

STABLE BATCHED BANDIT: OPTIMAL REGRET WITH FREE INFERENCE

Anonymous authors

Paper under double-blind review

ABSTRACT

In this paper, we discuss statistical inference when using a sequential strategy to collect data. While inferential tasks become challenging with sequentially collected data, we argue that this problem can be alleviated when the sequential algorithm satisfies certain stability properties; we call such algorithms stable bandit algorithms. Focusing on batched bandit problems, we first demonstrate that popular algorithms including the greedy-UCB algorithm and ϵ -greedy ETC algorithms are not stable, complicating downstream inferential tasks. Our main result shows that a form of elimination algorithm is stable in the batched bandit setup, and we characterize the asymptotic distribution of the sample means. This result allows us to construct asymptotically exact confidence intervals for arm-means which are sharper than existing concentration-based bounds. As a byproduct of our main results, we propose an Explore and Commit (ETC) strategy, which is stable — thus allowing easy statistical inference— and also attains optimal regret up to a factor of 4.

Our work connects two historically conflicting paradigms in sequential learning environments: regret minimization and statistical inference. Ultimately, we demonstrate that it is possible to minimize regret without sacrificing the ease of performing statistical inference, bridging the gap between these two important aspects of sequential decision-making.

1 INTRODUCTION

Reinforcement learning (RL) has emerged as a pivotal paradigm in artificial intelligence, driving significant advancements across diverse domains. Its impact spans from theoretical computer science to practical applications in robotics, control systems, and beyond. At the core of RL lies the fundamental challenge of balancing exploration and exploitation - a dilemma that encapsulates the agent’s need to gather new information about its environment while simultaneously leveraging existing knowledge to maximize rewards. This balance is crucial for developing effective decision-making strategies through environmental interaction, positioning RL as a cornerstone technology in the evolution of autonomous systems.

In many real-world applications of reinforcement learning, data is collected sequentially and often in batches, reflecting practical constraints and operational realities. This batched approach to data collection is particularly prevalent in domains such as online education Kizilcec et al. (2020), mobile health interventions Liao et al. (2020); Klasnja et al. (2019); Yom-Tov et al. (2017), and digital marketing Li et al. (2010), where multiple users interact with systems simultaneously. While traditional RL algorithms excel at optimizing performance within a specific problem instance, there is a growing need for methods that can extract generalizable insights from the collected data. Statistical inference on sequentially collected data becomes crucial when the goal extends beyond mere performance optimization to include scientific discovery and informed decision-making for future implementations. Consider a mobile app designed to improve dental hygiene habits Trella et al. (2024); Nahum-Shani et al. (2024). The app uses RL to personalize reminders and brushing technique tips. Beyond maximizing daily app engagement, researchers and dentists would be interested in understanding which interventions most effectively promote long-term oral health improvements. They might want to determine if gamified brushing sessions are more impactful than educational content, or if the frequency of reminders significantly affects adherence to recommended brushing

054 duration. This knowledge could guide the development of future dental health interventions, allow
055 for refinement of less effective strategies, and contribute to our understanding of habit formation in
056 oral care.

057 In this paper, we focus on the problem of statistical inference in bandits problems with data collected
058 in batches; colloquially known as batched bandit problems. While bandit strategies focus on mini-
059 mizing regret, the sequential (non-iid) nature of bandit algorithms make the down-stream statistical
060 inference much more challenging. For instance, sample means maybe biased for bandit data Nie
061 et al. (2018), and the sample means may not be asymptotically normal Zhang et al. (2020); Ying
062 et al. (2024). In the following section, we provide a brief survey of batched bandit algorithms, with
063 a special focus on explore and commit (ETC) strategies, and on statistical inference with the data
064 collected from a sequential procedure, akin to a bandit algorithm.

066 1.1 RELATED WORK

068 1.1.1 BATCHED BANDITS AND EXPLORE-THEN-COMMIT ALGORITHMS

069 The study of batched bandits has gained significant attention in recent years, with a focus on algo-
070 rithms that balance exploration and exploitation in a limited number of interaction rounds. Explore-
071 Then-Commit (ETC) algorithms represent a special case of batched bandits where the learning pro-
072 cess is divided into two distinct phases: an exploration phase followed by a commitment phase. See
073 the work of Robbins (1952); Anscombe (1963). Perchet et al. (2016) proposed a general strategy
074 for constructing batched bandit algorithms, including ETC-type approaches. Their work addressed
075 the crucial aspect of batch size selection, which may vary across batches to obtain minimax regret
076 bounds. Building on this foundation, Gao et al. (2019) investigated whether adaptively chosen batch
077 sizes could further reduce regret in batched settings. Exploring different aspects of batched bandits,
078 Jin et al. (2021) examined a scenario with a random horizon, ensuring asymptotically optimal regret
079 for exponential families as reward distributions. This work highlighted the flexibility of batched ap-
080 proaches in handling uncertain time horizons. The algorithm that we study in this paper is motivated
081 from the work of Auer & Ortner (2010), where the authors discussed an elimination-based algorithm
082 for batched bandits.

084 1.1.2 STATISTICAL INFERENCE WITH BANDIT DATA

085 The challenge of performing valid statistical inference with sequentially collected data, particularly
086 in batched bandit settings, has become an important area of research. Zhang et al. (2020) demon-
087 strated that the average reward obtained from batched bandit algorithms is not necessarily asymp-
088 totically normal, and proposed a batched OLS estimator for inference in non-stationary settings. To
089 address these challenges, researchers have developed two main approaches: non-asymptotic meth-
090 ods based on concentration bounds for self-normalized martingales Abbasi-Yadkori et al. (2011),
091 and asymptotic methods exploiting the martingale nature of the data and debiasing techniques. See
092 the works in Hall & Heyde (2014); Zhang & Zhang (2014); Khamaru et al. (2021); Ying et al. (2024);
093 Lin et al. (2023); Bibaut et al. (2021); Hadad et al. (2021); Zhang et al. (2021); Abbasi-Yadkori et al.
094 (2011) and references therein.

096 1.2 CONTRIBUTIONS

097 Our approach to inference in the bandit problem is significantly different from existing approaches.
098 As we already pointed out in the previous related work section, most of the inference methods are
099 *post-processing* methods; meaning they utilize very little information of the bandit algorithm itself,
100 and rely on the Martingale structure present in the sequentially collected data. While this approach is
101 more flexible, the worst-case guarantees for such methods can be pessimistic; see the paper Khamaru
102 et al. (2021); Lattimore (2023) for worst-case lower bounds.

103 In contrast, we discuss classes of algorithms, which we call *stable bandit algorithms*, where no such
104 post-processing is needed, and classical statistical methods — which are used for iid data — can be
105 used. At a very high level,

106 *We can treat bandit data as iid data (asymptotically) when the bandit algorithm is stable.*
107

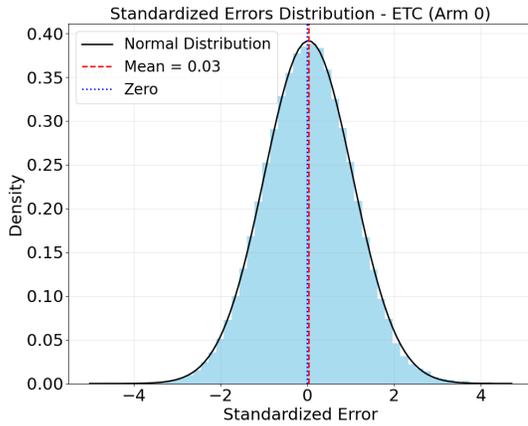


Figure 1: Stable ETC arm 1

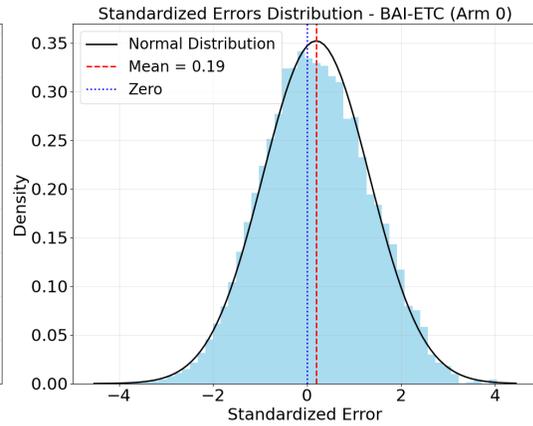


Figure 2: BAI ETC arm 1

Figure 3: Comparison of error distributions for stable-ETC Algorithm 2 and BAI ETC algorithm for a two-armed-bandit with Gaussian rewards and $\mu_1 = \mu_2 = 1$. We see that the asymptotic distribution of the arm-means are close to Gaussian when stable-ETC — a *stable algorithm*— is used. But, the distributions of arm means are not Gaussian when BAI-ETC algorithm — which provides optimal regret — is used; the mean of standardized noise are significantly positive, close to 0.20. We also show (in Corollary 2) that the regret of stable-ETC is no more than 4-times the the optimal-regret. The simulation results are average of 5000 repetitions and the horizon is set to $T = 1000$. See Appendix B for a detailed simulation.

The notion of stable bandit algorithm is motivated from the seminal work of Lai & Wei (1982). To the best of our knowledge, this work is the first to show stable bandit algorithms in batched settings.

Our main contributions are as follows:

- First, we introduce a class of bandit algorithms for multi-armed bandits, which we call *stable bandit algorithms*, and argue that the sample means for each arm are asymptotically normal, when the bandit data is collected using a stable bandit algorithm.
- In Section 3.2 we focus on 2-batch algorithms. We demonstrate that the vanilla ϵ -greedy explore-then-commit (ETC) algorithm is not stable, and we propose a modification of the explore-then-commit algorithm which is stable. An interesting result in this section is a stable ETC algorithm whose regret is optimal up to a factor 4.
- In Section 3.3 we focus on B -batch algorithms. In Algorithm 3, we propose a B -batch algorithm, in Theorem 2 we discuss the stability property of this algorithm, and characterize the asymptotic distribution of the sample-means.

2 PROBLEM SET UP

In this paper we focus on multiarmed bandit algorithm where the data is collected in multiple batches. For sake of exposition, we discuss the two-armed case in full details, though many of our results extend to the K -armed setting. At each round $1 \leq t \leq T$, we select an arm $A_t \in \{1, 2\}$ and receive a reward $Y_t \in \mathbb{R}$ from the distribution \mathcal{P}_{A_t} . We assume

- Let μ_a and σ_a^2 , respectively, denote the mean and variance of the distribution \mathcal{P}_a . We assume that \mathcal{P}_a is a sub-Gaussian random variable with sub-Gaussian parameter λ_a . The parameters $(\mu_a, \sigma_a^2, \lambda_a)$ are unknown and without loss of generality we assume that $\lambda_a \leq 1$ for $a = 1, 2$.

The focus of this paper is to understand bandit algorithms where the data is collected in batches. We consider two types of batched bandit algorithms:

- 162
163
164
165
166
167
168
169
170
1. **Two batch algorithm:** In Section 3.2 we focus on a two batch algorithm where the number of arms within each batch goes to ∞ . See Algorithms 1 and 2 for more details. The algorithm discussed in this section are motivated from Explore Then Commit (ETC) strategies Robbins (1952); Anscombe (1963), and draws inspiration from the ETC type algorithm discussed in Auer & Ortner (2010).
 2. **B -batch algorithm:** In Section 3, we focus on algorithms where the data is collected in B batches. The number of rounds in each batch, which we denote by $2m$ remains fixed, and we let the number of batches B to ∞ . We detail our B -batch procedure in Algorithm 3.

171 **Goal:** The goal in both cases is to understand the asymptotic properties of the samples means for both arms defined as

172
173
174
175

$$\bar{\mu}_{a,T} = \frac{1}{n_{a,T}} \cdot \sum_{t=1}^T Y_t \cdot \mathbf{1}_{A_t=a} \quad \text{where} \quad n_{a,T} = \sum_{t=1}^T \mathbf{1}_{A_t=a}.$$

176 We are interested to understand the asymptotic behavior of the sample means $(\bar{\mu}_{1,T}, \bar{\mu}_{2,T})$. This, for example, will allow us to construct confidence intervals of (μ_1, μ_2) .

179 3 MAIN RESULTS

180
181 Before moving onto the details of the algorithm we introduce a class of bandit algorithms which we call *stable bandit algorithms*. Our first result, stated in Lemma 1, proves that stable algorithms ensures that the sample means $(\bar{\mu}_{1,T}, \bar{\mu}_{2,T})$ are asymptotically normal.

185 3.1 STABLE BANDIT ALGORITHMS

186
187 Throughout, we use \mathcal{M}_T to denote a generic bandit algorithm with horizon T . Let $n_{a,t}(\mathcal{M}_T)$ denote the number of arm pulls of arm a in t rounds. We say an algorithm \mathcal{M}_T is *stable* if for arms $a \in \{1, 2\}$ there exists *non-random* scalars $n_a^*(\mathcal{M}_T)$ satisfying

188
189
190
191 (stability:)
$$\frac{n_{a,T}(\mathcal{M}_T)}{n_a^*(\mathcal{M}_T)} \xrightarrow{p} 1 \quad \text{for some} \quad n_a^*(\mathcal{M}_T) \rightarrow \infty \quad \text{as} \quad T \rightarrow \infty. \quad (1)$$

192 Here, the constants $\{n_a^*(\mathcal{M}_T)\}_{a=1,2}$ above may depend on the parameters associated to reward distributions $\mathcal{P}_1, \mathcal{P}_2$ or other tuning parameters that are independent of the data collected using algorithm \mathcal{M}_T . Throughout, we hide the dependence of the algorithm \mathcal{M}_T in $n_{a,T}$ and n_a^* for notational simplicity. Let us first prove a simple yet useful Lemma for stable algorithms:

193
194
195
196 **Lemma 1** *If an algorithm \mathcal{M}_T is stable and the third moment of the arm-reward distribution \mathcal{P}_a is bounded. Then for all arms $a \in \{1, 2\}$ the sample means are asymptotically normal. Concretely,*

197
198
199
200
$$\sqrt{n_{a,T}} \cdot (\bar{\mu}_{a,T} - \mu_a) \xrightarrow{p} \mathcal{N}(0, \sigma_a^2) \quad (2)$$

201 **Proof of Lemma 1** Fix an arm a . Define the partial sum:

202
203
$$S_{a,t} = \sum_{\ell \leq t} (Y_\ell - \mu_a) \cdot \mathbf{1}_{\{A_\ell=a\}}$$

204 By construction, $S_{a,t}$ is a sum of Martingale difference sequence. Additionally, using the notation $\mathcal{F}_t := \sigma\{(Y_\ell, A_\ell)_{\ell \leq t}\}$ for the σ -field generated by data-set obtained up to stage t , we have

205
206
207
208
$$\sum_{1 \leq t \leq T} \text{Var} \left(\frac{1}{\sigma_a \cdot \sqrt{n_a^*}} \cdot (Y_t - \mu_a) \cdot \mathbf{1}_{\{A_t=a\}} \mid \mathcal{F}_{t-1} \right) = \frac{n_{a,T}}{n_a^*} \xrightarrow{p} 1.$$

209 In words, the sum of the conditional variances of the Martingale difference array stabilizes. Combining this with the assumption $n_a^* \rightarrow \infty$ and using the fact that the third moment of the reward distribution is bounded (recall that rewards are sub-Gaussian) we see that the Lindeberg conditions of the Martingale Central Limit Theorem Dvoretzky (1972) are satisfied. Thus, applying the Martingale CLT from Dvoretzky (1972) we conclude

210
211
212
213
214
215
$$\sqrt{n_{a,T}} \cdot (\bar{\mu}_{a,T} - \mu_a) \xrightarrow{p} \mathcal{N}(0, \sigma_a^2) \quad (3)$$

This completes the proof of Lemma 1.

Confidence interval for μ_a : Of course, we can estimate the reward variance by the sample variance estimate, and utilize Lemma 1 to construct confidence intervals for the unknown sample means μ_1 and μ_2 . For instance, for any consistent estimator of $\hat{\sigma}_{a,T}$ of σ_a , and given any target $\alpha \in (0, 1)$ using Slutsky’s theorem we conclude that

$$\lim_{T \rightarrow \infty} \mathbb{P} \left(\left[\bar{\mu}_{a,T} - \hat{\sigma}_{a,T} \cdot \frac{z_{\alpha/2}}{\sqrt{n_{a,T}}}, \bar{\mu}_{a,T} + \hat{\sigma}_{a,T} \cdot \frac{z_{\alpha/2}}{\sqrt{n_{a,T}}} \right] \ni \mu_a \right) = 1 - \alpha.$$

See, the comments after Theorem 1. for a discussion on consistent estimator of σ_a .

3.2 INFERENCE IN 2-BATCH BANDITS: EXPLORE THEN COMMIT (ETC) STRATEGIES

In this section, we focus on explore then commit-type strategies. Before doing so, we argue that many naive and intuitive algorithms are not stable.

3.2.1 INSTABILITY OF VANILLA-ETC STRATEGY:

Arguably, the most naive and intuitive strategy is the explore then commit strategy which uses sample mean to decide which arm to commit to. Concretely, consider an ϵ -greedy ETC algorithm where we

- Pull both arms with probability $1/2$ for a total of $2m$ times in the first batch.
- Define

$$(\epsilon\text{-greedy}) \quad \hat{a}_{\max} = \arg \max_{a \in \{1,2\}} \bar{\mu}_{a,2m} \quad (4)$$

- For the remaining $T - 2m$ rounds, pull the arm \hat{a}_{\max} with higher sample mean with probability $1 - \epsilon$, and the arm with lower mean with probability $\epsilon > 0$.

Let us discuss the stability property of the above algorithm. For simplicity, let us assume the reward distributions are Gaussian; i.e., $\mathcal{P}_1 \equiv \mathcal{N}(\mu_1, 1)$, $\mathcal{P}_2 \equiv \mathcal{N}(\mu_2, 1)$, and $m = \lceil T/4 \rceil$. In the case when the margin $\Delta = \mu_1 - \mu_2 = 0$, by symmetry we have

$$\mathbb{P}(\bar{\mu}_{1,2m} > \bar{\mu}_{2,2m}) = \mathbb{P}(\bar{\mu}_{2,2m} > \bar{\mu}_{1,2m}) = \frac{1}{2}$$

Thus for both arms $a \in \{1, 2\}$ we have, as $T \rightarrow \infty$

$$\frac{n_{a,T}}{T} \xrightarrow{p} \begin{cases} \frac{1}{4} + \frac{\epsilon}{2} & \text{with probability } \frac{1}{2} \\ \frac{3}{4} - \frac{\epsilon}{2} & \text{with probability } \frac{1}{2} \end{cases}$$

Stated differently, the ϵ -greedy ETC algorithm from equation 4 is not stable when $\Delta = 0$. Invoking (Zhang et al., 2020, Theorem 6) we have the following lemma:

Lemma 2 (Zhang et al., 2020, Theorem 6) *Suppose the data is collected using the ETC algorithm from equation 4. Then the sample mean for arm 1 is not asymptotically normal when $\Delta = \mu_1 - \mu_2 = 0$. In particular,*

$$\sqrt{n_{1,T}} \cdot (\bar{\mu}_{1,T} - \mu_1) \xrightarrow{d} Y \quad (5)$$

where $Y = \sqrt{\frac{1}{3-\epsilon}} (Z_1 - \sqrt{2-\epsilon}Z_3) \mathbb{I}_{Z_1 > Z_2} + \sqrt{\frac{1}{1+\epsilon}} (Z_1 - \sqrt{\epsilon}Z_3) \mathbb{I}_{Z_1 < Z_2}$, and Z_1, Z_2, Z_3 are iid standard Gaussian random variables.

This instability property of the ϵ -greedy ETC algorithm also extends to other natural algorithms like ϵ -greedy upper confidence bound (UCB) algorithm, and the non-normality of the sample means phenomenon still persists. See Appendix C of the paper Zhang et al. (2020) for more details.

3.2.2 A STABLE ETC-STRATEGY:

We are now ready to discuss a modification of the ϵ -greedy ETC (displayed in equation 4) which is stable. The algorithm proceeds in two stages:

- In the first stage, we select both arms m times.
- At the end of first stage, we collect arms with high rewards and create an active set $\mathcal{A} \subseteq \{1, 2\}$. In the second stage, we select all the arms in the active set \mathcal{A} equally often.

The details of this two-batch method is detailed in Algorithm 1. We point out the strategy in Algorithm 1 draws motivation from elimination types algorithms studies Auer & Ortner (2010). We are now ready to analyze the stability of Algorithm 1.

Algorithm 1 An Explore then Commit strategy

Inputs: Pair of integers (T, m) with $1 \leq m \leq T/2$

Batch 1

Pull both arm m times, and construct the active set after the total $2m$ arm-pulls

$$\mathcal{A} := \left\{ a \mid \bar{\mu}_{a,2m} + \sqrt{\frac{2 \log T}{n_{a,2m}}} \geq \max \left\{ \bar{\mu}_{1,2m} - \sqrt{\frac{2 \log T}{n_{1,2m}}}, \bar{\mu}_{2,2m} - \sqrt{\frac{2 \log T}{n_{2,2m}}} \right\} \right\} \quad (6)$$

Batch 2:

if $T - 2m \geq 1$ **then**

If the set \mathcal{A} is singleton, pull the arm in \mathcal{A} remaining $T - 2m$ times, or pull both arms with probability $1/2$, a total of $T - 2m$ times.

end if

Condition on m : Let, Θ denote the collection of all problem dependent parameters. In Theorem 1, we allow any sequence of $m \equiv m(T, \Theta)$ that satisfies the following property:

$$\frac{m \cdot \Delta^2}{8 \log T} \rightarrow \beta \quad \text{for some } 0 \leq \beta \leq \infty \quad \text{as } T \rightarrow \infty. \quad (7)$$

Here, the condition for $\beta = \infty$ means $\frac{m \cdot \Delta^2}{8 \log T} \rightarrow \infty$. The condition above, for instance allows for $m = T^\alpha$ for some $1 > \alpha > 0$, $m = \frac{2 \log T}{\Delta^2}$, or any constant value of m . Additionally, the condition 7 is always satisfied with $\beta = 0$ when $\Delta = 0$ for any value of m . Condition 7 rules out choices of m — changing with T — for which the ratio in 7 oscillates. The condition equation 7 allows for most choices of m that are used in practice, especially when Δ is kept fixed as the number of T increases. As we discuss later, the above condition also allows us to analyze the case when Δ is allowed to scale with the sample size T .

Theorem 1 *Suppose m satisfies condition equation 7 for some $0 \leq \beta \leq \infty$, and $T - 2m \rightarrow \infty$. Then Algorithm 1 is stable with the following choices of n_1^*, n_2^**

$$\text{If } \beta \leq 1, \quad n_1^* = n_2^* = \frac{T}{2} \quad (8)$$

$$\text{If } \beta > 1, \quad \text{then } n_1^* = T - \frac{8\beta \log T}{\Delta^2} \quad \text{and} \quad n_2^* = \frac{8\beta \log T}{\Delta^2} \quad (9)$$

Consequently, for both arms

$$\hat{\sigma}_{a,T} \cdot \sqrt{n_{a,T}} \cdot (\bar{\mu}_{a,T} - \mu_a) \xrightarrow{d} \mathcal{N}(0, 1).$$

Here, $\bar{\mu}_{a,T}$ denotes the sample mean, and $\hat{\sigma}_{a,T}$ is any consistent estimator of variance σ_a .

See Section A.2 for a proof of this theorem ¹. A few comments regarding Theorem 1 are in order.

¹When $\beta = \infty$, we replace $\frac{8 \log T}{\Delta^2}$ by m in equation 9.

Estimating variance and statistical inference: It turns out that under the assumptions of Theorem 1 sample variance estimate $\hat{\sigma}_{a,T}$ is a consistent estimator of σ_a . Here,

$$\hat{\sigma}_{a,T} = \frac{1}{n_{a,T}} \sum_{t=1}^T (Y_t - \bar{\mu}_{a,T})^2 \cdot \mathbf{1}_{A_t=a} \quad (10)$$

See Corollary 1 in the paper Khamaru & Zhang (2024) for a proof of consistency for $\hat{\sigma}_{a,T}$. One can now easily create asymptotically exact $1 - \alpha$ confidence interval. In particular, given any $1 > \alpha > 0$ define the confidence interval $\mathcal{C}_{a,\alpha}$

$$\mathcal{C}_{a,\alpha} = \left[\bar{\mu}_{a,T} - \hat{\sigma}_{a,T} \cdot \frac{z_{1-\alpha/2}}{\sqrt{n_{a,T}}}, \bar{\mu}_{a,T} + \hat{\sigma}_{a,T} \cdot \frac{z_{1-\alpha/2}}{\sqrt{n_{a,T}}} \right] \quad (11)$$

where $z_{1-\alpha/2}$ is the $(1 - \alpha/2)^{th}$ quantile of the standard Gaussian distribution. Then, we have that for both arms $a \in \{1, 2\}$ $\lim_{T \rightarrow \infty} \mathbb{P}(\mu_a \in \mathcal{C}_{a,\alpha}) = 1 - \alpha$.

Stability its and connections to Law of Iterated Logarithm: It is interesting understand whether the bonus factor $\sqrt{2 \log \log T}$ plays any special role in stability, and whether we can replace it some other bonus factor. A careful look at the proof (see Section A.2) reveals that one can replace $\sqrt{2 \log \log T}$ in the equation 6 by any other bonus-term q_T satisfying

$$\frac{\sqrt{2 \log \log T}}{q_T} \rightarrow 0 \quad \text{as } T \rightarrow \infty. \quad (12)$$

The term $\sqrt{2 \log \log T}$ above comes from the Law of Iterated Logarithm (LIL). In other words, the stability of Algorithm 1 is guaranteed as long as the bonus factor is q_T *over-powers* the fluctuations in the sample means — which is governed by the Law of Iterated Logarithm. In such case, modifying the argument of Theorem 1 we obtain the following corollary:

Corollary 1 *Suppose condition 12 holds, and $\frac{m\Delta^2}{4q_T} \rightarrow \beta$ for some $0 \leq \beta \leq \infty$. Then Algorithm 1 with bonus-term q_T in place of $\sqrt{2 \log \log T}$ is stable. We have*

$$\begin{aligned} n_1^* = n_2^* &= \frac{T}{2} \quad \text{If } \beta \leq 1, \quad \text{and} \\ n_1^* &= T - \frac{4\beta q_T^2}{\Delta^2} \quad \text{and} \quad n_2^* = \frac{4\beta q_T^2}{\Delta^2} \quad \text{If } \beta > 1. \end{aligned}$$

3.2.3 DATA DEPENDENT STOPPING: OPTIMAL-REGRET WITH FREE INFERENCE

In Theorem 1 we assume that the choice of m in Algorithm 1 is a pre-determined input to the algorithm, and it *does not depend* on the data collected by Algorithm 1. In this section, we analyze a two-stage algorithm where m is dependent on the data, more formally a stopping time.

Corollary 2 *Algorithm 2 is stable with*

$$\frac{n_{1,T}}{T - 4 \log T / \Delta^2} \xrightarrow{p} 1 \quad \text{and} \quad \frac{n_{2,T}}{4 \log T / \Delta^2} \xrightarrow{p} 1$$

Additionally, assuming $T\Delta^2 \geq 4e^2$, the regret \mathbb{R}_T of Algorithm can be upper bounded by

$$\mathbb{R}_T \leq \frac{16 \log T}{\Delta} + \frac{120e \sqrt{\log(\Delta^2 T / 4)} + 64e + 32}{\Delta} + 2\Delta$$

See Section A.3.1 for a proof of this corollary.

Comparison with lower bound: It is interesting to compare the regret of the Algorithm 2 with a lower bound for explore and commit-type algorithms. Following the work of (Garivier et al., 2016, Theorem 4) we have that for any uniformly efficient ETC strategies Lai & Robbins (1985); Garivier et al. (2016) we have that

$$\text{(Lower bound) : } \quad \liminf_{T \rightarrow \infty} \frac{\mathbb{R}_T}{\log T} \geq \frac{4}{\Delta}. \quad (13)$$

Algorithm 2 ETC with stopping time

Inputs: Integer $T \geq 2$
 Set $A_1 = 1, A_2 = 2$ and set $t = 2$
while $|\bar{\mu}_{1,t} - \bar{\mu}_{2,t}| \leq \sqrt{\frac{4 \log T}{(t/2)}}$ **do**
 Use $A_{t+1} = 1$ and $A_{t+2} = 2$, and set $t = t + 2$
end while

$$\mathcal{A} := \left\{ a \mid \bar{\mu}_{a,t} + \sqrt{\frac{\log T}{(t/2)}} \geq \max \left\{ \bar{\mu}_{1,t} - \sqrt{\frac{\log T}{(t/2)}}, \bar{\mu}_{2,t} - \sqrt{\frac{\log T}{(t/2)}} \right\} \right\}$$

if $T - t \geq 1$ **then**

 If the set \mathcal{A} is singleton, pull the arm in \mathcal{A} remaining $T - t$ times, or pull both arms with probability $1/2$, a total of $T - t$ times.

end if

See Section 3 of work by Garivier et al. (2016) for more discussion on the lower bound. It is now interesting to understand the asymptotic behavior of the Algorithm 2. Assuming Δ is bounded by a constant, simple algebra yields

$$\limsup_{T \rightarrow \infty} \frac{\mathbb{R}_T}{\log T} \leq \frac{16}{\Delta}$$

Stated differently, Algorithm 2 ensures accurate asymptotic inference while matching the minimax-optimal regret up to a factor of 4.

3.3 INFERENCE IN B-BATCHED BANDITS

In this section, we focus on a batched bandit algorithm with B batches. In each batch $1 \leq b \leq B$, we perform arm pulls a total of $2m$ times. Throughout this section, we assume m is fixed, and we let the number of $B \rightarrow \infty$. We give details about our algorithm in Algorithm 3. Akin to the last section, we are interested in the stability of Algorithm 3.

We point-out that unlike Section 3.2.2 the number of arm pulls within each batch is *fixed*, i.e., not data-dependent.

Algorithm 3 B -batch algorithm

Input: Pair of integer (m, B) with $m, B \geq 1$.
 Set $T = 2mB$, $\mathcal{A}_1 = \{1, 2\}$ and pull both arms m times.
for $b = 1$ **to** $B - 1$ **do**
 Construct the active set

$$\mathcal{A}_{b+1} := \left\{ a \mid \bar{\mu}_{a,2mb} + \sqrt{\frac{2 \log T}{n_{a,2mb}}} \geq \max \left\{ \bar{\mu}_{1,2mb} - \sqrt{\frac{2 \log T}{n_{1,2mb}}}, \bar{\mu}_{2,2mb} - \sqrt{\frac{2 \log T}{n_{2,2mb}}} \right\} \right\}$$

 If the set \mathcal{A}_{b+1} is singleton, pull the arm in \mathcal{A}_{b+1} $2m$ times, or pull both arms m times.

end for

Theorem 2 Suppose $B \rightarrow \infty$, then Algorithm 3 is stable with

$$\begin{aligned} n_1^* &= \frac{T}{2} \cdot \mathbf{1}_{\{\Delta=0\}} + \left(T - \frac{8 \log T}{\Delta^2} \right) \cdot \mathbf{1}_{\{\Delta>0\}} \quad \text{and} \\ n_2^* &= \frac{T}{2} \cdot \mathbf{1}_{\{\Delta=0\}} + \frac{8 \log T}{\Delta^2} \cdot \mathbf{1}_{\{\Delta>0\}}. \end{aligned} \quad (14)$$

Consequently, for each arm $a \in \{1, 2\}$

$$\hat{\sigma}_{a,T} \cdot \sqrt{n_{a,T}} \cdot (\bar{\mu}_{a,T} - \mu_a) \xrightarrow{d} \mathcal{N}(0, 1).$$

Here, $\bar{\mu}_{a,T}$ denotes the sample mean, and $\hat{\sigma}_{a,T}$ is any consistent estimator of variance σ_a .

See Section A.3 for a proof of this theorem. Just like our previous section, the results in Theorem 2 can be generalized to a general bonus term q_T satisfying

$$\frac{\sqrt{2 \log \log T}}{q_T} \rightarrow 0 \quad \text{as } T \rightarrow \infty.$$

In particular, for any q_T , the expression of n_a^* in equation 14 changes to

$$\begin{aligned} n_1^* &= \frac{T}{2} \cdot \mathbf{1}_{\{\Delta=0\}} + \left(T - \frac{4q_T^2}{\Delta^2}\right) \cdot \mathbf{1}_{\{\Delta>0\}} \quad \text{and} \\ n_2^* &= \frac{T}{2} \cdot \mathbf{1}_{\{\Delta=0\}} + \frac{4q_T^2}{\Delta^2} \cdot \mathbf{1}_{\{\Delta>0\}}. \end{aligned}$$

4 PROOFS

In this section, we prove our main Theorem 1, in part. Complete proof of all the results are deferred to the Appendix.

Define

$$g_T = \sqrt{2 \log T} \quad \text{and} \quad h_T := \sqrt{7 \log \log(4T) + 3 \log 2} \quad (15)$$

$$\mathcal{E}_T := \left\{ |\bar{\mu}_{1,n_{1,t}} - \mu_1| \leq \lambda_1 \frac{h_T}{\sqrt{n_{1,t}}} \quad \text{and} \quad |\bar{\mu}_{2,n_{2,t}} - \mu_2| \leq \lambda_2 \frac{h_T}{\sqrt{n_{2,t}}} \quad \forall t \in [T] \right\} \quad (16)$$

The proof utilizes the following lemma from (Khamaru & Zhang, 2024, Lemma 5.1). See also the work by Balsubramani (2014).

Lemma 3 *Let X_1, X_2, \dots be i.i.d. λ_a -sub-Gaussian random variable with zero mean. Then the sample-mean $\bar{X}_t := (X_1 + \dots + X_t)/t$ satisfies the following bound*

$$\mathbb{P} \left(\exists t \geq 1 : |\bar{X}_t| \geq \lambda_a \sqrt{\frac{9}{4t} \cdot \log \frac{(\log_2 4t)^2}{\delta}} \right) \leq 2\delta$$

By assumptions the arm-means are sub-Gaussian with sub-Gaussian parameter bounded by 1. Thus, substituting $\delta = 1/\log(4T)$ in Lemma 3 and taking a union bound over both arms we obtain

$$\mathbb{P}(\mathcal{E}_T) = \mathbb{P}(|\bar{\mu}_{a,t} - \mu_a| \leq h_T, \text{ for all } 1 \leq t \leq T, 1 \leq a \leq 2) \geq 1 - \frac{6}{\log(4T)} \quad (17)$$

4.1 PARTIAL PROOF OF THEOREM 1

Let us define two indicator variables

$$I_1 := \mathbf{1}_{\{1 \in \mathcal{A}\}} \quad \text{and} \quad I_2 := \mathbf{1}_{\{2 \in \mathcal{A}\}}.$$

From Algorithm 1 we have,

$$n_{1,T} = \begin{cases} T - m & \text{if } I_2 = 0, \\ m & \text{if } I_1 = 0, \\ m + \sum_{i=1}^{T-2m} V_i & \text{if } I_1 = I_2 = 1. \end{cases} \quad (18)$$

where $V_i \sim \text{Bern}(0, \frac{1}{2})$ for $1 \leq i \leq T - 2m$

The proof follows by analyzing the random variable $n_{1,T}$ under the high probability event \mathcal{E}_T . Using the bound equation 16, we get that $\mathbb{P}(\mathcal{E}_T) \geq 1 - \frac{6}{\log 4T}$. Thus, it suffices to study the behavior of $n_{1,T}$ on the high-probability event \mathcal{E}_T .

486 CASE 1: $\beta \leq 1$:

487 We have that for large T s

$$\begin{aligned}
488 \{I_2 = 0\} \cap \mathcal{E}_T &= \left\{ \bar{\mu}_{2,2m} + \frac{g_T}{\sqrt{n_{2,2m}}} < \bar{\mu}_{1,2m} - \frac{g_T}{\sqrt{n_{1,2m}}} \right\} \cap \mathcal{E}_T \\
489 &\stackrel{(i)}{\subseteq} \left\{ \mu_2 - \lambda_2 \frac{h_T}{\sqrt{n_{2,2m}}} + \frac{g_T}{\sqrt{n_{2,2m}}} < \mu_1 + \lambda_1 \frac{h_T}{\sqrt{n_{1,2m}}} - \frac{g_T}{\sqrt{n_{1,2m}}} \right\} \\
490 &= \left\{ \mu_2 - \lambda_2 \frac{h_T}{\sqrt{m}} + \frac{g_T}{\sqrt{m}} < \mu_1 + \lambda_1 \frac{h_T}{\sqrt{m}} - \frac{g_T}{\sqrt{m}} \right\} \\
491 &= \left\{ \frac{1}{\sqrt{m}}(-(\lambda_1 + \lambda_2)h_T + 2g_T) < \Delta \right\}
\end{aligned}$$

492 The step (i) uses the property of the event \mathcal{E}_T . Now note that the set in the last line is empty when
493 $\Delta = 0$ for large T ; this is because $\frac{g_T}{h_T} \rightarrow 0$ as $T \rightarrow \infty$. When $\Delta > 0$ with $\beta \leq 1$, we have using
494 condition equation 7 and $h_T/g_T \rightarrow 0$ we have

$$\frac{1}{\Delta\sqrt{m}}(-(\lambda_1 + \lambda_2)h_T + 2g_T) \rightarrow 1/\sqrt{\beta} \geq 1.$$

495 where we have used the notation $1/0 \equiv \infty$ for $\beta = 0$, and we have $\mathbb{P}(\mathcal{E}_T \cap \{I_2 = 0\}) \rightarrow 0$ when
496 $0 \leq \beta \leq 1$. Since $\mu_1 \geq \mu_2$ by assumption, it is immediate to verify that $\mathbb{P}(\mathcal{E}_T \cap \{I_1 = 0\}) \rightarrow 0$.
497 Thus we have

$$\mathbb{P}(\{I_1 = 1\} \cap \{I_2 = 1\} \cap \mathcal{E}_T) \rightarrow 1, \tag{19}$$

498 When $I_1 = I_2 = 1$, we have $n_{1,T} = m + \sum_{i \leq T-2m} V_i$, and we have

$$\frac{m + \sum_{i \leq T-2m} V_i}{T/2} = 1 + \frac{\sum_{i \leq T-2m} (V_i - \frac{1}{2})}{\frac{T}{2}} \xrightarrow{p} 1$$

499 where the last deduction above uses $T - 2m \rightarrow \infty$ and the weak law of large numbers. Using a
500 similar argument for $n_{1,T}$, and putting together the pieces we conclude

$$\frac{n_{1,T}}{T/2} \xrightarrow{p} 1 \quad \text{and} \quad \frac{n_{2,T}}{T/2} \xrightarrow{p} 1.$$

501 The proof of the case $\beta > 1$ is similar, and the details are moved to Appendix .

522 DISCUSSION

523 In this paper, we discussed the problem of statistical inference when data is collected using a batched
524 bandit algorithm. We introduced the concept of stable bandit algorithms, which allows for straight-
525 forward statistical inference even when the dataset is not i.i.d. For instance, the sample arm means
526 are asymptotically normal when data is collected using a stable bandit algorithm. We also argue that
527 such stable algorithms do not sacrifice regret and are optimal up to a constant factor in certain cases.
528 Our work bridges the gap between regret minimization and statistical inference, two historically
529 conflicting paradigms in sequential learning environments.

530 While we focused on two-armed bandit problems in this paper, several interesting questions remain.
531 For instance, it would be interesting to extend our results to the K -armed case. In our B batched
532 Algorithm 3, the number of arm-pulls (m) in each batch is kept fixed. It would be interesting to
533 understand the stability properties of our algorithm when the number of arm-pulls are allowed to
534 grow with T .

537 REFERENCES

538 Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic
539 bandits. *Advances in neural information processing systems*, 24, 2011.

- 540 FJ Anscombe. Sequential medical trials. *Journal of the American Statistical Association*, 58(302):
541 365–383, 1963.
- 542 Peter Auer and Ronald Ortner. Ucb revisited: Improved regret bounds for the stochastic multi-armed
543 bandit problem. *Periodica Mathematica Hungarica*, 61(1-2):55–65, 2010.
- 544 Akshay Balsubramani. Sharp finite-time iterated-logarithm martingale concentration. *arXiv preprint*
545 *arXiv:1405.2639*, 2014.
- 546 Aurélien Bibaut, Maria Dimakopoulou, Nathan Kallus, Antoine Chambaz, and Mark van Der Laan.
547 Post-contextual-bandit inference. *Advances in neural information processing systems*, 34:28548–
548 28559, 2021.
- 549 Aryeh Dvoretzky. Asymptotic normality for sums of dependent random variables. In *Proceedings of*
550 *the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability*
551 *Theory*, volume 6, pp. 513–536. University of California Press, 1972.
- 552 Zijun Gao, Yanjun Han, Zhimei Ren, and Zhengqing Zhou. Batched multi-armed bandits problem,
553 2019. URL <https://arxiv.org/abs/1904.01763>.
- 554 Aurélien Garivier, Tor Lattimore, and Emilie Kaufmann. On explore-then-commit strategies. *Ad-*
555 *vances in Neural Information Processing Systems*, 29, 2016.
- 556 Vitor Hadad, David A Hirshberg, Ruohan Zhan, Stefan Wager, and Susan Athey. Confidence in-
557 tervals for policy evaluation in adaptive experiments. *Proceedings of the national academy of*
558 *sciences*, 118(15):e2014602118, 2021.
- 559 Peter Hall and Christopher C Heyde. *Martingale limit theory and its application*. Academic press,
560 2014.
- 561 Tianyuan Jin, Jing Tang, Pan Xu, Keke Huang, Xiaokui Xiao, and Quanquan Gu. Almost op-
562 timal anytime algorithm for batched multi-armed bandits. In Marina Meila and Tong Zhang
563 (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of
564 *Proceedings of Machine Learning Research*, pp. 5065–5073. PMLR, 18–24 Jul 2021. URL
565 <https://proceedings.mlr.press/v139/jin21c.html>.
- 566 Koulik Khamaru and Cun-Hui Zhang. Inference with ucb. *arXiv*, 2024.
- 567 Koulik Khamaru, Yash Deshpande, Tor Lattimore, Lester Mackey, and Martin J Wainwright. Near-
568 optimal inference in adaptive linear regression. *arXiv preprint arXiv:2107.02266*, 2021.
- 569 René F Kizilcec, Justin Reich, Michael Yeomans, Christoph Dann, Emma Brunskill, Glenn Lopez,
570 Selen Turkay, Joseph Jay Williams, and Dustin Tingley. Scaling up behavioral science interven-
571 tions in online education. *Proceedings of the National Academy of Sciences*, 117(26):14900–
572 14905, 2020.
- 573 Predrag Klasnja, Shawna Smith, Nicholas J Seewald, Andy Lee, Kelly Hall, Brook Luers, Eric B
574 Hekler, and Susan A Murphy. Efficacy of contextually tailored suggestions for physical activity:
575 a micro-randomized optimization trial of heartsteps. *Annals of Behavioral Medicine*, 53(6):573–
576 582, 2019.
- 577 Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances*
578 *in applied mathematics*, 6(1):4–22, 1985.
- 579 Tze Leung Lai and Ching Zong Wei. Least squares estimates in stochastic regression models with
580 applications to identification and control of dynamic systems. *The Annals of Statistics*, 10(1):
581 154–166, 1982.
- 582 Tor Lattimore. A lower bound for linear and kernel regression with adaptive covariates. In *The*
583 *Thirty Sixth Annual Conference on Learning Theory*, pp. 2095–2113. PMLR, 2023.
- 584 Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to
585 personalized news article recommendation. In *Proceedings of the 19th international conference*
586 *on World wide web*, pp. 661–670, 2010.

- 594 Peng Liao, Kristjan Greenewald, Predrag Klasnja, and Susan Murphy. Personalized heartsteps: A
595 reinforcement learning algorithm for optimizing physical activity. *Proceedings of the ACM on*
596 *Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(1):1–22, 2020.
- 597 Licong Lin, Koulik Khamaru, and Martin J Wainwright. Semi-parametric inference based on adap-
598 tively collected data. *arXiv preprint arXiv:2303.02534*, 2023.
- 600 Inbal Nahum-Shani, Zara M Greer, Anna L Trella, Kelly W Zhang, Stephanie M Carpenter, Dennis
601 Ruenger, David Elashoff, Susan A Murphy, and Vivek Shetty. Optimizing an adaptive digital
602 oral health intervention for promoting oral self-care behaviors: Micro-randomized trial protocol.
603 *Contemporary clinical trials*, 139:107464, 2024.
- 604 Xinkun Nie, Xiaoying Tian, Jonathan Taylor, and James Zou. Why adaptively collected data have
605 negative bias and how to correct for it. In *International Conference on Artificial Intelligence and*
606 *Statistics*, pp. 1261–1269. PMLR, 2018.
- 607 Vianney Perchet, Philippe Rigollet, Sylvain Chassang, and Erik Snowberg. Batched bandit prob-
608 lems. *The Annals of Statistics*, 44(2), April 2016. ISSN 0090-5364. doi: 10.1214/15-aos1381.
609 URL <http://dx.doi.org/10.1214/15-AOS1381>.
- 611 Herbert Robbins. Some aspects of the sequential design of experiments. 1952.
- 612 Anna L Trella, Kelly W Zhang, Stephanie M Carpenter, David Elashoff, Zara M Greer, Inbal
613 Nahum-Shani, Dennis Ruenger, Vivek Shetty, and Susan A Murphy. Oralytics reinforcement
614 learning algorithm. *arXiv preprint arXiv:2406.13127*, 2024.
- 615 Mufang Ying, Koulik Khamaru, and Cun-Hui Zhang. Adaptive linear estimating equations. *Ad-*
616 *vances in Neural Information Processing Systems*, 36, 2024.
- 617 Elad Yom-Tov, Guy Feraru, Mark Kozdoba, Shie Mannor, Moshe Tennenholtz, and Irit Hochberg.
618 Encouraging physical activity in patients with diabetes: intervention using a reinforcement learn-
619 ing system. *Journal of medical Internet research*, 19(10):e338, 2017.
- 620 Cun-Hui Zhang and Stephanie S Zhang. Confidence intervals for low dimensional parameters in
621 high dimensional linear models. *Journal of the Royal Statistical Society Series B: Statistical*
622 *Methodology*, 76(1):217–242, 2014.
- 623 Kelly Zhang, Lucas Janson, and Susan Murphy. Inference for batched bandits. *Advances in neural*
624 *information processing systems*, 33:9818–9829, 2020.
- 625 Kelly Zhang, Lucas Janson, and Susan Murphy. Statistical inference with m-estimators on adaptively
626 collected data. *Advances in neural information processing systems*, 34:7460–7471, 2021.
- 627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647

A PROOFS

In this section, we prove Theorems and Corollaries from the main text.

A.1 PROOF OF THEOREM 1

Here, we provide the complete proof for Theorem 1. For the sake of readability we repeat some parts of the proof from the main section of the paper.

Define

$$g_T = \sqrt{2 \log T} \quad \text{and} \quad h_T := \sqrt{7 \log \log(4T) + 3 \log 2} \quad (20)$$

$$\mathcal{E}_T := \left\{ |\bar{\mu}_{1,n_{1,t}} - \mu_1| \leq \lambda_1 \frac{h_T}{\sqrt{n_{1,t}}} \quad \text{and} \quad |\bar{\mu}_{2,n_{2,t}} - \mu_2| \leq \lambda_2 \frac{h_T}{\sqrt{n_{2,t}}} \quad \forall t \in [T] \right\} \quad (21)$$

The proof utilizes the following lemma from (Khamaru & Zhang, 2024, Lemma 5.1). See also the work by Balsubramani (2014).

Lemma 4 *Let X_1, X_2, \dots be i.i.d. λ_a -sub-Gaussian random variable with zero mean. Then the sample-mean $\bar{X}_t := (X_2 + \dots + X_t)/t$ satisfies the following bound*

$$\mathbb{P} \left(\exists t \geq 1 : |\bar{X}_t| \geq \lambda_a \sqrt{\frac{9}{4t} \cdot \log \frac{(\log_2 4t)^2}{\delta}} \right) \leq 2\delta$$

By assumptions the arm-means are sub-Gaussian with sub-Gaussian parameter bounded by 1. Thus, substituting $\delta = 1/\log(4T)$ in Lemma 4 and taking a union bound over both arms we obtain

$$\mathbb{P}(\mathcal{E}_T) = \mathbb{P}(|\bar{\mu}_{a,t} - \mu_a| \leq h_T, \text{ for all } 1 \leq t \leq T, 1 \leq a \leq 2) \geq 1 - \frac{6}{\log(4T)} \quad (22)$$

A.2 PROOF OF THEOREM 1

Let us define two indicator variables

$$I_1 := \mathbf{1}_{\{1 \in \mathcal{A}\}} \quad \text{and} \quad I_2 := \mathbf{1}_{\{2 \in \mathcal{A}\}}.$$

From Algorithm 1 we have,

$$n_{1,T} = \begin{cases} T - m & \text{if } I_2 = 0, \\ m & \text{if } I_1 = 0, \\ m + \sum_{i=1}^{T-2m} V_i & \text{if } I_1 = I_2 = 1. \end{cases} \quad (23)$$

where $V_i \sim \text{Bern}(0, \frac{1}{2})$ for $1 \leq i \leq T - 2m$

The proof follows by analyzing the random variable $n_{1,T}$ under the high probability event \mathcal{E}_T . Using the bound equation 16, we get that $\mathbb{P}(\mathcal{E}_T) \geq 1 - \frac{6}{\log 4T}$. Thus, it suffices to study the behavior of $n_{1,T}$ on the high-probability event \mathcal{E}_T .

CASE 1: $\beta \leq 1$:

We have that for large T s

$$\begin{aligned} \{I_2 = 0\} \cap \mathcal{E}_T &= \left\{ \bar{\mu}_{2,2m} + \frac{g_T}{\sqrt{n_{2,2m}}} < \bar{\mu}_{1,2m} - \frac{g_T}{\sqrt{n_{1,2m}}} \right\} \cap \mathcal{E}_T \\ &\stackrel{(i)}{\subseteq} \left\{ \mu_2 - \lambda_2 \frac{h_T}{\sqrt{n_{2,2m}}} + \frac{g_T}{\sqrt{n_{2,2m}}} < \mu_1 + \lambda_1 \frac{h_T}{\sqrt{n_{1,2m}}} - \frac{g_T}{\sqrt{n_{1,2m}}} \right\} \\ &= \left\{ \mu_2 - \lambda_2 \frac{h_T}{\sqrt{m}} + \frac{g_T}{\sqrt{m}} < \mu_1 + \lambda_1 \frac{h_T}{\sqrt{m}} - \frac{g_T}{\sqrt{m}} \right\} \\ &= \left\{ \frac{1}{\sqrt{m}} (-(\lambda_1 + \lambda_2)h_T + 2g_T) < \Delta \right\} \end{aligned}$$

The step (i) uses the property of the event \mathcal{E}_T . Now note that the set in the last line is empty when $\Delta = 0$ for large T ; this is because $\frac{g_T}{h_T} \rightarrow 0$ as $T \rightarrow \infty$. When $\Delta > 0$ with $\beta \leq 1$, we have using condition equation 7 and $h_T/g_T \rightarrow 0$ we have

$$\frac{1}{\Delta\sqrt{m}}(-(\lambda_1 + \lambda_2)h_T + 2g_T) \rightarrow 1/\sqrt{\beta} \geq 1.$$

where we have used the notation $1/0 \equiv \infty$ for $\beta = 0$, and we have $\mathbb{P}(\mathcal{E}_T \cap \{I_2 = 0\}) \rightarrow 0$ when $0 \leq \beta \leq 1$. Since $\mu_1 \geq \mu_2$ by assumption, it is immediate to verify that $\mathbb{P}(\mathcal{E}_T \cap \{I_1 = 0\}) \rightarrow 0$. Thus we have

$$\mathbb{P}(\{I_1 = 1\} \cap \{I_2 = 1\} \cap \mathcal{E}_T) \rightarrow 1, \quad (24)$$

When $I_1 = I_2 = 1$, we have $n_{1,T} = m + \sum_{i \leq T-2m} V_i$, and we have

$$\frac{m + \sum_{i \leq T-2m} V_i}{T/2} = 1 + \frac{\sum_{i \leq T-2m} (V_i - \frac{1}{2})}{\frac{T}{2}} \xrightarrow{p} 1$$

where the last deduction above uses $T - 2m \rightarrow \infty$ and the weak law of large numbers. Using a similar argument for $n_{1,T}$, and putting together the pieces we conclude

$$\frac{n_{1,T}}{T/2} \xrightarrow{p} 1 \quad \text{and} \quad \frac{n_{2,T}}{T/2} \xrightarrow{p} 1.$$

CASE 2: $\beta > 1$:

First note that $\beta > 1$ implies $\Delta > 0$. Next, we have for T large

$$\begin{aligned} \{I_2 = 1\} \cap \mathcal{E}_T &= \left\{ \bar{\mu}_{2,2m} + \frac{g_T}{\sqrt{n_{2,2m}}} \geq \bar{\mu}_{1,2m} - \frac{g_T}{\sqrt{n_{1,2m}}} \right\} \cap \mathcal{E}_T \\ &\stackrel{(i)}{\subseteq} \left\{ \mu_2 + \lambda_2 \frac{h_T}{\sqrt{n_{2,2m}}} + \frac{g_T}{\sqrt{n_{2,2m}}} \geq \mu_1 - \lambda_1 \frac{h_T}{\sqrt{n_{1,2m}}} - \frac{g_T}{\sqrt{n_{1,2m}}} \right\} \\ &= \left\{ \mu_2 + \lambda_2 \frac{h_T}{\sqrt{m}} + \frac{g_T}{\sqrt{m}} \geq \mu_1 - \lambda_1 \frac{h_T}{\sqrt{m}} - \frac{g_T}{\sqrt{m}} \right\} \\ &= \left\{ \frac{1}{\Delta\sqrt{m}}((\lambda_1 + \lambda_2)h_T + 2g_T) \geq 1 \right\} \end{aligned}$$

Now using $\beta > 1$ and noting $h_T/g_T \rightarrow 0$ we have that

$$\lim_{T \rightarrow \infty} \frac{1}{\Delta\sqrt{m}}((\lambda_1 + \lambda_2)h_T + 2g_T) \rightarrow 1/\sqrt{\beta} < 1.$$

Thus we conclude $\mathbb{P}(\{I_2 = 1\} \cap \mathcal{E}_T) \rightarrow 0$. Again, noting $\mu_1 \geq \mu_2$ it is immediate to see that $\mathbb{P}(\{I_2 = 1\} \cap \mathcal{E}_T) \rightarrow 1$. Taking the intersection of the the last two events we conclude

$$\mathbb{P}(\mathcal{E}_T \cap \{I_1 = 1\} \cap \{I_2 = 0\}) \rightarrow 1.$$

On the event $\mathcal{E}_T \cap \{I_1 = 1\} \cap \{I_2 = 0\}$ we have $n_{1,T} = T - m$ and $n_{2,T} = m$, and using $\frac{m\Delta^2}{8 \log T} \rightarrow \beta$ we conclude

$$\frac{n_{1,T}}{T - (8\beta \log T)/\Delta^2} \xrightarrow{p} 1 \quad \text{and} \quad \frac{n_{2,T}}{(8\beta \log T)/\Delta^2} \xrightarrow{p} 1.$$

This completes the proof of Theorem 1.

A.3 PROOF OF THEOREM 2

We start by defining the events E_1 and E_2

$$\begin{aligned} E_1 &= \{\exists 1 \leq b \leq B - 1 \text{ such that } 1 \notin \mathcal{A}_b\} \quad \text{and} \\ E_2 &= \{\exists 1 \leq b \leq B - 1 \text{ such that } 2 \notin \mathcal{A}_b\} \end{aligned} \quad (25)$$

In words, the sets E_1 and E_2 , respectively, correspond to the event that arm 1 and 2 is eliminated in one of the first $B - 1$ batches of Algorithm 3; note that there no elimination at the end of B^{th} batch. Throughout the proof, we assume T is large enough such that

$$g_T - (\lambda_1 \vee \lambda_2)h_T > 0 \quad (26)$$

756 CASE 1: $\Delta = 0$:

757 We show that when $\Delta = 0$, none of the arms are eliminated with high probability. Note that

$$\begin{aligned}
758 E_1 \cap \mathcal{E}_T &\subseteq \left\{ \bar{\mu}_{1,2mb} + \frac{g_T}{\sqrt{n_{1,2mb}}} < \bar{\mu}_{2,2mb} - \frac{g_T}{\sqrt{n_{2,2mb}}} \text{ for some } b \in [B-1] \right\} \cap \mathcal{E}_T \\
759 &\stackrel{(i)}{\subseteq} \left\{ \mu_1 - \lambda_1 \frac{h_T}{\sqrt{n_{1,2mb}}} + \frac{g_T}{\sqrt{n_{1,2mb}}} < \mu_2 + \lambda_2 \frac{h_T}{\sqrt{n_{2,2mb}}} - \frac{g_T}{\sqrt{n_{2,2mb}}} \text{ for some } b \in [B-1] \right\} \\
760 &\subseteq \left\{ 0 < - \left(\frac{1}{\sqrt{n_{1,2mb}}} (g_T - \lambda_1 h_T) + \frac{1}{\sqrt{n_{2,2mb}}} (g_T - \lambda_2 h_T) \right) \text{ for some } b \in [B-1] \right\} \stackrel{(ii)}{=} \phi \\
761 & \\
762 & \\
763 & \\
764 & \\
765 & \\
766 & \\
767 & \tag{27}
\end{aligned}$$

768 where, in step (i), we use the description of event \mathcal{E}_T , and the implication (ii) uses that T is large
769 enough such that the property 26 is satisfied. Similarly we get $E_2 \cap \mathcal{E}_T = \phi$ for large T . Additionally,
770 using the Lemma 3 we conclude

$$771 \mathbb{P}(E_1 \cap E_2^C \cap \mathcal{E}_T) \geq 1 - \frac{6}{\log 4T} \tag{28}$$

772 Finally, on the event $E_1 \cap E_2^C \cap \mathcal{E}_T$ we have $n_{1,T} = n_{2,T} = \frac{T}{2}$, and we conclude $\frac{n_{1,T}}{T/2} \xrightarrow{P} 1$ and
773 $\frac{n_{2,T}}{T/2} \xrightarrow{P} 1$.

774 CASE 2: $\Delta > 0$:

775 Following an argument similar to equation 27, we have that when $\mu_1 \geq \mu_2$, the arm is never elimi-
776 nated in B batches with probability at least $1 - \frac{6}{\log T}$. Next we show that arm 2 is eliminated with
777 high probability in one of the B rounds with high probability. If possible, let arm 2 is not eliminated
778 in $B - 1$ rounds. Then, we have $n_{1,T} = n_{2,T} = m(B - 1) = T/2 - m$.

$$\begin{aligned}
779 E_1^C \cap E_2^C \cap \mathcal{E}_T & \\
780 &\stackrel{(i)}{\subseteq} \left\{ \bar{\mu}_{2,2m(B-1)} + \frac{g_T}{\sqrt{n_{2,2m(B-1)}}} \geq \bar{\mu}_{1,2m(B-1)} - \frac{g_T}{\sqrt{n_{1,2m(B-1)}}} \right\} \cap \mathcal{E}_T \\
781 &\stackrel{(ii)}{\subseteq} \left\{ \mu_2 + \lambda_2 \frac{h_T}{\sqrt{m(B-1)}} + \frac{g_T}{\sqrt{m(B-1)}} \geq \mu_1 - \lambda_1 \frac{h_T}{\sqrt{m(B-1)}} - \frac{g_T}{\sqrt{m(B-1)}} \right\} \\
782 &= \left\{ \frac{1}{\sqrt{T/2 - m}} [(\lambda_1 + \lambda_2)h_T + 2g_T] \geq \Delta \right\} \downarrow \phi \\
783 & \\
784 & \\
785 & \\
786 & \\
787 & \\
788 & \\
789 & \\
790 & \\
791 & \\
792 & \\
793 & \\
794 & \tag{29}
\end{aligned}$$

795 Here, step (i) uses the fact that none of the arms are eliminated at the end of batch $B - 1$, step (ii)
796 uses the property of the set \mathcal{E}_T , and the final step above uses the fact that $\frac{\sqrt{T/2 - m}}{[(\lambda_1 + \lambda_2)h_T + 2g_T]} \rightarrow 0$
797 as $T \rightarrow \infty$ (Recall $T=2mB$ and $B \rightarrow \infty$) and $\Delta > 0$. Putting together the pieces we conclude
798 $\mathbb{P}(E_1^C \cap E_2^C \cap \mathcal{E}_T) \rightarrow 0$, meaning, arm 1 is never eliminated and arm 2 is eliminated with high
799 probability.

800 Next, we define b^* as the time arm 2 is eliminated

$$801 b^* = \min \left\{ b : \bar{\mu}_{2,2mb} + \frac{g_T}{\sqrt{mb}} < \bar{\mu}_{1,2mb} - \frac{g_T}{\sqrt{mb}} \right\} \text{ in } E_2 \tag{30}$$

802 We show that b^* is upper and lower bounded respectively, by \bar{b} and \underline{b} .

$$803 \underline{b} := \min \left\{ b : \mu_2 - \lambda_2 \frac{h_T}{\sqrt{mb}} + \frac{g_T}{\sqrt{mb}} < \mu_1 + \lambda_1 \frac{h_T}{\sqrt{mb}} - \frac{g_T}{\sqrt{mb}} \right\} \tag{31a}$$

$$804 \bar{b} := \min \left\{ b : \mu_2 + \lambda_2 \frac{h_T}{\sqrt{mb}} + \frac{g_T}{\sqrt{mb}} < \mu_1 - \lambda_1 \frac{h_T}{\sqrt{mb}} - \frac{g_T}{\sqrt{mb}} \right\} \tag{31b}$$

Simple algebra yields

$$\bar{b} \leq \frac{((\lambda_1 + \lambda_2)h_T + 2g_T)^2}{m\Delta^2} + 1; \quad \underline{b} \geq \frac{(-(\lambda_1 + \lambda_2)h_T + 2g_T)^2}{m\Delta^2} \quad (32)$$

If possible $E_2 \cap \mathcal{E}_T$ we have is true as both the arms are not eliminated throughout. Implication (v) holds due to the concentration in \mathcal{E}_T for large enough T . Implication (vi) is true for large enough B as $\log(T)$ is $o(B-1)$. So in \mathcal{E}_T both Arms cannot remain active throughout the T rounds. The only possibility in \mathcal{E}_T is that arm 2 gets eliminated.

Define,

$$b^* = \min \left\{ b : \bar{\mu}_{2,2mb} + \frac{g_T}{\sqrt{mb}} < \bar{\mu}_{1,2mb} - \frac{g_T}{\sqrt{mb}} \right\} \quad \text{in } E_2 \quad (33)$$

b^* is the last batch where arm 2 is pulled and well defined since arm 2 is eliminated in E_2 .

$$\implies n_{2,T} = n_{2,2mb^*} = mb^* \quad \text{in } E_2 \quad (34)$$

Now we will try to bound b^* from both sides.

Define

$$\underline{b} := \min \left\{ b : \mu_2 - \lambda_2 \frac{h_T}{\sqrt{mb}} + \frac{g_T}{\sqrt{mb}} < \mu_1 + \lambda_1 \frac{h_T}{\sqrt{mb}} - \frac{g_T}{\sqrt{mb}} \right\} \quad (35a)$$

$$\bar{b} := \min \left\{ b : \mu_2 + \lambda_2 \frac{h_T}{\sqrt{mb}} + \frac{g_T}{\sqrt{mb}} < \mu_1 - \lambda_1 \frac{h_T}{\sqrt{mb}} - \frac{g_T}{\sqrt{mb}} \right\} \quad (35b)$$

Observe

$$\bar{b} \leq \frac{((\lambda_1 + \lambda_2)h_T + 2g_T)^2}{m\Delta^2} + 1; \quad \underline{b} \geq \frac{(-(\lambda_1 + \lambda_2)h_T + 2g_T)^2}{m\Delta^2} \quad (36)$$

In $E_{2,T} \cap \mathcal{E}_T$,

$$\begin{aligned} & \bar{\mu}_{2,2mb^*} + \frac{g_T}{\sqrt{mb^*}} < \bar{\mu}_{1,2mb^*} - \frac{g_T}{\sqrt{mb^*}} \\ & \stackrel{(i)}{\implies} \mu_2 - \lambda_2 \frac{h_T}{\sqrt{mb^*}} + \frac{g_T}{\sqrt{mb^*}} < \mu_1 + \lambda_1 \frac{h_T}{\sqrt{mb^*}} - \frac{g_T}{\sqrt{mb^*}} \\ & \implies \underline{b} \leq b^* \end{aligned} \quad (37)$$

Implication (i) is due to the concentration in \mathcal{E}_T . Now we need to show $\bar{b} \geq b^*$ in $E_2 \cap \mathcal{E}_T$. We will show this using contradiction.

In $E_{2,T} \cap \mathcal{E}_T$ if $\bar{b} < b^*$,

$$\implies \bar{\mu}_{2,2m\bar{b}} + \frac{g_T}{\sqrt{m\bar{b}}} \geq \bar{\mu}_{1,2m\bar{b}} - \frac{g_T}{\sqrt{m\bar{b}}} \quad (38)$$

$$\implies \mu_2 + \lambda_2 \frac{h_T}{\sqrt{m\bar{b}}} + \frac{g_T}{\sqrt{m\bar{b}}} \geq \mu_1 - \lambda_1 \frac{h_T}{\sqrt{m\bar{b}}} - \frac{g_T}{\sqrt{m\bar{b}}} \quad (39)$$

which contradicts the definition of \bar{b} .

\therefore In $E_2 \cap \mathcal{E}_T$ we have $\underline{b} \leq b^* \leq \bar{b}$

$$\begin{aligned} & \stackrel{(ii)}{\implies} \frac{(-(\lambda_1 + \lambda_2)h_T + 2g_T)^2}{m\Delta^2} \leq b^* \leq \frac{((\lambda_1 + \lambda_2)h_T + 2g_T)^2}{m\Delta^2} + 1 \\ & \stackrel{(iii)}{\implies} \frac{(-(\lambda_1 + \lambda_2)h_T + 2g_T)^2}{\Delta^2} \leq n_{2,T} \leq \frac{((\lambda_1 + \lambda_2)h_T + 2g_T)^2}{\Delta^2} + m \\ & \implies (-(\lambda_1 + \lambda_2) \frac{h_T}{2g_T} + 1)^2 \leq \frac{n_{2,T}}{\frac{4g_T^2}{\Delta^2}} \leq ((\lambda_1 + \lambda_2) \frac{h_T}{2g_T} + 1)^2 + \frac{m\Delta^2}{4g_T^2} \\ & \implies \frac{n_{2,T}}{\frac{8 \log T}{\Delta^2}} \rightarrow 1 \quad \implies \frac{n_{1,T}}{T - \frac{8 \log T}{\Delta^2}} \rightarrow 1 \end{aligned} \quad (40)$$

864 Implication (ii) follows from 36 and implication (iii) follows from 34.
 865 From 17, 27 and 29 we get $\mathbb{P}(E_2 \cap \mathcal{E}_T) \rightarrow 1$

$$866 \implies \frac{n_{2,T}}{\frac{8 \log T}{\Delta^2}} \xrightarrow{p} 1 \quad \implies \frac{n_{1,T}}{T - \frac{8 \log T}{\Delta^2}} \xrightarrow{p} 1$$

870 A.3.1 PROOF OF COROLLARY 2

871 If $\Delta = \mu_1 - \mu_2 = 0$, using Lemma 3 and equation 17 we deduce that

$$872 \mathbb{P}(|\bar{\mu}_{1,2t} - \bar{\mu}_{2,2t}| \leq \sqrt{\frac{4 \log T}{t}} \text{ for all } 1 \leq t \leq T/2) \geq 1 - \frac{6}{\log 4T}$$

873 Thus, $\mathbb{P}(n_{1,T} = n_{2,T} = T/2) \geq 1 - \frac{6}{\log 4T}$, and algorithm 2 is stable when $\Delta = 0$.

874 Now let $\Delta > 0$. Throughout, we use the shorthand $h_T = \sqrt{7 \log \log(4T)} + 3 \log 2$. We define t_{\min} and t_{\max} as

$$875 t_{\min} = \min \left\{ t \mid \Delta + \frac{h_T}{\sqrt{t}} > \sqrt{\frac{4 \log T}{t}} \right\} \text{ and } t_{\max} = \min \left\{ t \mid \Delta - \frac{h_T}{\sqrt{t}} > \sqrt{\frac{4 \log T}{t}} \right\}$$

876 Let t_Δ be the first time (random) such that $|\bar{\mu}_{1,2t_\Delta} - \bar{\mu}_{2,2t_\Delta}| \geq \sqrt{\frac{4 \log T}{t_\Delta}}$. We claim that

$$877 t_{\min} \leq t_\Delta \leq t_{\max} \quad \text{with probability} \geq 1 - \frac{6}{\log 4T}. \quad (41)$$

$$878 \hat{a} = 1 \quad \text{with probability} \geq 1 - \frac{6}{\log 4T}. \quad (42)$$

879 Let us prove the stability of algorithm 2 using the last two bounds first. Simple algebra yields:

$$880 t_{\min} \geq \frac{(\sqrt{4 \log T} - h_T)^2}{\Delta^2} - 1 \quad \text{and} \quad t_{\max} \leq \frac{(\sqrt{4 \log T} + h_T)^2}{\Delta^2} + 1$$

881 We then have

$$882 \frac{(\sqrt{4 \log T} - h_T)^2}{\Delta^2} - 1 \leq \frac{t_\Delta}{4 \log T / \Delta^2} \leq \frac{(\sqrt{4 \log T} + h_T)^2}{\Delta^2} + 1 \quad \text{with prob.} \geq 1 - \frac{6}{\log 4T}.$$

883 The upper and lower bound above converges to 1 as $T \rightarrow \infty$ and with $\Delta > 0$ fixed. Finally, using $\hat{a} = 1$ with high probability from equation 42 we have that $n_{2,T} = t_\Delta$ and $n_{1,T} = T - t_\Delta$ with probability at least $1 - \frac{12}{\log T}$. Putting together the pieces we conclude

$$884 \frac{n_{2,T}}{4 \log T / \Delta^2} \xrightarrow{p} 1 \quad \text{and} \quad \frac{n_{1,T}}{T - 4 \log T / \Delta^2} \xrightarrow{p} 1.$$

885 It remains to prove the claims 41 and equation 42.

886 **Proof of claim 41:** The proof of of this part is similar to that of Theorem 2; see the arguments from equation 35a- equation 39. The main idea is that on the event \mathcal{E}_T , defined in equation 16, it is easy to verify that $t_{\min} \leq t_\Delta \leq t_{\max}$. Finally, note that $\mathbb{P}(\mathcal{E}_T) \geq 1 - \frac{6}{\log 4T}$, and the claim 42 now follows.

887 **Proof of claim 42:** The proof of this step is similar to that of equation 27.

888 ²we assume Δ does not change with any T in this corollary.

Regret bound: Let $\tau \equiv 2t_\Delta$ be the time that round 1 ends, and $\Delta = \mu_1 - \mu_2 > 0$. Note that, simple algebra yields at the end of batch 1 – assuming it ends before round T – the active set \mathcal{A} cannot contain both arms. Define the event $\mathcal{F} = (1 \notin \mathcal{A}, \tau < T)$, we have

$$\mathbb{R}_T \leq T\Delta\mathbb{P}(\mathcal{F}) + \frac{\Delta}{2} \cdot \mathbb{E}[\tau \wedge T] \quad (43)$$

We show that

$$\mathbb{P}(\mathcal{F}) \leq \frac{120e\sqrt{\log(\Delta^2 T/4)}}{\Delta^2 T} + \frac{64e}{\Delta^2 T} \quad (44a)$$

$$\mathbb{E}[\tau \wedge T] \leq 4 + \frac{32 \log T}{\Delta^2} + \frac{64}{\Delta^2} \quad (44b)$$

Combining bounds 43, 44a, and 44b we conclude

$$\mathbb{R}_T \leq \frac{16 \log T}{\Delta} + \frac{120e\sqrt{\log(\Delta^2 T/4)} + 64e + 32}{\Delta} + 2\Delta \quad (45)$$

This completes the proof of the corollary 2. It remains to prove the bounds 44a and 44b.

Deriving the bound 44a: Throughout, we assume $T\Delta^2 \geq 4e^2$. We have

$$\begin{aligned} \mathbb{P}(\mathcal{F}) &\leq \mathbb{P}\left(\bar{\mu}_{1,2s} - \bar{\mu}_{2,2s} \leq -\sqrt{\frac{4 \log T}{s}} \text{ for some } s \text{ with } 1 \leq 2s \leq T\right) \\ &\leq \mathbb{P}\left(\frac{\bar{\mu}_{1,2s} - \bar{\mu}_{2,2s}}{\sqrt{2}} \leq -\sqrt{\frac{2 \log(T/2s)}{s}} \text{ for some } s \text{ with } 1 \leq 2s \leq T\right) \\ &\leq \frac{120e\sqrt{\log(\Delta^2 T/4)}}{\Delta^2 T} + \frac{64e}{\Delta^2 T}. \end{aligned}$$

where the last bound utilizes Lemma 1.c from the paper Garivier et al. (2016).

Deriving the bound 44b: By definition, $2 \leq \tau \leq T$, and the τ is always even by construction. We have

$$\begin{aligned} \mathbb{E}[\tau \wedge T] &\leq \sum_{s=1}^T \mathbb{P}(\tau \geq s) \\ &\leq 2 + 2 \sum_{s=1}^{T/2} \mathbb{P}(\tau \geq 2s) \\ &\leq 2 + \frac{32 \log T}{\Delta^2} + 2 \sum_{s \geq \frac{16 \log T}{\Delta^2}}^{T/2} \mathbb{P}\left(\frac{\bar{\mu}_{1,2s} - \bar{\mu}_{1,2s} - \Delta}{\sqrt{2}} \leq \sqrt{\frac{2 \log T}{s}} - \frac{\Delta}{\sqrt{2}}\right) \\ &\leq 2 + \frac{32 \log T}{\Delta^2} + 2 \sum_{s \geq \frac{16 \log T}{\Delta^2}}^{T/2} \mathbb{P}\left(\frac{\bar{\mu}_{1,2s} - \bar{\mu}_{1,2s} - \Delta}{\sqrt{2}} \leq -\frac{\Delta}{2\sqrt{2}}\right) \\ &\stackrel{(i)}{\leq} 2 + \frac{32 \log T}{\Delta^2} + 2 \sum_{s=1}^{T/2} \exp\left\{-\frac{s\Delta^2}{16}\right\} \\ &\leq 2 + \frac{32 \log T}{\Delta^2} + \frac{2}{1 - \exp\{-\Delta^2/16\}} \leq 2 + \frac{32 \log T}{\Delta^2} + \frac{64}{\Delta^2} \wedge 2 \leq 4 + \frac{32 \log T}{\Delta^2} + \frac{64}{\Delta^2} \end{aligned}$$

The implication (i) follows from tail probability bounds of sub-Gaussian random variables.

B NUMERICAL SIMULATIONS

In this section we provide a few simple simulations demonstrating the utility of our theory. We compare the asymptotic distribution of sample means for a two arm bandit problem, with reward distribution $\mathcal{P}_1 \equiv \mathcal{N}(\mu_1, 1)$ and $\mathcal{P}_2 \equiv \mathcal{N}(\mu_2, 1)$. In the simulations below we considered $\mu_1 = \mu_2 = 1$.

Algorithms: We compare two algorithms:

- The stable-ETC strategy, discussed in Algorithm 2.
- The BAI-ETC algorithm by Garivier et al. (2016).

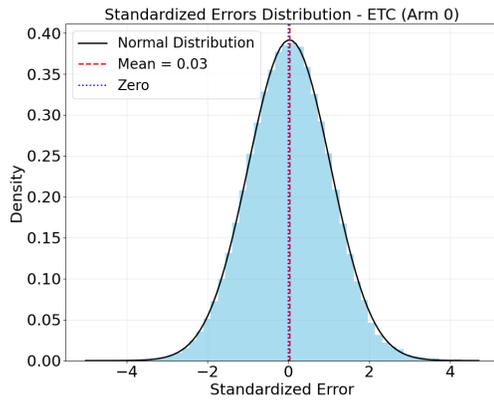


Figure 4: Stable ETC arm 1

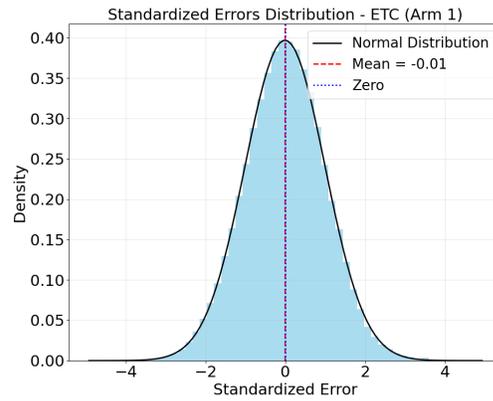


Figure 5: Stable ETC arm 2

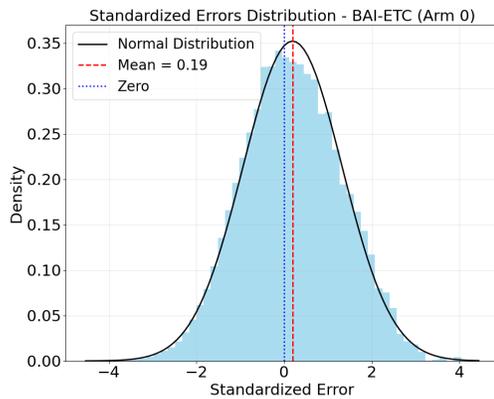


Figure 6: BAI ETC arm 1

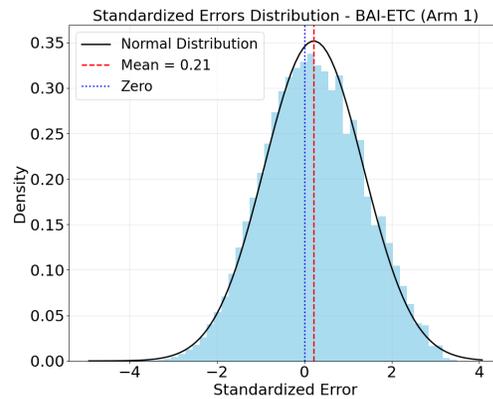


Figure 7: BAI ETC arm 2

Figure 8: Comparison of error distributions for stable-ETC Algorithm 2 and BAI ETC algorithm for a two-armed-bandit. We see that the asymptotic distribution of the arm-means are close to Gaussian when the *stable*-ETC is used. But, the distributions of arm means are not Gaussian when BAI-ETC algorithm is used; the mean of standardized noise are significantly positive, close to 0.20. The simulation results are average of 5000 repetitions and the horizon is set to $T = 1000$.