\min\left\{\frac{1}{4}C_0, \sqrt{\frac{3C_0^2}{8C_1}}\right\}$. We have,

$$\begin{aligned}&\left|\mathbb{E}\left[\left(\hat{X}_{1,t} - \hat{X}_{2,t} - \Delta\right)^2\middle|\mathcal{F}_{t-1}\right] - \mathbb{E}\left[\left(\hat{X}_{1,t} - \hat{X}_{2,t} - \Delta\right)^2\right]\right|\\ &\leq \left|\mathbb{E}\left[\left(\hat{X}_{1,t} - \hat{X}_{2,t} - \Delta\right)^2\middle|\mathcal{F}_{t-1}\right] - (\sigma_1+\sigma_2)^2\right| + \left|\mathbb{E}\left[\left(\hat{X}_{1,t} - \hat{X}_{2,t} - \Delta\right)^2\right] - (\sigma_1+\sigma_2)^2\right|\\ &= \left|\mathbb{E}\left[\left(\hat{X}_{1,t} - \hat{X}_{2,t} - \Delta\right)^2\middle|\mathcal{F}_{t-1}\right] - (\sigma_1+\sigma_2)^2\right| + \left|\mathbb{E}\left[\mathbb{E}\left[\left(\hat{X}_{1,t} - \hat{X}_{2,t} - \Delta\right)^2\middle|\mathcal{F}_{t-1}\right] - (\sigma_1+\sigma_2)^2\right]\right|\\ &= o_p(1),\end{aligned}$$

where the last equality is from Lemma D.3 and the asymptotic notation $o_p(\cdot)$ is as $t \to \infty$. This implies

$$\exp\left(\overline{\lambda}^2\left(\mathbb{E}[\xi_t^2|\mathcal{F}_{t-1}] - \mathbb{E}[\xi_t^2]\right)\right) = o_p(1).$$

By Proposition A.2 (a), the sequence

$$\left\{\exp\left(\overline{\lambda}^2\left(\mathbb{E}[\xi_t^2|\mathcal{F}_{t-1}] - \mathbb{E}[\xi_t^2]\right)\right)\right\}_{t\geq 1}$$

is uniformly integrable. From L^r-convergence theorem (Proposition A.3),

$$\mathbb{E}\left[\exp\left(\overline{\lambda}^2\left(\mathbb{E}[\xi_t^2|\mathcal{F}_{t-1}] - \mathbb{E}[\xi_t^2]\right)\right)\right] = o(1).$$

Therefore,

$$\prod_{t=1}^T\left(\mathbb{E}\left[\exp\left(\overline{\lambda}^2\left(\mathbb{E}[\xi_t^2|\mathcal{F}_{t-1}] - \mathbb{E}[\xi_t^2]\right)\right)\right]\right)^{\frac{1}{2}} = o(1). \qquad (17)$$

Using (16) and (17), we get, with some constants $\tilde{C}_2, \tilde{C}_3 > 0$,

$$\frac{\mathbb{E}\left[\exp\left(\overline{\lambda}\sum_{t=1}^{T}\xi_t\right)\right]}{\prod_{t=1}^{T}\mathbb{E}\left[\exp\left(\overline{\lambda}\xi_t\right)\right]} \leq \exp\left(\tilde{C}_2\overline{\lambda}^4/T + \tilde{C}_3\overline{\lambda}^3/\sqrt{T}\right). \tag{18}$$

In summary, by (15) and (18), for all $1 \leq u \leq \sqrt{T}\min\left\{\frac{1}{4}C_0, \sqrt{\frac{3C_0^2}{8C_1}}\right\}$,

$$\frac{\mathbb{P}\left(Z_T > u\right)}{1 - \Phi(\overline{\lambda})} \leq \tilde{C}\overline{\lambda}\exp\left(\tilde{C}_2\overline{\lambda}^4/T + \tilde{C}_3\overline{\lambda}^3/\sqrt{T} + C\left(\overline{\lambda}^3/\sqrt{T} + \overline{\lambda}^2 V_T\right)\right). \tag{19}$$

Next, we compare $1 - \Phi(\overline{\lambda})$ with $1 - \Phi(u)$. Recall the following upper bound and lower bound on $1 - \Phi(x) = \Phi(-x)$ in (8):

$$\frac{1}{\sqrt{2\pi}(1+x)}\exp\left(-\frac{x^2}{2}\right) \leq \Phi(-x) \leq \frac{1}{\sqrt{\pi}(1+x)}\exp\left(-\frac{x^2}{2}\right),\ x \geq 0.$$

For all $1 \leq u \leq \sqrt{T}\min\left\{\frac{1}{4}C_0, \sqrt{\frac{3C_0^2}{8C_1}}\right\}$,

$$\begin{aligned}
1 &\leq \frac{\int_{\overline{\lambda}}^{\infty}\exp(-t^2/2)\mathrm{d}t}{\int_{u}^{\infty}\exp(-t^2/2)\mathrm{d}t} \\
&\leq \frac{\frac{1}{\sqrt{\pi}(1+\overline{\lambda})}\exp(-\overline{\lambda}^2/2)}{\frac{1}{\sqrt{2\pi}(1+u)}\exp(-u^2/2)} \\
&= \sqrt{2}\frac{1+u}{1+\overline{\lambda}}\exp((u^2 - \overline{\lambda}^2)/2).
\end{aligned}$$

From (12), we have

$$\begin{aligned}
u^2 - \overline{\lambda}^2 &= (u + \overline{\lambda})(u - \overline{\lambda}) \\
&\leq 2u(C\theta(uV_T + u^2/\sqrt{T})) \\
&= 2C\theta(u^2V_T + u^3/\sqrt{T}).
\end{aligned}$$

Therefore, with some constant $\tilde{C}_4 > 0$

$$\frac{\int_{\overline{\lambda}}^{\infty}\exp(-t^2/2)\mathrm{d}t}{\int_{u}^{\infty}\exp(-t^2/2)\mathrm{d}t} \leq \exp\left(\tilde{C}_4\left(u^2V_T + u^3/\sqrt{T}\right)\right).$$

We find that

$$1 - \Phi(\overline{\lambda}) \leq \left(1 - \Phi(u)\right)\exp\left(\tilde{C}_4\left(u^2V_T + u^3/\sqrt{T}\right)\right). \tag{20}$$

By combining (19), (20), and (11), for all $1 \leq u \leq \sqrt{T}\min\left\{\frac{1}{4}C_0, \sqrt{\frac{3C_0^2}{8C_1}}\right\}$, there exists a constant $\tilde{C}_5 > 0$ such that

$$\begin{aligned}
\frac{\mathbb{P}\left(Z_T > u\right)}{1 - \Phi(u)} &\leq \tilde{C}\overline{\lambda}\exp\left(C\left(\overline{\lambda}^3/\sqrt{T} + \overline{\lambda}^2V_T\right) + \tilde{C}_2\overline{\lambda}^4/T + \tilde{C}_3\overline{\lambda}^3/\sqrt{T} + \tilde{C}_4\left(u^2V_T + u^3/\sqrt{T}\right)\right) \\
&\leq \tilde{C}u\exp\left(\tilde{C}_5\left(u^2V_T + u^3/\sqrt{T} + u^4/T\right)\right).
\end{aligned}$$

Applying the same argument to the martingale $-Z_T$, we conclude the proof.

□