On Instability of Minimax Optimal Optimism-Based Bandit Algorithms

Samya Praharaj[†] Koulik Khamaru[†]
[†]*Department of Statistics, Rutgers University*

November 25, 2025

Abstract

Statistical inference from data generated by multi-armed bandit (MAB) algorithms is challenging due to their adaptive, non-i.i.d. nature. A classical manifestation is that sample averages of arm-rewards under bandit sampling may fail to satisfy a central limit theorem. Lai and Wei's *stability* condition Lai and Wei (1982) provides a sufficient—and essentially necessary—criterion for asymptotic normality in bandit problems. While the celebrated Upper Confidence Bound (UCB) algorithm satisfies this stability condition, it is not minimax-optimal, raising the question of whether minimax optimality and statistical stability can be achieved simultaneously. In this paper, we analyze the stability properties of a broad class of bandit algorithms that are based on the *optimism-principle*. We establish general structural conditions under which such algorithms violate the Lai–Wei stability criterion. As a consequence, we show that widely used minimax-optimal UCB-style algorithms—including MOSS, Anytime-MOSS, Vanilla-MOSS, ADA-UCB, OC-UCB, KL-MOSS, KL-UCB++, KL-UCB-SWITCH and Anytime KL-UCB-SWITCH—are unstable. We further complement our theoretical results with numerical simulations demonstrating that, in all these cases, the sample means fail to exhibit asymptotic normality.

Overall, our findings suggest a fundamental tension between stability and minimax-optimal regret, raising the question of whether it is possible to design bandit algorithms that achieve both. Understanding whether such simultaneously stable and minimax-optimal strategies exist remains an important open direction.

1 Introduction

Sequential inference (Wald (1992), Lai and Robbins (1985), Siegmund (2013)) and online decision making (Supancic III and Ramanan (2017), Bastani et al. (2025), Bastani and Bayati (2020)) concern the study of how learners update their beliefs and choose actions in real time as new information arrives. Unlike batch settings, where all data is available at once, sequential settings require processing streaming data efficiently, often under uncertainty and resource constraints. Multi-armed bandit algorithms (MAB) are a class of online strategies that have found wide applicability across a diverse range of domains which include online advertising (Wen et al. (2017), Vaswani et al. (2017)), adaptive clinical trials (Bastani and Bayati (2020)), personal recommendation systems (Bouneffouf

et al. (2012), Bouneffouf et al. (2013), Zhou et al. (2017)) and financial portfolio optimization (Shen et al. (2015), Huo and Fu (2017)).

The general K-armed MAB problem can be formulated as follows. A learner has K reward distributions with means $\mu_1, \mu_2, \ldots, \mu_K$, respectively. The set of indices $\{1, 2, \ldots, K\}$ is called the *action space*. At time t, the learner observes the history $\mathcal{F}_{t-1} := \sigma(A_1, X_1, \ldots, A_{t-1}, X_{t-1})$—the σ-field generated by the data collected up to time $t-1$—and chooses an action A_t. The function mapping the history to the action space is called a *policy*. If $A_t = a$, then the learner receives a reward X_t from arm a. Formally,

$$X_t = \mu_a + \epsilon_{a,t} \qquad \text{if } A_t = a. \tag{1}$$

The zero-mean noise $\epsilon_{a,t}$ is assumed to be independent of the history $\mathcal{F}_{t-1}$ at every round. To make the presentation succinct, we assume that all arms have the same variance, i.e., $\mathbb{E}[\epsilon_{a,t}^2] = \sigma^2$ for all arms $a \in [K]$ and rounds $t \geq 1$. We use $n_{a,t}$ to denote the number of pulls of arm a up to time t, and T to denote the total number of interactions with the environment. The loss function for this problem is called the *regret*, defined as

$$\mathrm{Reg}(T) := \sum_{t=1}^{T} (\mu^\star - \mu_{A_t}), \tag{2}$$

where $\mu^\star := \max_{a \in [K]} \mu_a$ is the maximum arm mean.

The core challenge in the MAB problem is the exploration-exploitation tradeoff. As the reward distributions are unknown, non-adequate arm pulls may lead to sub-optimal arm choice, whereas over-pulling of sub-optimal arms may lead to higher regret. To address this issue, a popular class of bandit algorithms called Upper Confidence Bound (UCB) was proposed by Lai and Robbins (Lai and Robbins (1985)) and later popularized by Lai (1987); Katehakis and Robbins (1995); Auer et al. (2002). In the UCB algorithm at time t, the learner creates high probability upper confidence bounds $U_a(t)$ of the mean μ_a for each arm a, and selects the arm with highest bound. For different choice of $U_a(t)$ one obtains different UCB strategies: some prominent examples include UCB (Lai (1987), Auer et al. (2002)), MOSS (Audibert and Bubeck (2009)), UCB-V (Audibert et al. (2009)) and KL-UCB (Garivier and Cappé (2011)).

While bandit algorithms are traditionally motivated by regret minimization, in several studies experimenters and scientists are also interested to conduct statistical inference from the data collected via running a bandit algorithm (Chow et al. (2005); Berry (2012); Zhang et al. (2022); Trella et al. (2023, 2024); Zhang et al. (2024); Trella et al. (2025)). However, due to the adaptive structure of the bandit strategies the data collected is highly non-i.i.d. As a consequence, standard techniques to conduct inference may lead to erroneous conclusions. In particular, the sample means—computed from data collected from adaptive algorithms—may not converge to normal distribution (Zhang et al. (2020); Deshpande et al. (2023); Khamaru et al. (2021); Lin et al. (2023, 2025)). To tackle this issue, a growing line of research (Kalvit and Zeevi (2021),Khamaru and Zhang (2024); Fan et al. (2024); Han et al. (2024); Halder et al. (2025)) focuses on a notion called *stability* of bandit algorithms which was introduced by Lai and Wei (1982). It is defined as follows:

Definition 1. A K-armed bandit algorithm $\mathcal{A}$ is called stable if for all arm $a \in [K]$, there exist *non-random* scalars $n^*_{a,T}(\mathcal{A})$ such that

$$\frac{n_{a,T}(\mathcal{A})}{n^*_{a,T}(\mathcal{A})} \xrightarrow{\mathrm{P}} 1 \quad \text{and} \quad n^*_{a,T}(\mathcal{A}) \to \infty \ \text{ as } \ T \to \infty. \tag{3}$$

Here $n_{a,T}(\mathcal{A})$ denotes the number of times arm a has been pulled up to time T.

Stability provides a sufficient condition ensuring that sample means exhibit asymptotic normality as $T \to \infty$ (see Section 3 for a detailed discussion). Conversely, as discussed in Appendix C, the absence of stability may prevent the applicability of a central limit theorem, due to non-normality. Thus, determining whether a given bandit algorithm is stable is a question of both theoretical and practical significance. While the stability of UCB-1 has been established (Fan and Glynn (2022); Khamaru and Zhang (2024); Han et al. (2024); Chen and Lu (2025)), this algorithm does not attain the minimax lower bound of $\Theta(\sqrt{KT})$ (Han et al. (2024), Lattimore and Szepesvári (2020)). In this paper we consider a popular class of bandit algorithms which are minimax optimal (Audibert and Bubeck (2009)), which we call the `MOSS-family` . Our goal is to investigate the extent to which this `MOSS-family` of bandit strategies exhibit the stability notion (equation (3)).

1.1 Related Work

When data is collected adaptively, local dependencies in sample points lead to non-i.i.d structure where standard techniques can fail. Such problems are known to fail to hold when applied to data collected from time-series and econometric studies (Dickey and Fuller (1979); White (1958, 1959)). In particular, for bandit algorithms Lai and Wei (1982) introduced the notion of stability and established that asymptotic normality holds under this condition. Building on this foundation, subsequent research has leveraged stability to construct asymptotically valid confidence intervals (Kalvit and Zeevi (2021), Khamaru and Zhang (2024), Fan and Glynn (2022), Fan et al. (2024), Han et al. (2024), Halder et al. (2025)).

This work aims to bridge the gap in understanding between the class of minimax-optimal bandit algorithms and asymptotic inference through Lai's stability condition. The MOSS algorithm was the first to be shown to achieve minimax regret (Audibert and Bubeck (2009)). Since then, several variants of MOSS (Garivier et al. (2016), Ménard and Garivier (2017), Lattimore and Szepesvári (2020)), Lattimore (2015)) have been proposed to address its limitations (see Section 2 for details). We note that results concerning the instability of multi-armed bandit algorithms remain relatively scarce. Recently Fan et al. (2024) have identified that the UCB-V algorithm can be unstable in some cases.

We note that there exists some alternate approaches to construction of valid confidence intervals (CI). One is the non-asymptotic approach which constructs finite-sample CI via the application of concentration inequalities for self-normalized martingales. This line of work is build on the foundational analysis of de la Pena et al (de la Pena et al. (2004), De la Pena et al. (2009)), which hold uniformly over time (see Abbasi-Yadkori et al. (2011); Howard et al. (2020); Waudby-Smith et al. (2024), Waudby-Smith et al. (2024), and references therein). However, the CI obtained via stability are often substantially shorter than those derived from anytime-valid approaches.

As the data collected via bandit algorithms are adaptive in nature, the sample means $\widehat{\mu}_a$ is biased. To tackle this issue, several researchers have studied bandits from the debiasing or doubly-robust perspective (Zhang and Zhang (2014); Krishnamurthy et al. (2018); Kim and Paik (2019); Chen et al. (2022); Kim et al. (2023)). Furthermore, in the context of Off Policy Evaluation (OPE) the uncertainty quantification problem has been approached via inverse-propensity scores and importance sampling techniques (Hadad et al. (2021); Zhang et al. (2022); Nair and Janson (2023); Leiner et al. (2025)).

1.2 Contributions

In this paper, we study the instability properties of a broad class of optimism principle (11) based multi-armed bandit algorithms. This class includes the UCB-style minimax-optimal algorithms in Table 1—such as MOSS, Anytime-MOSS, Vanilla-MOSS, ADA-UCB, and OC-UCB—as well as the Kulback-Leibler (KL) divergene based minimax optimal algorithms in Table 2, including KL-MOSS, KL-UCB++, KL-UCB-SWITCH, and Anytime-KL-UCB-switch. These algorithms are introduced formally in Section 2. Section 3 provides a detailed discussion of the Lai–Wei stability condition for multi-armed bandits and its implications for asymptotic inference. Our main theoretical result, Theorem 1, establishes general sufficient conditions under which an optimism-based algorithm fails to satisfy the Lai–Wei stability criterion. As a consequence, Corollary 1 shows that all algorithms in Tables 1 are unstable when the reward distribution is 1 sub-Gaussian. Furthermore, Corollary 2 shows that algorithms listed in Table 2 are unstable under one parameter exponential family of distribution. Finally, in Section 4.1, we complement our theoretical findings with numerical experiments demonstrating that the sample means produced by these algorithms fail to exhibit asymptotic normality.

2 The MOSS Family of Algorithms

The multi-armed bandit and reinforcement learning based algorithms have gained renewed interest in modern times due to their various applications. One popular class of algorithms is the upper confidence bound family. An important and widely used method in this class is the Upper Confidence Bound (UCB)-1 algorithm. At every round t, the UCB-1 algorithm Auer et al. (2002) selects arm A_{t+1} according to

$$A_{t+1} = \arg\max_{a\in[K]} \left\{ \widehat{\mu}_a(t) + \sqrt{\frac{2\log T}{n_{a,t}}} \right\}, \tag{4}$$

where $\widehat{\mu}_a(t)$ is the empirical mean of the observed rewards for arm a up to time t. The regret of the UCB-1 algorithm is $O(\sqrt{KT\log T})$, which is sub-optimal by a factor $\sqrt{\log T}$. The Minimax Optimal Strategy in the Stochastic case (MOSS) algorithm, introduced by Lai (1987) and later refined and popularized by Audibert and Bubeck (2009), resolves this sub-optimality and attains a minimax regret of $O(\sqrt{KT})$. The MOSS algorithm selects arms using the update rule

$$A_{t+1} := \arg\max_{a\in[K]} \left\{ \widehat{\mu}_a(t) + \sqrt{\frac{\max\{0, \log((T/K)/n_{a,t})\}}{n_{a,t}}} \right\}. \tag{5}$$

At a high level,

- In the *early stages*—when $n_{a,t}$ is small—the MOSS algorithm behaves similarly to the UCB-1 algorithm.
- In the *later stages*—when $n_{a,t}$ is large—MOSS reduces the uncertainty for heavily explored arms. This modification leads to improved regret bounds compared to UCB-1.

We detail the MOSS algorithm in Algorithm 1.

Algorithm 1: MOSS (Minimax Optimal Strategy in the Stochastic case)

Input : Number of arms K; time horizon T.
Output: Chosen arms $(A_t)_{t=1}^T$.

For arm a at round t, let $n_{a,t}$ be pulls so far, $\widehat{\mu}_a(t)$ its mean, and define

$$U_a(t) = \widehat{\mu}_a(t) + \sqrt{\frac{\max\{0, \log(\frac{T/K}{n_{a,t}})\}}{n_{a,t}}}.$$

Init: Pull each arm once.
for $t = K{+}1$ **to** T **do**
  Select $A_t \in \arg\max_a U_a(t-1)$.
  Pull A_t, observe X_t.
  $n_{a,t} \leftarrow n_{a,t-1} + 1\{A_t = a\}, \quad \widehat{\mu}_{a,t} \leftarrow \frac{n_{a,t-1}}{n_{a,t}}\widehat{\mu}_{a,t-1} + \frac{X_t}{n_{a,t}}\mathbb{I}_{\{A_t=a\}}$.

While MOSS attains the minimax optimal regret, it also exhibits several important limitations. First, the algorithm requires prior knowledge of the horizon T, which is unavailable in many practical settings. Second, MOSS is not *instance-optimal* Lattimore and Szepesvári (2020); despite its favorable minimax performance, it may behave poorly on specific bandit instances. These issues have motivated the development of refined variants that retain (or nearly retain) the minimax optimality of MOSS while improving performance in more nuanced regimes. Table 1 provides detailed descriptions of these algorithms for 1-sub-Gaussian reward distributions.

Table 1: Upper Confidence Bound for MOSS variants

Algorithm	**Upper Confidence Bound** $U_a(t)$
MOSS	$\widehat{\mu}_a(t) + \sqrt{\frac{\max\{0,\ \log(\frac{T}{Kn_{a,t}})\}}{n_{a,t}}}$
Anytime-MOSS	$\widehat{\mu}_a(t) + \sqrt{\frac{\max\{0,\ \log(\frac{t}{Kn_{a,t}})\}}{n_{a,t}}}$
Vanilla-MOSS	$\widehat{\mu}_a(t) + \sqrt{\frac{\log(\frac{T}{n_{a,t}})}{n_{a,t}}}$
OC-UCB	$\widehat{\mu}_a(t) + \sqrt{\frac{2(1+\epsilon)}{n_{a,t}}\log(\frac{T}{t})}$
ADA-UCB	$\widehat{\mu}_a(t) + \sqrt{\frac{2}{n_{a,t}}\log\left(\frac{T}{H_a}\right)}$, where $H_a := \sum_{j=1}^k \min(n_{a,t}, \sqrt{n_{a,t}, n_{j,t}})$

Lattimore (2015) proposed OC-UCB (Optimally Confident UCB), which achieves the optimal worst-case regret $O(\sqrt{KT})$ while also attaining nearly optimal problem-dependent regret in

Gaussian bandits. Subsequently, Lattimore (2018) introduced ADA-UCB, which is both minimax-optimal and asymptotically optimal. To address the dependence on the horizon, an *anytime* version of MOSS can be obtained by replacing T with t in the exploration term. Moreover, Section 9.2 of Lattimore and Szepesvári (2020) documents that MOSS may incur worse regret than UCB in certain regimes. This can be mitigated by removing the dependence on the number of arms K in the exploration function, yielding a variant which we call *Vanilla-MOSS*.

All such algorithms arise from replacing the Upper Confidence Bound $U_a(t)$ in step 1 of Algorithm 1 with suitable modifications, each grounded in the principle of *optimism in the face of uncertainty* Lattimore and Szepesvári (2020). These algorithms are known to perform well when the reward distribution is σ sub-Gaussian. For a sub-class of reward distributions, the UCB-1 algorithm can be improved by using the notion of Kullback-Leibler (KL) divergence:

Definition 2 (Kullback–Leibler Divergence)**.** Let P and Q be two probability distributions on the same measurable space, with densities p and q with respect to a common dominating measure. The *Kullback–Leibler divergence* of Q from P is defined as

$$\mathrm{KL}(P, Q) = \int p(x) \log\left(\frac{p(x)}{q(x)}\right) dx.$$

When the domain of the random variable is discrete, it reduces to

$$\mathrm{KL}(P, Q) = \sum_x p(x) \log\left(\frac{p(x)}{q(x)}\right).$$

For one-dimensional canonical exponential distributions the KL-divergence has a simplified form. Suppose P_θ is absolutely continuous w.r.t. some dominating measure ρ on $\mathbb{R}$ with density $\exp(x\theta - b(\theta))$. Then the KL-divergence becomes a function of the population means

$$\mathrm{KL}(P_{\theta_1}, P_{\theta_2}) = b(\theta_2) - b(\theta_1) - b'(\theta_1)(\theta_2 - \theta_1)$$

By reparameterization $b'(\theta_1) = \mu_1, b'(\theta_2) = \mu_2$, we can rewrite the KL-divergence as follows:

$$\mathrm{KL}(P_{\theta_1}, P_{\theta_2}) = \mathrm{KL}(\mu_1, \mu_2)$$

For Bernoulli distribution, one can obtain tighter confidence bounds in terms of the KL-divergence, than the one obtained for general sub-Gaussian distribution. This notion has been extended to bounded and one-parameter exponential family of distributions (Ménard and Garivier (2017)). By combining this idea with the UCB-1 algorithm (equation (4)), we have the KL-UCB algorithm (Filippi, 2010; Maillard et al., 2011; Garivier and Cappé, 2011). Let I be the compact set of possible values of population mean μ. Then at every round t, select arm A_{t+1} according to

$$A_{t+1} = \arg\max_{a \in [K]} \left\{ \sup \left\{ q \in I : \mathrm{KL}(\widehat{\mu}_{a,t}, q) \leq \frac{f(t)}{n_{a,t}} \right\} \right\}, \tag{6}$$

where $f(t) = \log t + \log\log t$

As the minimax regret of KL-UCB is also sub-optimal, subsequent work combined insights from both MOSS and KL-UCB, yielding refined algorithms such as KL-MOSS, KL-UCB-SWITCH, and KL-UCB++ (Garivier et al., 2016; Ménard and Garivier, 2017; Garivier et al., 2022). KL-UCB-SWITCH and Anytime KL-UCB-SWITCH achieve optimal distribution-free worst-case regret, while KL-UCB++ is both minimax-optimal and asymptotically optimal for single-parameter exponential families. Table 2 summarizes their UCB indices.

Table 2: Upper Confidence Bound for KL-based MOSS variants

Algorithm	**Upper Confidence Bound** $U_a(t)$
KL-MOSS	$\sup\left\{q \in I : \mathrm{KL}(\widehat{\mu}_a(t), q) \le \dfrac{\log^+(T/n_{a,t})}{n_{a,t}}\right\}$
KL-UCB++	$\sup\{q \in I : \mathrm{KL}(\widehat{\mu}_a(t), q) \le \dfrac{h(n_{a,t})}{n_{a,t}}\}$, where $h(n) := \log^+\left(\dfrac{T/K}{n}\left(\log^+\dfrac{T/K}{n}\right)^2 + 1\right)$
KL-UCB-Switch	$\begin{cases} U_a^{KM}(t), & n_{a,t} \le \lfloor (T/K)^{1/5} \rfloor \\ U_a^{M}(t), & n_{a,t} > \lfloor (T/K)^{1/5} \rfloor \end{cases}$, where $U_a^{KM}(t), U_a^{M}(t)$ are the UCB of KL-MOSS and MOSS respectively
Anytime–KL-UCB-Switch	$\begin{cases} \sup\left\{q \in I : \mathrm{KL}(\widehat{\mu}_a(t), q) \le \dfrac{\phi\left(\frac{t}{Kn_{a,t}}\right)}{n_{a,t}}\right\}, & \text{if } n_{a,t} \le f(t,K), \\ \widehat{\mu}_a(t) + \sqrt{\dfrac{1}{2n_{a,t}}\phi\left(\frac{t}{Kn_{a,t}}\right)}, & \text{if } n_{a,t} > f(t,K), \end{cases}$ where $\phi(x) := \log^+(x)(1 + (\log^+ x)^2)$ and $f(t,K) = [(t/K)^{1/5}]$.

3 Inference in Multi-Armed Bandits via Stability

The ability to conduct valid inference is of particular importance in settings such as clinical trials (Marschner, 2024; Trella et al., 2023, 2024; Zhang et al., 2024; Trella et al., 2025) and A/B testing ((Dubarry, 2015)). A central difficulty is that data generated by multi-armed bandit algorithms are inherently non-i.i.d., rendering classical estimators biased and inconsistent in general (Zhang et al., 2020; Deshpande et al., 2018). Consequently, the direct application of standard asymptotic theory for hypothesis testing and uncertainty quantification can lead to severely distorted inference.

To address these challenges, recent work has focused on the notion of *stability* in bandit algorithms, a concept originally introduced in Lai and Wei (1982). Informally, a multi-armed bandit algorithm $\mathcal{A}$ is stable if the number of times each arm is pulled concentrates around certain algorithm-dependent deterministic quantities. Formally, we recall from the definition(3) that a K-armed bandit algorithm $\mathcal{A}$ is called stable if, for all arms $a \in [K]$, there exist *non-random* scalars $n^*_{a,T}(\mathcal{A})$ such that

$$\frac{n_{a,T}(\mathcal{A})}{n^*_{a,T}(\mathcal{A})} \xrightarrow{\mathbb{P}} 1, \qquad n^*_{a,T}(\mathcal{A}) \to \infty \quad \text{as } T \to \infty. \tag{7}$$

Here $n_{a,T}(\mathcal{A})$ denotes the number of times arm a has been pulled up to time T.

A key implication of stability is that when data are collected via a stable bandit algorithm $\mathcal{A}$, classical asymptotic inference becomes valid despite the adaptive data collection. In particular, one obtains an i.i.d.-like asymptotic normality result:

Lemma 1. *Suppose we have a stable K-armed MAB algorithm with sub-Gaussian reward distributions of variance 1 for all arms. Then the following distributional convergence holds:*

$$\begin{bmatrix} \sqrt{n_{1,T}(\mathcal{A})}\,(\widehat{\mu}_{1,T} - \mu_1) \\ \vdots \\ \sqrt{n_{K,T}(\mathcal{A})}\,(\widehat{\mu}_{K,T} - \mu_K) \end{bmatrix} \xrightarrow{d} N(0, I_K). \tag{8}$$

This result follows from (Lai and Wei, 1982, Theorem 3), combined with the martingale central limit theorem (Hall and Heyde, 2014; Dvoretzky, 1972).

A growing line of recent work has examined the stability of widely used bandit algorithms. Kalvit and Zeevi (2021) establish stability of the UCB-1 algorithm (Auer et al., 2002) in the two-armed case under certain conditions on the arm means, with extensions to the unequal-variance setting Fan et al. (2024); Chen and Lu (2025). Fan and Glynn (2022) prove stability of UCB-1 and Thompson Sampling under the assumption of a unique optimal arm. The most general results to date are due to Khamaru and Zhang (2024) and Han et al. (2024), who establish stability for a broad class of UCB-type algorithms with potentially growing numbers of arms K, arbitrary arm means, and provide rates of convergence for the central limit theorem in Lemma 1.

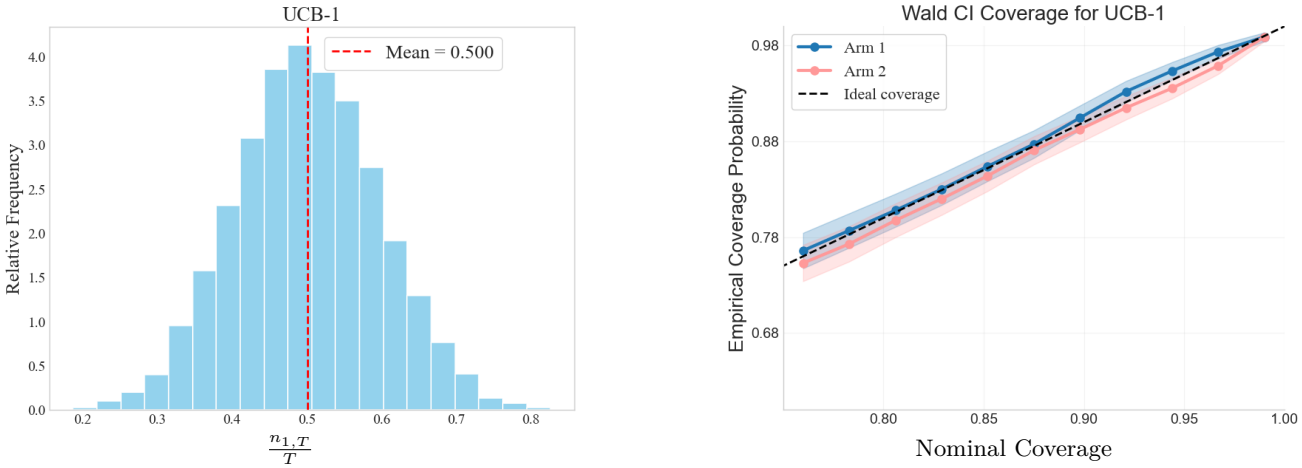


Figure 1: **Left**: Empirical results for UCB-1 based on 5000 independent sample runs with $T = 10000$ pulls. **Right**: Coverage of Wald confidence interval of MOSS with $T = 10000$.

Is Stability Necessary for Asymptotic Normality?

Lemma 1 shows that stability is a sufficient condition to achieve asymptotic normality. This naturally leads to the question of whether stability is, in some sense, also necessary for a central limit theorem to hold. To understand this, it is useful to peek at the proof of Lemma 1. Consider a K-armed bandit problem from (1). Given a bandit algorithm $\mathcal{A}$, and for an arm $a \in [K]$, the

standardized estimation error admits the decomposition

$$\sqrt{n_{a,T}(\mathcal{A})}\,(\widehat{\mu}_{a,T}-\mu_a) = \frac{1}{\sqrt{n_{a,T}(\mathcal{A})}}\sum_{t=1}^{T}\xi_{a,t}\,\mathbb{I}\{A_t=a\}$$

$$= \underbrace{\sqrt{\frac{n^*_{a,T}(\mathcal{A})}{n_{a,T}(\mathcal{A})}}}_{\approx 1}\cdot\underbrace{\frac{1}{\sqrt{n^*_{a,T}(\mathcal{A})}}\sum_{t=1}^{T}\xi_{a,t}\,\mathbb{I}\{A_t=a\}}_{\text{(normalized sum of an MDS)}}. \tag{9}$$

The first factor in (9) converges to 1 in probability by stability. The second factor is a normalized sum of martingale differences adapted to $(\mathcal{F}_t)$. Its conditional variance is given by

$$\sum_{t=1}^{T}\mathrm{Var}\left(\frac{1}{\sqrt{n^*_{a,T}(\mathcal{A})}}\xi_{a,t}\,\mathbb{I}\{A_t=a\}\,\middle|\,\mathcal{F}_{t-1}\right)=\frac{n_{a,T}(\mathcal{A})}{n^*_{a,T}(\mathcal{A})}, \tag{10}$$

which again converges in probability to 1 under stability. Thus, in the multi-armed bandit setting, stability is precisely the condition that the sum of conditional variances of the underlying martingale difference sequence stabilizes, allowing a martingale central limit theorem to yield asymptotic normality.

Conversely, violations of such variance stabilization are well known to lead to non-Gaussian limits in classical martingale settings, including Gaussian mixture limits (Aldous and Eagleson (1978); Hall and Heyde (2014); Lai and Wei (1982)) and limits arising from Brownian motion indexed by a stopping time (Hall and Heyde (2014)). In the context of bandit algorithms, Zhang et al. (2020) show that several standard batched policies exhibit non-normal limiting behavior. As we discuss in Appendix C, these algorithms fail the stability condition.

4 Instability of MOSS and it's Variants

While UCB-1 is provably stable Khamaru and Zhang (2024); Han et al. (2024), its regret is not optimal Lattimore and Szepesvári (2020); Han et al. (2024). This motivates the search for bandit algorithms which are both stable and attains the minimax-optimal regret of $\sqrt{KT}$. A class of bandit algorithms that attain the minimax optimal regret is the MOSS-family (see Table 1). The main result of this paper is to prove that MOSS family of algorithms, while attaining minimax-optimal regret, are not stable for a K-armed bandit problem. Furthermore, in Section 4.1, we also demonstrate through extensive simulations that the normality-based confidence intervals do not provide correct coverage for the MOSS-family of algorithms.

Our main result applies to a wide class of algorithms that select an arm according to the following *optimism-based* rule

$$A_t := \arg\max_{a\in[K]} U_a(t) \qquad \texttt{(Optimism-rule)} \tag{11}$$

where $U_a(t)$ is an appropriate upper confidence bound of arm mean μ_a at time t. We assume that our optimism-based rule (11) satisfies the following conditions:

Assumption A.

(A1) For any arm $a \in [K]$ at any time $t \in [T]$, the UCB $U_a(t)$ is a function of $\{\widehat{\mu}_a(t) : \ a \in [K]\}$ and $\{n_{a,t} : a \in [K]\}$.

(A2) Fix a time $t \in [T]$ and arm $a \in [K]$. If arm a is not pulled at time t, then $U_a(t) \leq U_a(t-1)$.

(A3) We assume that on the event defined below:

$$\mathcal{E}_T := \left\{ n_{a,\frac{T}{2K}} \in \left[\frac{T}{4K^2}, \ \frac{3T}{4K^2} \right] \ \text{ for all } \ a \in [K] \right\} \tag{12a}$$

the following sandwich relation holds for $t \geq T/2K$

$$\widehat{\mu}_{a,t} + \frac{\beta_1}{\sqrt{n_{a,t}}} \leq \ U_a(t) \ \leq \ \widehat{\mu}_{a,t} + \frac{\beta_2}{\sqrt{n_{a,t}}}. \tag{12b}$$

Here β_1, β_2 are suitable real (algorithm dependent) constants.

Assumption (A1) states that once we know the sample means $\widehat{\mu}_{a,t}$ and the number of arm pulls $n_{a,t}$ for each arm a at time t, then we exactly know which arm to pull at next step $t+1$. We note that if for some algorithm $\mathcal{A}$ this assumption is not satisfied, our proof may still go through provided we have more explicit information on the UCB of $\mathcal{A}$. Assumption (A2) posits a mild assumption on the behavior of the upper confidence bound $U_a(t)$ of arms that are not pulled. This assumption is satisfied for all Algorithms in Table 1 except Anytime MOSS. Our instability proofs also work for Anytime-MOSS with a slight modification of our argument. The requirement in Assumption (A3) is more subtle. This assumption requires that for large value of t, the upper confidence bound (UCB) $U_a(t)$ for all arms are sandwiched between two UCBs with *constant bonus* factors. We demonstrate that all algorithms in Table 1 satisfy this condition. A closer look at the proof of Theorem 1 reveals that this constant-bonus sandwich behavior of the UCB's in the later stage of the algorithm leads to their instability; we return to this point in the comments following Theorem 1.

Theorem 1. *Consider a K armed bandit problem where all arms have identical 1-sub-Gaussian reward distribution. Then any optimism based rule (11) satisfying Assumptions (A1), (A2) and (A3) is unstable.*

The proof of this theorem is provided in Section 5. Let us now discuss how Assumption (A3) breaks the stability of *optimism-based* rule 11. Observe that under Assumption (A3) for large values of t the upper confidence bound is safely approximated (sandwiched) by an optimism rule with the following UCB:

$$\widetilde{U}_{a,\beta}(t) = \widehat{\mu}_{a,t} + \frac{\beta}{\sqrt{n_{a,t}}} \tag{14}$$

where β is an appropriate constant that is independent on $n_{a,t}$ and T. It is instructive to compare the rule 14 with the UCB-1 algorithm analyzed in Khamaru and Zhang (2024); Han et al. (2024). The authors show that any UCB style algorithm with UCB's of the form

$$\mathrm{UCB}_a(t) := \widehat{\mu}_{a,t} + \frac{q_T}{\sqrt{n_{a,t}}} \tag{15}$$

is stable if $q_T \gg \sqrt{2\log\log T}$. This class of algorithms include for instance the popular UCB-1 algorithm by Auer et al. (2002). The term $\sqrt{2\log\log T}$ originates from the law of iterated logarithm (LIL) which captures the fluctuation of the sample mean from the population mean over T rounds. In the proof of Theorem 1 we show that if the bonus term q_T in rule 15 is replaced by a constant (independent of T), then the resulting optimism based rule becomes unstable.

All algorithms considered in Tables 1 and 2 satisfy Assumption (A1). Furthermore, a simple inspection of algorithms in Table 1 satisfy Assumptions (A2) and (A3) except Anytime MOSS. In Appendix D we prove that a simple modification of the proof argument of Theorem 1 also shows the instability of Anytime Moss. Overall we conclude.

Corollary 1. *Consider a K armed bandit problem where all arms have identical 1-sub-Gaussian reward distribution. Then algorithms stated in Table 1, while achieving minimax optimal regret, are unstable.*

Note that the KL-divergence based algorithms in Table 2 (except Anytime KL-UCB-SWITCH) satisfy Assumption (A2). However, Assumption (A3) is not necessarily satisfied for any sub-Gaussian reward distribution. Here we outline a sub-class of reward distributions for which Assumption (A3) holds. Suppose the reward distribution is such that the following inequality is satisfied for some constant $c > 0$ (which is distribution dependent):

$$\mathrm{KL}(\widehat{\mu}_a(t), q) \;\geq\; c\,(\widehat{\mu}_a(t) - q)^2. \tag{16}$$

Garivier et al. (2019) show that any distribution supported on $[0, 1]$ satisfies (16). In particular, Bernoulli and uniform distributions fall into this class. Moreover, within canonical exponential families, inequality (16) also holds whenever the mean parameter space I is bounded and the variance is uniformly bounded. This contains a large class of distributions including Bernoulli and Gaussian random variables (with I bounded). Now, for any such reward distribution, the KL-based algorithms—KL-MOSS, KL-UCB++, and KL-UCB-SWITCH—listed in Table 2 obey the sandwich bound:

$$\widehat{\mu}_a(t) - \sqrt{\frac{f(T/n_{a,t})}{n_{a,t}}} \;\leq\; U_a(t) \;\leq\; \widehat{\mu}_a(t) + \sqrt{\frac{f(T/n_{a,t})}{n_{a,t}}} \tag{17}$$

for some appropriate choices of f. Finally, note that the UCB indices of Anytime KL-UCB-SWITCH are sandwiched between the UCB of two *unstable* algorithms which are Anytime in nature. For a rigorous argument see Appendix D. The discussion above leads us to our second corollary.

Corollary 2. *Consider a K-armed bandit problem where all arms have identical 1-sub-Gaussian reward distributions satisfying equation (16). Then the algorithms stated in Table 2, while achieving minimax optimal regret, are unstable.*

4.1 Illustrative simulations

In this section, we present simulation results that illustrate both the instability of the algorithms in Tables 1 and 2 and the failure of CLT-based confidence intervals (Wald-type intervals) for estimating the arm means. All experiments consider a two-armed bandit with identical reward distributions $\mathcal{N}(0, 1)$. For each algorithm, we report two plots. The left panel displays the empirical *relative*

frequencies of arm pulls, i.e., $n_{1,T}/T$. The right panel reports the empirical coverage of the standard Wald confidence interval for the mean reward. In every case, the Wald interval exhibits substantial undercoverage, indicating significant deviations from normality. These findings provide empirical support for our theoretical results establishing instability of these algorithms, as well as direct evidence that the central limit theorem does not hold for the sample arm-means.

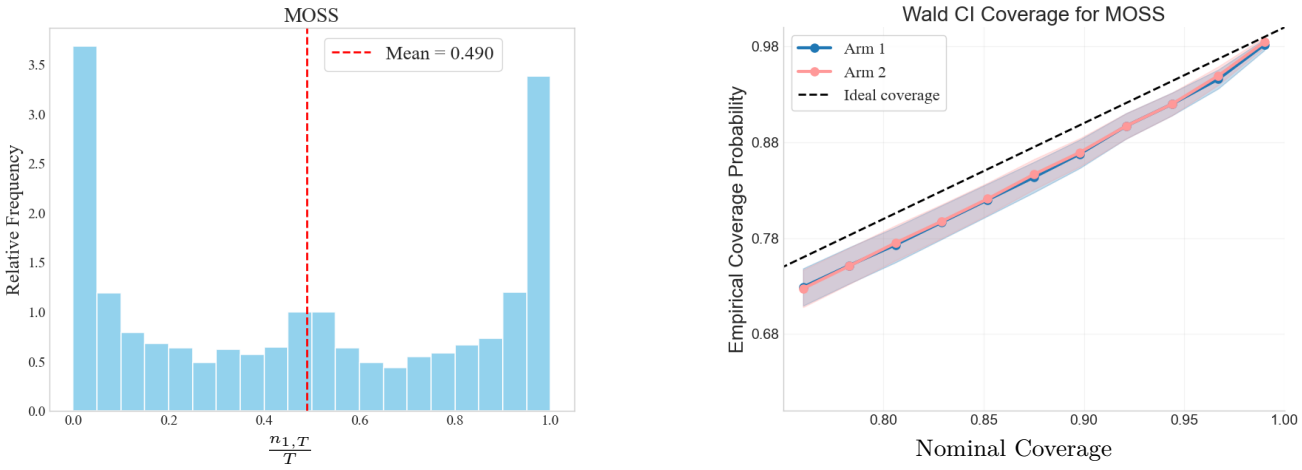


Figure 2: **Left**: Empirical results for MOSS based on 5000 independent sample runs with $T = 10000$ pulls. **Right**: Coverage of Wald confidence interval of MOSS with $T = 10000$.

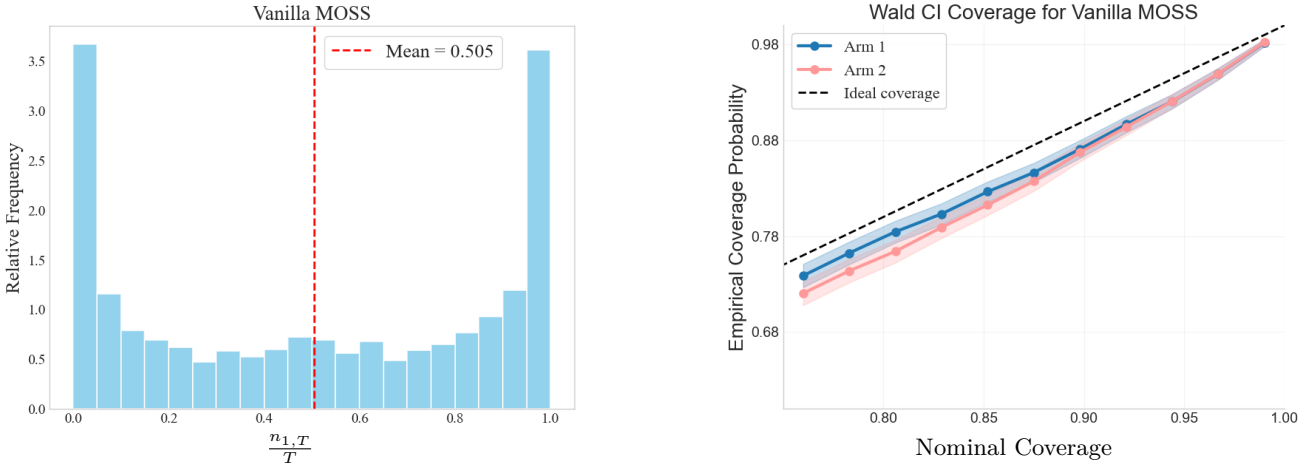
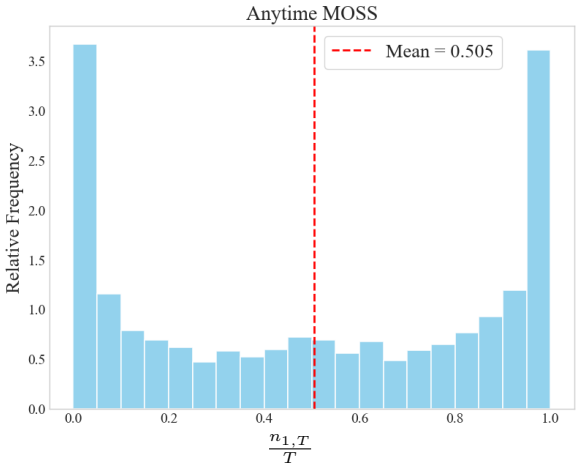


Figure 3: **Left**: Empirical results for Vanilla MOSS based on 5000 independent sample runs with $T = 10000$ pulls. **Right**: Coverage of Wald confidence interval of Vanilla MOSS with $T = 10000$.



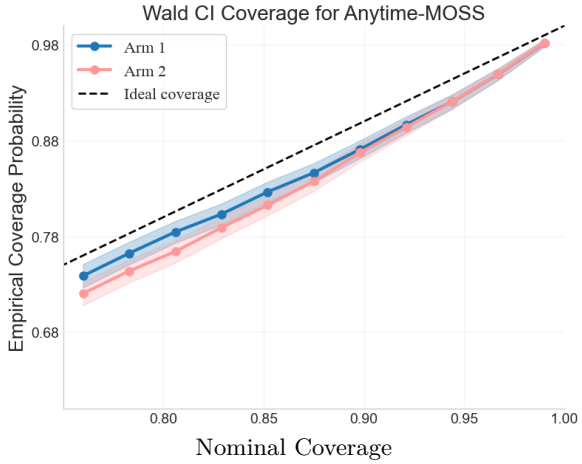
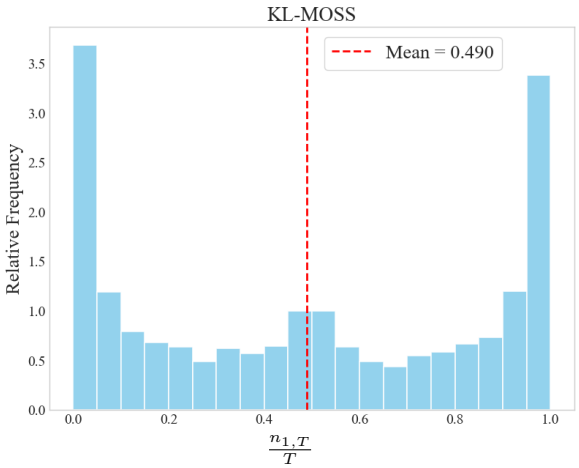


Figure 4: **Left**: Empirical results for Anytime MOSS based on 5000 independent sample runs with $T = 10000$ pulls. **Right**: Coverage of Wald confidence interval of Anytime MOSS with $T = 10000$.



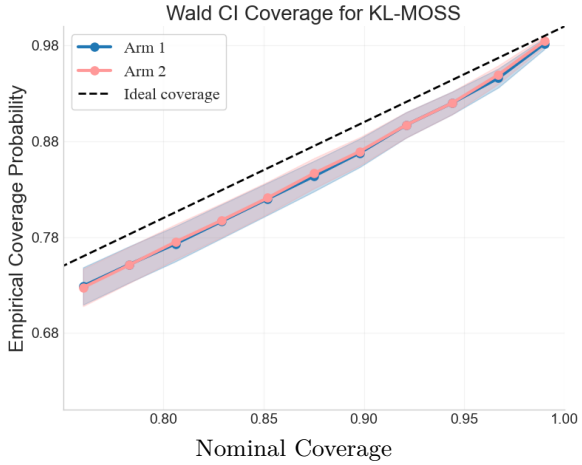
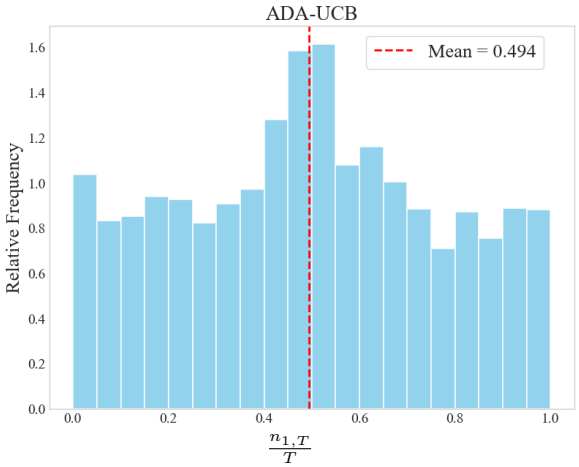


Figure 5: **Left**: Empirical results for KL-MOSS based on 5000 independent sample runs with $T = 10000$ pulls. **Right**: Coverage of Wald confidence interval of KL-MOSS with $T = 10000$.



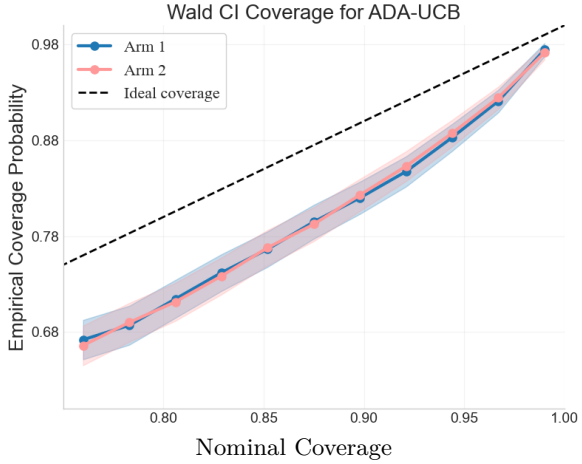
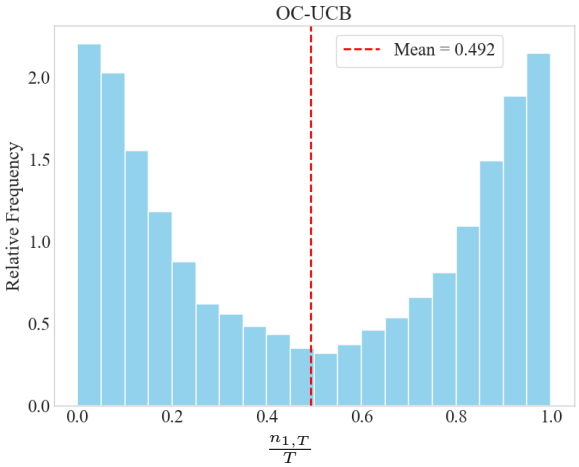


Figure 6: **Left**: Empirical results for ADA-UCB based on 5000 independent sample runs with $T = 10000$ pulls. **Right**: Coverage of Wald confidence interval of ADA-UCB with $T = 10000$.



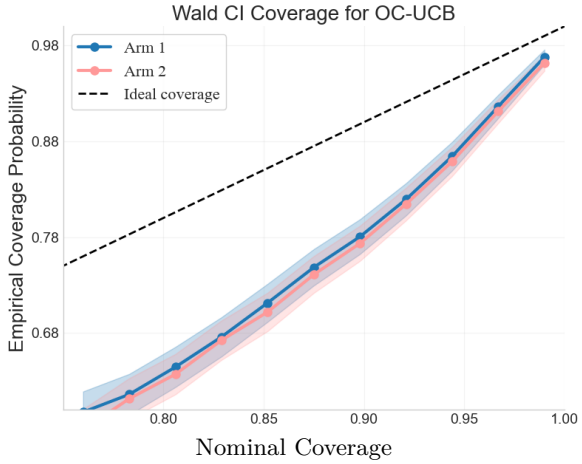
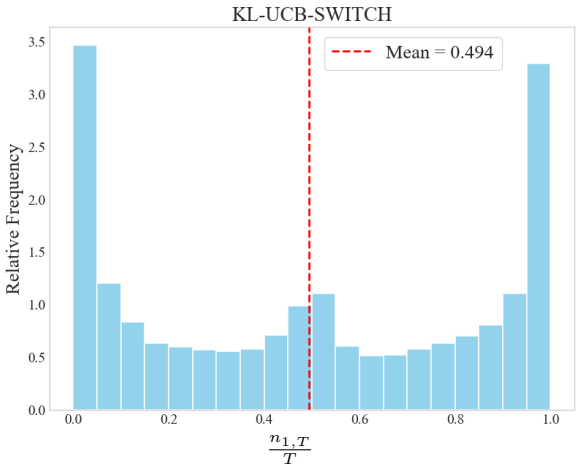


Figure 7: **Left**: Empirical results for OC-UCB based on 5000 independent sample runs with $T = 10000$ pulls. **Right**: Coverage of Wald confidence interval of OC-UCB with $T = 10000$. We have chosen $\epsilon = 0.1$.



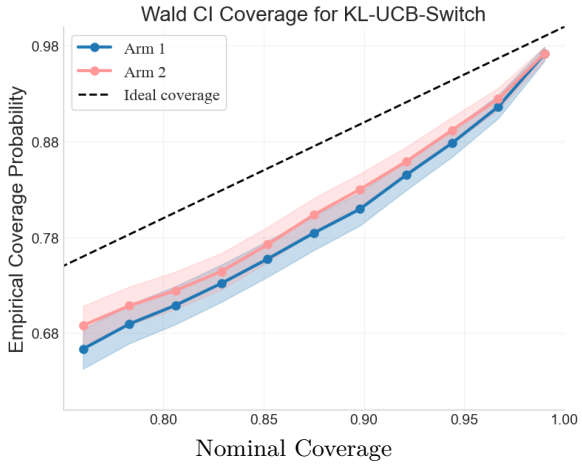
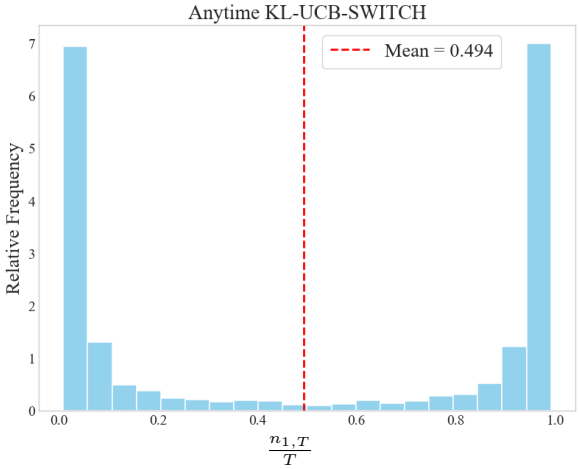


Figure 8: **Left**: Empirical results for KL-UCB-Switch based on 5000 independent sample runs with $T = 10000$ pulls. **Right**: Coverage of Wald confidence interval of KL-UCB-Switch with $T = 10000$.



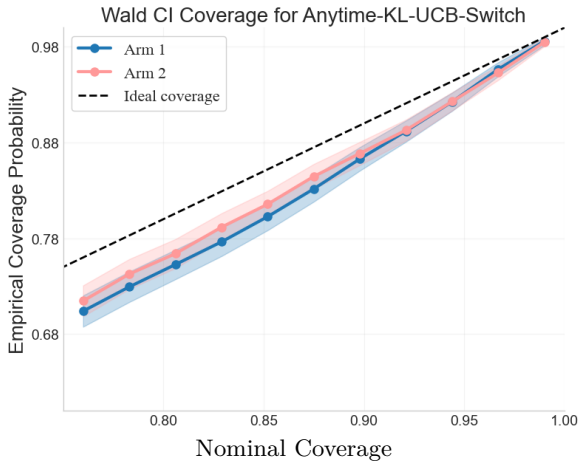


Figure 9: **Left**: Empirical results for Anytime-KL-UCB-Switch based on 5000 independent sample runs with $T = 10000$ pulls. **Right**: Coverage of Wald confidence interval of KL-UCB-Switch with $T = 10000$.

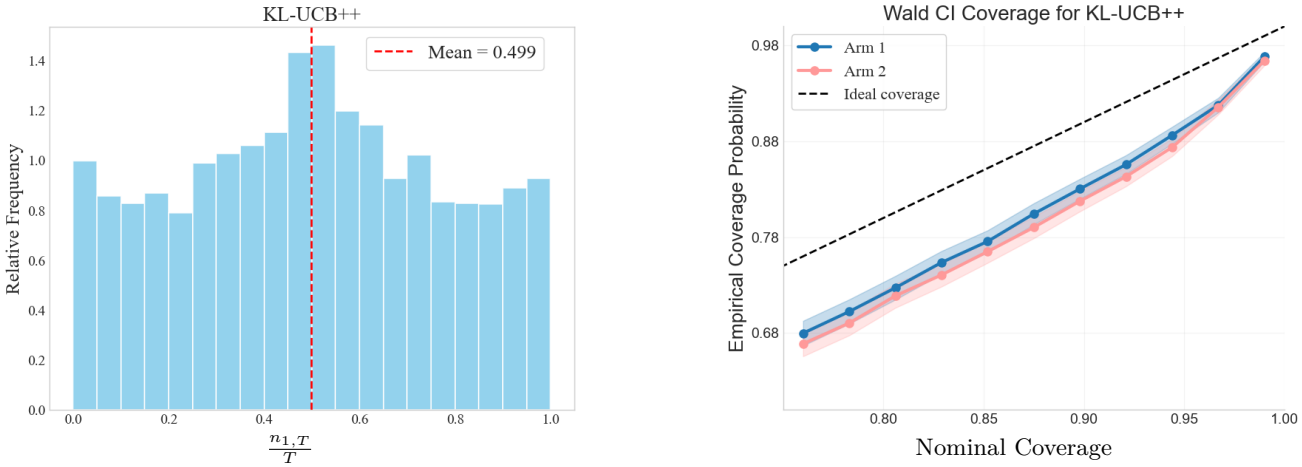


Figure 10: **Left**: Empirical results for KL-UCB++ based on 5000 independent sample runs with $T = 10000$ pulls. **Right**: Coverage of Wald confidence interval of KL-UCB++ with $T = 10000$.

5 Proof of Theorem 1

We prove this theorem by contradiction. We use $\mathcal{A}$ to denote a generic optimism rule (11). We prove Theorem 1 in two steps:

(a) In Lemma 2, we show that if an algorithm $\mathcal{A}$ is stable under the set up of Theorem 1, then $\mathcal{A}$ must be stable with $n^\star_{a,T} = T/K$ for all arm a.

(b) Next we argue that if algorithm $\mathcal{A}$ is stable in the first $T/2K$ rounds, then under Assumptions (A1), (A2) and (A3) the algorithm $\mathcal{A}$ must violate the stability condition (3) in later rounds ($t \geq T/2K$) with positive probability.

Lemma 2. *Under the set up of Theorem 1, any stable algorithm $\mathcal{A}$ must be stable with $n^\star_{a,T} = T/K$ for all arm $a \in [K]$.*

We prove Lemma 2 in Appendix D. Given this lemma at hand it suffices to prove that

$$\liminf_{T\to\infty} \mathrm{P}\left(\frac{n_{1,T}(\mathcal{A})}{T} \geq \frac{K+1}{2K},\ \frac{n_{a,T}(\mathcal{A})}{T} \leq \frac{1}{2K} \ \text{ for all } \ a \in \{2,\ldots,K\}\right) > 0. \tag{18}$$

Indeed, it is easy to see that the last bound violates the stability condition (3). Throughout we assume that $\beta_1, \beta_2 > 0$ in Assumption (A3). The same proof goes through with arbitrary reals β_1, β_2. Let the mean of all the arms be μ and define the boundary $\lambda_T := \mu - c/\sqrt{T}$ for some $c > 0$; we select the value of c at a later stage of the proof. Now consider the event

$$\mathcal{E}_0(T) := \left\{U_a\left(\frac{T}{2K}\right) < \lambda_T \ \forall a \in \{2,\ldots,K\},\ \min_{\frac{T}{2K}\leq t\leq T} U_1(t) > \lambda_T\right\}. \tag{19}$$

Observe that on the event $\{U_a(T/2K) < \lambda_T \ \text{ for all } \ a \geq 2 \ \text{ and } \ U_1(T/2K) > \lambda_T\}$, arm 1 is pulled at time $T/(2K) + 1$. Consequently, we have for all arms $a \geq 2$,

$$n_{a,\frac{T}{2K}+1}(\mathcal{A}) = n_{a,\frac{T}{2K}}(\mathcal{A}), \qquad \text{and} \qquad \widehat{\mu}_{a,\frac{T}{2K}+1} = \widehat{\mu}_{a,\frac{T}{2K}}.$$

Hence, it follows from Assumption (A2) that

$$U_a\Big(\frac{T}{2K}+1\Big) \le U_a\Big(\frac{T}{2K}\Big) < \lambda_T.$$

By applying this argument iteratively it follows that on event $\mathcal{E}_0(T)$ arm 1 is pulled at every step from $T/(2K)$ to T. Hence,

$$\mathcal{E}_0(T) \subset \left\{ \frac{n_{1,T}(\mathcal{A})}{T} \ge \frac{K+1}{2K},\ \frac{n_{a,T}(\mathcal{A})}{T} \le \frac{1}{2K}\ \forall a \ge 2 \right\}. \tag{20}$$

Now, let us define the event

$$\mathcal{E}(T) = \Big\{ n_{a,\frac{T}{2K}} \in \left[\frac{T}{4K^2}, \frac{3T}{4K^2} \right] \text{ for all } a \in [K] \Big\}.$$

Now suppose that an algorithm satisfying Assumptions (A1), (A2) and (A3) is *stable*. By Lemma 2 we have

$$\frac{n_{a,T/2K}}{T/2K^2} \xrightarrow{\mathrm{P}} 1$$

Consequently, we have $\mathrm{P}(\mathcal{E}(T)) \to 1$ as $T \to \infty$, and for any sequence of events $(A_T)_{T\ge 1}$,

$$\liminf_{T\to\infty} \mathrm{P}(A_T) = \liminf_{T\to\infty} \mathrm{P}(A_T \cap \mathcal{E}(T)). \tag{21}$$

In particular, for the choice of $A_T = \mathcal{E}_0(T)$ we obtain,

$$\liminf_{T\to\infty} \mathrm{P}(\mathcal{E}_0(T)) = \liminf_{T\to\infty} \mathrm{P}(\mathcal{E}_0(T) \cap \mathcal{E}(T)). \tag{22}$$

On the event $\mathcal{E}_0(T) \cap \mathcal{E}(T)$, Assumption (A3) implies that the upper confidence bounds $U_a(t)$ for all $a \in [K]$ lie between $\widetilde{U}_{a,\beta_1}(t)$ and $\widetilde{U}_{a,\beta_2}(t)$ where,

$$\widetilde{U}_{a,\beta}(t) = \widehat{\mu}_{a,t} + \frac{\beta}{\sqrt{n_{a,t}}}$$

Moreover, on $\mathcal{E}(T)$, there exist constants γ_1, γ_2 (depending on β_1, β_2) such that for each arm $a \in [K]$,

$$\widehat{\mu}_{a,t} + \frac{\gamma_1}{\sqrt{t}} \le \widetilde{U}_{a,\beta_1}(t) \le U_a(t) \le \widetilde{U}_{a,\beta_2}(t) \le \widehat{\mu}_{a,t} + \frac{\gamma_2}{\sqrt{t}}. \tag{23}$$

Motivated by this, we define the *pseudo upper confidence bound* below

$$V_{a,\gamma}(t) := \widehat{\mu}_{a,t} + \frac{\gamma}{\sqrt{t}}.$$

Consider the event

$$\mathcal{E}_1(T) := \left\{ V_{a,\gamma_2}\left(\frac{T}{2K}\right) < \lambda_T\ \forall a \in \{2,\dots,K\},\ \min_{\frac{T}{2K}+1 \le t \le T} V_{1,\gamma_1}(t) > \lambda_T \right\}. \tag{24}$$

Note that equation (23) and Assumption (A3) imply that

$$\mathcal{E}_1(T) \cap \mathcal{E}(T) \subset \mathcal{E}_0(T) \cap \mathcal{E}(T). \tag{25}$$

Therefore we obtain the following string of inequalities.

$$\begin{aligned}
&\liminf_{T\to\infty} \mathrm{P}\left(\frac{n_{1,T}(\mathcal{A})}{T} \geq \frac{K+1}{2K},\ \frac{n_{a,T}(\mathcal{A})}{T} \leq \frac{1}{2K} \ \text{ for all } \ a \in \{2,\dots,K\} \right) \\
&\overset{(i)}{\geq} \liminf_{T\to\infty} \mathrm{P}(\mathcal{E}_0(T)) \\
&\overset{(ii)}{=} \liminf_{T\to\infty} \mathrm{P}(\mathcal{E}_0(T) \cap \mathcal{E}(T)) \\
&\overset{(iii)}{\geq} \liminf_{T\to\infty} \mathrm{P}(\mathcal{E}_1(T) \cap \mathcal{E}(T)) \\
&\overset{(iv)}{=} \liminf_{T\to\infty} \mathrm{P}(\mathcal{E}_1(T)).
\end{aligned} \tag{26}$$

Inequalities (i) and (iii) follow from equation (20) and (25), whereas equalities (ii) and (iv) follow from equations (21) and (22). Now we apply the following lemma, which states that $\mathrm{P}(\mathcal{E}_1(T))$ can be written in terms of probabilities of three different events which we then analyze separately.

Lemma 3. *Assume that assumption (A1) holds. Then we have*

$$\mathrm{P}(\mathcal{E}_1(T)) = \frac{D_{1,T}}{D_{2,T}}\, D_{3,T}, \tag{27}$$

where

$$\begin{aligned}
D_{1,T} &= \mathrm{P}\left(V_{a,\gamma_2}\left(\frac{T}{2K}\right) < \lambda_T \ \forall a \geq 2,\ V_{1,\gamma_1}\left(\frac{T}{2K}\right) > \lambda_T,\ V_{1,\gamma_1}\left(\frac{T}{2K}+1\right) > \lambda_T \right), \\
D_{2,T} &= \mathrm{P}\left(V_{1,\gamma_1}\left(\frac{T}{2K}+1\right) > \lambda_T \right), \\
D_{3,T} &= \mathrm{P}\left(\min_{\frac{T}{2K}+1 \leq t \leq T} V_{1,\gamma_1}(t) > \lambda_T \right).
\end{aligned} \tag{28}$$

Lemma 3 enables us to obtain a non-degenerate lower bound on the term in equation (26) by controlling each of the sequences $D_{i,T}$ individually. This reduction is achieved using the following lemma.

Lemma 4. *Assume that Assumption (A3) holds and suppose $D_{1,T}, D_{2,T}$ and $D_{3,T}$ are as defined in equations (28). Then there exist positive constants $\kappa_1, \kappa_2, \kappa_3$ such that*

$$\liminf_{T\to\infty} D_{i,T} > \kappa_i \quad \textit{for each } i \in \{1,2,3\}. \tag{29}$$

We provide the proof of Lemmas 3 and 4 in Appendix A. This completes the proof.

6 Discussion

Stability is a sufficient and nearly necessary condition for a central limit theorem to hold in the multi-armed bandit problem. This paper establishes general and easily verifiable conditions under which an optimism-based K-armed bandit algorithm becomes unstable. These conditions cover a broad class of minimax-optimal UCB and KL-UCB type methods. In particular, we show that widely used minimax-optimal algorithms—including MOSS, Anytime-MOSS, Vanilla-MOSS, ADA-UCB, OC-UCB, KL-MOSS, KL-UCB++, KL-UCB-SWITCH, and Anytime KL-UCB-SWITCH—are unstable. In all these cases, we complement our theoretical guarantees with numerical simulations demonstrating that the sample means fail to exhibit asymptotic normality.

Our theoretical results and numerical experiments raise an interesting question: does there exist any optimism-based bandit algorithm that is both minimax optimal and stable? Additional directions include quantifying how instability depends on the reward gap Δ, and determining whether modifications of MOSS or related minimax strategies could achieve stability under alternative model assumptions. Overall, our findings highlight the need for new adaptive data collection mechanisms that preserve stability while retaining optimal regret performance.

Acknowledgment. This work was partially supported by the National Science Foundation Grant DMS-2311304 to Koulik Khamaru.

7 References

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

David J Aldous and Geoffrey Kennedy Eagleson. On mixing and stability of limit theorems. *The Annals of Probability*, pages 325–331, 1978.

Julyan Arbel, Olivier Marchal, and Hien D Nguyen. On strict sub-gaussianity, optimal proxy variance and symmetry for bounded random variables. *ESAIM: Probability and Statistics*, 24: 39–55, 2020.

Jean-Yves Audibert and Sébastien Bubeck. Minimax policies for adversarial and stochastic bandits. In *COLT*, pages 217–226, 2009.

Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19):1876–1902, 2009.

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2):235–256, 2002.

Hamsa Bastani and Mohsen Bayati. Online decision making with high-dimensional covariates. *Operations Research*, 68(1):276–294, 2020.

Hamsa Bastani, Osbert Bastani, and Wichinpong Park Sinchaisri. Improving human sequential decision making with reinforcement learning. *Management Science*, 2025.

Donald A Berry. Adaptive clinical trials in oncology. *Nature reviews Clinical oncology*, 9(4):199–207, 2012.

Djallel Bouneffouf, Amel Bouzeghoub, and Alda Lopes Gançarski. A contextual-bandit algorithm for mobile context-aware recommender system. In *International conference on neural information processing*, pages 324–331. Springer, 2012.

Djallel Bouneffouf, Amel Bouzeghoub, and Alda Lopes Gançarski. Contextual bandits for context-based information retrieval. In *International Conference on Neural Information Processing*, pages 35–42. Springer, 2013.

Ningyuan Chen, Xuefeng Gao, and Yi Xiong. Debiasing samples from online learning using bootstrap. In *International Conference on Artificial Intelligence and Statistics*, pages 8514–8533. PMLR, 2022.

Yilun Chen and Jiaqi Lu. A characterization of sample adaptivity in ucb data. *arXiv preprint arXiv:2503.04855*, 2025.

Shein-Chung Chow, Mark Chang, and Annpey Pong. Statistical consideration of adaptive methods in clinical development. *Journal of biopharmaceutical statistics*, 15(4):575–591, 2005.

Victor H de la Pena, Michael J Klass, and Tze Leung Lai. Self-normalized processes: exponential inequalities, moment bounds and iterated logarithm laws. 2004.

Victor H De la Pena, Tze Leung Lai, and Qi-Man Shao. *Self-normalized processes: Limit theory and Statistical Applications*. Springer, 2009.

Yash Deshpande, Lester Mackey, Vasilis Syrgkanis, and Matt Taddy. Accurate inference for adaptive linear models. In *International Conference on Machine Learning*, pages 1194–1203. PMLR, 2018.

Yash Deshpande, Adel Javanmard, and Mohammad Mehrabi. Online debiasing for adaptively collected high-dimensional data with applications to time series analysis. *Journal of the American Statistical Association*, 118(542):1126–1139, 2023.

David A Dickey and Wayne A Fuller. Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American statistical association*, 74(366a):427–431, 1979.

Cyrille Dubarry. Confidence intervals for ab-test. *arXiv preprint arXiv:1501.07768*, 2015.

Rick Durrett. *Probability: theory and examples*, volume 49. Cambridge university press, 2019.

Aryeh Dvoretzky. Asymptotic normality for sums of dependent random variables. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability Theory*, volume 6, pages 513–536. University of California Press, 1972.

Lin Fan and Peter W Glynn. The typical behavior of bandit algorithms. *arXiv preprint arXiv:2210.05660*, 2022.

Yingying Fan, Yuxuan Han, Jinchi Lv, Xiaocong Xu, and Zhengyuan Zhou. Precise asymptotics and refined regret of variance-aware ucb. *arXiv preprint arXiv:2412.08843*, 2024.

Sarah Filippi. Optimistic strategies in reinforcement learning. *HAL*, 2010, 2010.

Aurélien Garivier and Olivier Cappé. The kl-ucb algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th annual conference on learning theory*, pages 359–376. JMLR Workshop and Conference Proceedings, 2011.

Aurélien Garivier, Tor Lattimore, and Emilie Kaufmann. On explore-then-commit strategies. *Advances in Neural Information Processing Systems*, 29, 2016.

Aurélien Garivier, Pierre Ménard, and Gilles Stoltz. Explore first, exploit next: The true shape of regret in bandit problems. *Mathematics of Operations Research*, 44(2):377–399, 2019.

Aurélien Garivier, Hédi Hadiji, Pierre Menard, and Gilles Stoltz. Kl-ucb-switch: optimal regret bounds for stochastic bandits from both a distribution-dependent and a distribution-free viewpoints. *Journal of Machine Learning Research*, 23(179):1–66, 2022.

Aurélien Garivier and Olivier Cappé. The kl-ucb algorithm for bounded stochastic bandits and beyond. In Sham M. Kakade and Ulrike von Luxburg, editors, *Proceedings of the 24th Annual Conference on Learning Theory*, volume 19 of *Proceedings of Machine Learning Research*, pages 359–376, Budapest, Hungary, 09–11 Jun 2011. PMLR. URL https://proceedings.mlr.press/v19/garivier11a.html.

Vitor Hadad, David A Hirshberg, Ruohan Zhan, Stefan Wager, and Susan Athey. Confidence intervals for policy evaluation in adaptive experiments. *Proceedings of the national academy of sciences*, 118(15):e2014602118, 2021.

Budhaditya Halder, Shubhayan Pan, and Koulik Khamaru. Stable thompson sampling: Valid inference via variance inflation. *arXiv preprint arXiv:2505.23260*, 2025.

Peter Hall and Christopher C Heyde. *Martingale limit theory and its application*. Academic press, 2014.

Qiyang Han, Koulik Khamaru, and Cun-Hui Zhang. Ucb algorithms for multi-armed bandits: Precise regret and adaptive inference. *arXiv preprint arXiv:2412.06126*, 2024.

Steven R Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform chernoff bounds via nonnegative supermartingales. 2020.

Xiaoguang Huo and Feng Fu. Risk-aware multi-armed bandit problem with application to portfolio selection. *Royal Society open science*, 4(11):171377, 2017.

Anand Kalvit and Assaf Zeevi. A closer look at the worst-case behavior of multi-armed bandit algorithms. *Advances in Neural Information Processing Systems*, 34:8807–8819, 2021.

Michael N Katehakis and Herbert Robbins. Sequential choice from several populations. *Proceedings of the National Academy of Sciences*, 92(19):8584–8585, 1995.

Koulik Khamaru and Cun-Hui Zhang. Inference with the upper confidence bound algorithm. *arXiv preprint arXiv:2408.04595*, 2024.

Koulik Khamaru, Yash Deshpande, Tor Lattimore, Lester Mackey, and Martin J Wainwright. Near-optimal inference in adaptive linear regression. *arXiv preprint arXiv:2107.02266*, 2021.

Gi-Soo Kim and Myunghee Cho Paik. Doubly-robust lasso bandit. *Advances in Neural Information Processing Systems*, 32, 2019.

Wonyoung Kim, Kyungbok Lee, and Myunghee Cho Paik. Double doubly robust thompson sampling for generalized linear contextual bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 8300–8307, 2023.

Akshay Krishnamurthy, Zhiwei Steven Wu, and Vasilis Syrgkanis. Semiparametric contextual bandits. In *International Conference on Machine Learning*, pages 2776–2785. PMLR, 2018.

Tze Leung Lai. Adaptive treatment allocation and the multi-armed bandit problem. *The annals of statistics*, pages 1091–1114, 1987.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

Tze Leung Lai and Ching Zong Wei. Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *The Annals of Statistics*, pages 154–166, 1982.

Tor Lattimore. Optimally confident ucb: Improved regret for finite-armed bandits. *arXiv preprint arXiv:1507.07880*, 2015.

Tor Lattimore. Refining the confidence level for optimistic bandit strategies. *Journal of Machine Learning Research*, 19(20):1–32, 2018.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

James Leiner, Robin Dunn, and Aaditya Ramdas. Adaptive off-policy inference for m-estimators under model misspecification. *arXiv preprint arXiv:2509.14218*, 2025.

Fred C Leone, Lloyd S Nelson, and RB Nottingham. The folded normal distribution. *Technometrics*, 3(4):543–550, 1961.

Licong Lin, Mufang Ying, Suvrojit Ghosh, Koulik Khamaru, and Cun-Hui Zhang. Statistical limits of adaptive linear models: low-dimensional estimation and inference. *Advances in Neural Information Processing Systems*, 36:15965–15990, 2023.

Licong Lin, Koulik Khamaru, and Martin J Wainwright. Semiparametric inference based on adaptively collected data. *The Annals of Statistics*, 53(3):989–1014, 2025.

Odalric-Ambrym Maillard, Rémi Munos, and Gilles Stoltz. A finite-time analysis of multi-armed bandits problems with kullback-leibler divergences. In *Proceedings of the 24th annual Conference On Learning Theory*, pages 497–514. JMLR Workshop and Conference Proceedings, 2011.

Ian C Marschner. Confidence distributions for treatment effects in clinical trials: Posteriors without priors. *Statistics in Medicine*, 43(6):1271–1289, 2024.

Pierre Ménard and Aurélien Garivier. A minimax and asymptotically optimal algorithm for stochastic bandits. In *International Conference on Algorithmic Learning Theory*, pages 223–237. PMLR, 2017.

Yash Nair and Lucas Janson. Randomization tests for adaptively collected data. *arXiv preprint arXiv:2301.05365*, 2023.

Weiwei Shen, Jun Wang, Yu-Gang Jiang, and Hongyuan Zha. Portfolio choices with orthogonal bandit learning. In *IJCAI*, volume 15, pages 974–980, 2015.

David Siegmund. *Sequential analysis: tests and confidence intervals*. Springer Science & Business Media, 2013.

James Supancic III and Deva Ramanan. Tracking as online decision-making: Learning a policy from streaming videos with reinforcement learning. In *Proceedings of the IEEE international conference on computer vision*, pages 322–331, 2017.

Anna L Trella, Kelly W Zhang, Inbal Nahum-Shani, Vivek Shetty, Finale Doshi-Velez, and Susan A Murphy. Reward design for an online reinforcement learning algorithm supporting oral self-care. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 15724–15730, 2023.

Anna L Trella, Kelly W Zhang, Stephanie M Carpenter, David Elashoff, Zara M Greer, Inbal Nahum-Shani, Dennis Ruenger, Vivek Shetty, and Susan A Murphy. Oralytics reinforcement learning algorithm. *arXiv preprint arXiv:2406.13127*, 2024.

Anna L Trella, Kelly W Zhang, Hinal Jajal, Inbal Nahum-Shani, Vivek Shetty, Finale Doshi-Velez, and Susan A Murphy. A deployed online reinforcement learning algorithm in an oral health clinical trial. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 28792–28800, 2025.

Sharan Vaswani, Branislav Kveton, Zheng Wen, Mohammad Ghavamzadeh, Laks VS Lakshmanan, and Mark Schmidt. Model-independent online learning for influence maximization. In *International conference on machine learning*, pages 3530–3539. PMLR, 2017.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Abraham Wald. Sequential tests of statistical hypotheses. In *Breakthroughs in statistics: Foundations and basic theory*, pages 256–298. Springer, 1992.

Ian Waudby-Smith, Lili Wu, Aaditya Ramdas, Nikos Karampatziakis, and Paul Mineiro. Anytime-valid off-policy inference for contextual bandits. *ACM/IMS Journal of Data Science*, 1(3):1–42, 2024.

Zheng Wen, Branislav Kveton, Michal Valko, and Sharan Vaswani. Online influence maximization under independent cascade model with semi-bandit feedback. *Advances in neural information processing systems*, 30, 2017.

John S White. The limiting distribution of the serial correlation coefficient in the explosive case. *The Annals of Mathematical Statistics*, pages 1188–1197, 1958.

John S White. The limiting distribution of the serial correlation coefficient in the explosive case ii. *The Annals of Mathematical Statistics*, pages 831–834, 1959.

Cun-Hui Zhang and Stephanie S Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(1):217–242, 2014.

Kelly Zhang, Lucas Janson, and Susan Murphy. Inference for batched bandits. *Advances in neural information processing systems*, 33:9818–9829, 2020.

Kelly W Zhang, Lucas Janson, and Susan A Murphy. Statistical inference after adaptive sampling for longitudinal data. *arXiv preprint arXiv:2202.07098*, 2022.

Kelly W Zhang, Nowell Closser, Anna L Trella, and Susan A Murphy. Replicable bandits for digital health interventions. *arXiv preprint arXiv:2407.15377*, 2024.

Qian Zhou, XiaoFang Zhang, Jin Xu, and Bin Liang. Large-scale bandit approaches for recommender systems. In *International Conference on Neural Information Processing*, pages 811–821. Springer, 2017.

A Proofs of Important Lemmas

In this section, we provide the proofs of Lemma 3 and 4 which are the key ingredients in the proof of Theorem 1 in Section 5.

Proof of Lemma 3

We begin the proof by fixing two arbitrary constant real numbers γ_1 and γ_2. For simplicity, we first prove the theorem for 2-arms and then generalize it for the K-arm case.

2-arm case:

Consider the following process:

$$\mathbf{Z}_t(\gamma_1, \gamma_2) := \begin{bmatrix} V_{1,\gamma_1}(t) \\ V_{2,\gamma_2}(t) \\ \widehat{\mu}_{1,t} \\ \widehat{\mu}_{2,t} \\ n_{1,t} \\ n_{2,t} \end{bmatrix} \tag{30}$$

For better readibility of the proofs, we use the shorthand Z_t for $Z_t(\gamma_1, \gamma_2)$. To see why the introduction of this new process is advantageous, let us consider two projection maps g_1 and g_2 on $\mathbb{R}$, which are defined below.

$$g_1(\mathbf{z}) := \mathbf{e}_1^T \mathbf{z} \quad \text{and} \quad g_2(\mathbf{z}) := \mathbf{e}_2^T \mathbf{z}$$

where $\mathbf{e}_i^T$ is the i^{th} orthonormal basis vector of the standard basis in $\mathbb{R}^6$.

Now, let us define sets A_T, B_T, C_T such that $A_T := g_2^{-1}((-\infty, \lambda_T))$, $C_T := g_1^{-1}((\lambda_T, \infty))$ and $B_T := A_T \bigcap C_T$. Using these sets, we can rewrite the following three sets of our interest in terms of the process $(\mathbf{Z}_t)_{t\geq 1}$:

$$\begin{aligned} &\{V_{1,\gamma_1}(t) > \lambda_T\} \equiv \{\mathbf{Z}_t \in C_T\}, \\ &\{V_{2,\gamma_2}(t) < \lambda_T\} \equiv \{\mathbf{Z}_t \in A_T\} \quad \text{and,} \\ &\{V_{2,\gamma_2}(t) < \lambda_T < V_{1,\gamma_1}(t)\} \equiv \{\mathbf{Z}_t \in B_T\} \end{aligned} \tag{31}$$

This reformulation allows us to rewrite the event $\mathcal{E}_1(T)$ of our interest in the following form:

$$\begin{aligned} \mathcal{E}_1(T) &\equiv \left\{ V_{2,\gamma_2}\left(\frac{T}{4}\right) < \lambda_T,\ V_{1,\gamma_1}\left(\frac{T}{4}\right) > \lambda_T, \min_{\frac{T}{4}+2 \leq t \leq T} V_{1,\gamma_1}\left(\frac{T}{4}\right) > \lambda_T \right\} \\ &= \left\{ \mathbf{Z}_{\frac{T}{4}} \in B_T,\ \mathbf{Z}_{\frac{T}{4}+1} \in C_T, \mathbf{Z}_t \in C_T \ \ \forall \ \ t \in \left[\frac{T}{4}+2, T\right] \right\} \end{aligned} \tag{32}$$

The decomposition (32) is useful because of a Markov property of the process $\{\mathbf{Z}_t\}_{t\geq 0}$ which is a Markov chain. We now formally state these Markov properties below:

Lemma 5. *Suppose* $\{\boldsymbol{Z}_t(\gamma_1,\gamma_2)\}_{t\geq 1}$ *is the process defined in equation* (30). *Then under the assumption (A1) it is a discrete time Markov chain, where* $n_{a,t}$ *is the number of times arm* a *has been pulled up to time* t, $\widehat{\mu}_{a,t}$ *and* $U_a(t)$, *respectively, denote the sample mean and the UCB of arm* a *at time* t.

Lemma 5 allows us to apply Markov property which states that given the present, past and future events are independent (Durrett (2019)). Concretely,

Lemma 6. *Suppose that* $(\boldsymbol{Q}_t)_{t\geq 1}$ *is a finite-dimensional, discrete time Markov Chain. Let* A, B *and* C *be arbitrary sets comprising of states in the state space. Then, we have the following equality holds for any fixed value of* m*:*

$$\begin{aligned}&\mathrm{P}\left(\boldsymbol{Q}_{t-1}\in A, \boldsymbol{Q}_{t+i}\in C, i=1,\ldots,m \;\;\middle|\;\; \boldsymbol{Q}_t\in B\right)\\ &=\mathrm{P}\left(\boldsymbol{Q}_{t+i}\in C, i=1,\ldots,m\middle|\;\boldsymbol{Q}_t\in B\right)\times \mathrm{P}\left(\boldsymbol{Q}_{t-1}\in A \;\middle|\; \boldsymbol{Q}_t\in B\right)\end{aligned}$$

We prove Lemmas 5 and 6 in Appendix D. Applying Lemma 6 with $\mathbf{Q}_t=\mathbf{Z}_t$, $A=B_T, B=C_T, C=C_T$ and $t=\frac{T}{4}+1$ yields

$$\begin{aligned}&=\mathrm{P}\left(\mathbf{Z}_{\frac{T}{4}}\in B_T,\;\mathbf{Z}_{\frac{T}{4}+1}\in C_T,\;\mathbf{Z}_t\in C_T\;\forall t\;\in\left[\frac{T}{4}+2,T\right]\right)\\
&=\mathrm{P}\left(\mathbf{Z}_{\frac{T}{4}}\in B_T,\mathbf{Z}_t\in C_T\;\forall\; t\in\left[\frac{T}{4}+2,T\right]\;\middle|\;\mathbf{Z}_{\frac{T}{4}+1}\in C_T\right)\times\mathrm{P}\left(\mathbf{Z}_{\frac{T}{4}+1}\in C_T\right)\\
&\overset{(i)}{=}\mathrm{P}\left(\mathbf{Z}_{\frac{T}{4}}\in B_T\;\middle|\;\mathbf{Z}_{\frac{T}{4}+1}\in C_T\right)\times\mathrm{P}\left(\mathbf{Z}_{\frac{T}{4}+1}\in C_T\right)\\
&\times\mathrm{P}\left(\mathbf{Z}_t\in C_T\;\forall\; t\in\left[\frac{T}{4}+2,T\right]\;\middle|\;\mathbf{Z}_{\frac{T}{4}+1}\in C_T\right)\\
&=\frac{\mathrm{P}\left(\mathbf{Z}_{\frac{T}{4}}\in B_T,\mathbf{Z}_{\frac{T}{4}+1}\in C_T\right)}{\mathrm{P}\left(\mathbf{Z}_{\frac{T}{4}+1}\in C_T\right)}\times\mathrm{P}\left(\mathbf{Z}_t\in C_T\;\forall\; t\in\left[\frac{T}{4}+1,T\right]\right)\\
&=\frac{D_{1,T}}{D_{2,T}}\;D_{3,T}\end{aligned}$$

where equality (i) follows from Lemma 6. In the last equality, we have defined three real sequences $D_{1,T}, D_{2,T}$ and $D_{3,T}$ which are as follows:

$$D_{1,T}=\mathrm{P}\left(\mathbf{Z}_{\frac{T}{4}}\in B_T,\;\mathbf{Z}_{\frac{T}{4}+1}\in C_T\right)=\mathrm{P}\left(V_{2,\gamma_2}\left(\frac{T}{4}\right)<\lambda_T,\;\min_{\frac{T}{4}\leq t\leq\frac{T}{4}+1}V_{1,\gamma_1}(t)>\lambda_T\right)$$

$$D_{2,T}=\mathrm{P}\left(\mathbf{Z}_{\frac{T}{4}+1}\in C_T\in C_T\right)=\mathrm{P}\left(V_{1,\gamma_1}\left(\frac{T}{4}+1\right)>\lambda_T\right)$$

$$D_{3,T}=\mathrm{P}\left(\mathbf{Z}_t\in C_T,\;\forall\; t\in\left[\frac{T}{4}+1,T\right]\right)=\mathrm{P}\left(\min_{\frac{T}{4}+1\leq t\leq T}V_{1,\gamma_1}(t)>\lambda_T\right)$$

K-arm case:

The extension to the general K arm case follows analogously by defining the following vectors:

$$\mathbf{V}_t(\gamma) := \begin{bmatrix} V_{1,\gamma_1}(t) \\ \dots \\ V_{K,\gamma_K}(t) \end{bmatrix}, \quad \widehat{\mu}_t(\gamma) := \begin{bmatrix} \widehat{\mu}_{1,t} \\ \dots \\ \widehat{\mu}_{K,t} \end{bmatrix}, \quad \text{and} \quad n_t(\gamma) := \begin{bmatrix} n_{1,t} \\ \dots \\ n_{K,t} \end{bmatrix}$$

Now the sane steps of the proof follow by considering the projection maps on the following Markov Chain in the extended state space:

$$\mathbf{Z}_t(\gamma) := \begin{bmatrix} \mathbf{V}_{1,\gamma_1}(t) \\ \widehat{\mu}_t(\gamma) \\ n_t(\gamma) \end{bmatrix}$$

The proof of the K-arm case is now immediate using a similar argument as the 2-arm case. with the following choices of $D_{1,T}, D_{2,T}$ and $D_{3,T}$.

$$D_{1,T} = \mathrm{P}\left(V_{a,\gamma_2}\left(\frac{T}{2K}\right) < \lambda_T \ \forall a \in \{2,\dots,K\},\ V_{1,\gamma_1}\left(\frac{T}{2K}\right) > \lambda_T, V_{1,\gamma_1}\left(\frac{T}{2K}+1\right) > \lambda_T \right)$$

$$D_{2,T} = \mathrm{P}\left(V_{1,\gamma_1}\left(\frac{T}{2K}\right) > \lambda_T \right)$$

$$D_{3,T} = \mathrm{P}\left(\min_{\frac{T}{2K}+1 \le t \le T} V_{1,\gamma_1}(t) > \lambda_T \right)$$

This completes the proof.

Proof of Lemma 4

Before moving on to the proof of Lemma 4, we start a high level proof ideas and recall a few definitions.

Notations and high-level proof sketch: In this section, we prove that each of the real sequences defined before in equation (28) are lower bounded (in the limit) by a positive constant. For simplicity, we provide a detailed proof for the two arms case. The proof for the general K arm case follows using a similar argument. Before we begin the proof, let us recall a few key observations which shall be used repeatedly. Invoking Assumption (A3), we note that on the event $\{n_{a,\frac{T}{4}} \in [T/16, 3T/16],\ \text{for}\ \ a = 1,2\}$ and for any $t \ge T/4$ we have:

$$\widehat{\mu}_{a,t} + \frac{\gamma_1}{\sqrt{t}} \le\ U_a(t)\ \le\ \widehat{\mu}_{a,t} + \frac{\gamma_2}{\sqrt{t}} \tag{33}$$

Without loss of generality, let us assume that both γ_1 and γ_2 are positive as the same proof goes through for any arbitrary real numbers. Furthermore, for a given value of γ define a *pseudo upper confidence bounds* $V_{a,\gamma}(t)$ for each arm a as:

$$V_{a,\gamma}(t) := \widehat{\mu}_{a,t} + g_\gamma(t) \tag{34}$$

The main idea of each of the proofs is that under stability, one can sandwich $U_a(t)$ by the two processes $V_{a,\gamma_1}(t)$ and $V_{a,\gamma_2}(t)$ with high probability for large T. Finally, we recall the important events:

$$\mathcal{E}(T) = \Big\{ n_{a,\frac{T}{2K}} \in \left[\frac{T}{4K^2}, \frac{3T}{4K^2} \right] \text{ for all } a \in [K] \Big\}. \tag{35a}$$

$$\mathcal{E}_1(T) := \left\{ V_{a,\gamma_2}\left(\frac{T}{2K}\right) < \lambda_T \ \forall a \in \{2,\dots,K\}, \ \min_{\frac{T}{2K}+1 \le t \le T} V_{1,\gamma_1}(t) > \lambda_T \right\}. \tag{35b}$$

Main argument:

Without loss of generality we assume $\mu = 0$. We are now ready to prove the following three claims from Lemma 4:

$$\liminf_{T\to\infty} D_{1,T} > 0, \tag{36a}$$

$$\liminf_{T\to\infty} D_{2,T} > 0, \tag{36b}$$

$$\liminf_{T\to\infty} D_{3,T} > 0. \tag{36c}$$

Proof of claim (36a):

Consider the event

$$\Big\{ V_{2,\gamma_2}\left(\frac{T}{4}\right) < \lambda_T, \ V_{1,\gamma_1}\left(\frac{T}{4}\right) > \lambda_T \Big\} \bigcap \mathcal{E}(T). \tag{37}$$

On this event, it follows that at step $\frac{T}{4}+1$, arm 1 is pulled and the empirical mean of arm 1 is updated as

$$\widehat{\mu}_{1,\frac{T}{4}+1} = \frac{n_{1,\frac{T}{4}}}{n_{1,\frac{T}{4}}+1}\widehat{\mu}_{1,\frac{T}{4}} + \frac{X_{\frac{T}{4}+1}}{n_{1,\frac{T}{4}}+1}.$$

Under the event $\{V_{2,\frac{T}{4}}(\gamma_2) < \lambda_T, V_{1,\frac{T}{4}}(\gamma_1) > \lambda_T\} \cap \mathcal{E}(T)$ we obtain the following

$$\begin{aligned} V_{1,\gamma_1}\left(\frac{T}{4}+1\right) &= \widehat{\mu}_{1,\frac{T}{4}+1} + \frac{\gamma_2}{\sqrt{\frac{T}{4}+1}} \\ &= \frac{n_{1,\frac{T}{4}}}{n_{1,\frac{T}{4}}+1}\widehat{\mu}_{1,\frac{T}{4}} + \frac{X_{\frac{T}{4}+1}}{n_{1,\frac{T}{4}}+1} + \frac{\gamma_2}{\sqrt{\frac{T}{4}+1}} \end{aligned} \tag{38}$$

Therefore from equation (38) we observe that on event $\mathcal{E}(T)$

$$\begin{aligned} V_{1,\gamma_1}\left(\frac{T}{4}+1\right) &= \frac{n_{1,\frac{T}{4}}}{n_{1,\frac{T}{4}}+1}\widehat{\mu}_{1,\frac{T}{4}} + \frac{X_{\frac{T}{4}+1}}{n_{1,\frac{T}{4}}+1} + \frac{\gamma_1}{\sqrt{\frac{T}{4}+1}} \\ &= \frac{n_{1,\frac{T}{4}}}{n_{1,\frac{T}{4}}+1}V_{1,\gamma_1}\left(\frac{T}{4}\right) + \frac{X_{\frac{T}{4}+1}}{n_{1,\frac{T}{4}}+1} + g_{\gamma_1}\left(\frac{T}{4}+1\right) - \frac{n_{1,\frac{T}{4}}}{n_{1,\frac{T}{4}+1}+1}g_{\gamma_1}\left(\frac{T}{4}\right) \\ &= \frac{n_{1,\frac{T}{4}}}{n_{1,\frac{T}{4}}+1}V_{1,\gamma_1}\left(\frac{T}{4}\right) + \frac{X_{\frac{T}{4}+1}}{n_{1,\frac{T}{4}}+1} + \phi(T) \end{aligned} \tag{39}$$

where

$$\phi(T) := g_{\gamma_1}\left(\frac{T}{4}+1\right) - \frac{n_{1,\frac{T}{4}}}{n_{1,\frac{T}{4}+1}+1}g_{\gamma_1}\left(\frac{T}{4}\right)$$

By substituting the expression (39) under the joint event $\{V_{2,\gamma_2}\left(\frac{T}{4}\right) < \lambda_T, V_{1,\gamma_1}\left(\frac{T}{4}\right) > \lambda_T\} \cap \mathcal{E}(T)$, we can write

$$\begin{aligned} &\mathrm{P}\Big(V_{2,\gamma_2}\left(\frac{T}{4}\right) < \lambda_T,\ V_{1,\gamma_1}\left(\frac{T}{4}\right) > \lambda_T,\ V_{1,\gamma_1}\left(\frac{T}{4}+1\right) > \lambda_T,\ \mathcal{E}(T)\Big) \\ &= \mathrm{P}\Big(V_{2,\gamma_2}\left(\frac{T}{4}\right) < \lambda_T,\ V_{1,\gamma_1}\left(\frac{T}{4}\right) > \lambda_T,\ \frac{n_{1,\frac{T}{4}}}{n_{1,\frac{T}{4}}+1}V_{1,\gamma_1}\left(\frac{T}{4}\right) + \frac{X_{\frac{T}{4}}}{n_{1,\frac{T}{4}}+1} + \phi(T) > \lambda_T,\ \mathcal{E}(T)\Big) \end{aligned}$$

The last inequality follows from equations (39). To simplify further, we note the equivalences

$$\Big\{V_{1,\gamma_1}\left(\frac{T}{4}\right) > \lambda_T\Big\} = \Big\{\frac{n_{1,\frac{T}{4}}}{n_{1,\frac{T}{4}}+1}V_{1,\gamma_1}\left(\frac{T}{4}\right) > \lambda_T\frac{n_{1,\frac{T}{4}}}{n_{1,\frac{T}{4}}+1}\Big\},$$

$$\Big\{\frac{n_{1,\frac{T}{4}}}{n_{1,\frac{T}{4}}+1}\lambda_T + \frac{X_{\frac{T}{4}+1}}{n_{1,\frac{T}{4}}+1} > \lambda_T - \phi(T)\Big\} = \Big\{X_{\frac{T}{4}+1} > \lambda_T - (n_{1,\frac{T}{4}}+1)\phi(T)\Big\},$$

Therefore we have,

$$\begin{aligned} &\liminf_{T\to\infty} \mathrm{P}\Big(V_{2,\gamma_2}\left(\frac{T}{4}\right) < \lambda_T,\ V_{1,\gamma_1}\left(\frac{T}{4}\right) > \lambda_T,\ V_{1,\gamma_1}\left(\frac{T}{4}+1\right) > \lambda_T,\ \mathcal{E}(T)\Big) \\ &= \liminf_{T\to\infty} \mathrm{P}\Big(V_{2,\gamma_2}\left(\frac{T}{4}\right) < \lambda_T,\ V_{1,\gamma_1}\left(\frac{T}{4}\right) > \lambda_T,\ X_{\frac{T}{4}+1} > \lambda_T + (n_{1,\frac{T}{4}}+1)\phi(T),\ \mathcal{E}(T)\Big). \end{aligned}$$

Now we shall show that the correction term $(n_{1,\frac{T}{4}+1}+1)\phi(T)$ is asymptotically negligible. Evaluating it explicitly gives

$$\begin{aligned} (n_{1,\frac{T}{4}}+1)\phi(T) &= n_{1,\frac{T}{4}}\left[g_{\gamma_1}\left(\frac{T}{4}+1\right) - g_{\gamma_1}\left(\frac{T}{4}\right)\right] + g_{\gamma_1}\left(\frac{T}{4}+1\right) \\ &= -\frac{4n_{1,\frac{T}{4}}}{\sqrt{T(T+4)}}\frac{2\gamma_1}{\sqrt{T}+\sqrt{T+4}} + \frac{2\gamma_1}{\sqrt{T+4}}. \end{aligned}$$

Now on the event $\mathcal{E}(T)$, we observe that,

$$-(n_{1,\frac{T}{4}}+1)\phi(T) \leq \frac{3}{4}\frac{T}{\sqrt{T(T+4)}}\frac{2|\gamma_1|}{\sqrt{T}+\sqrt{T+4}} + \frac{2|\gamma_1|}{\sqrt{T+4}}.$$

Define,

$$r_T := \frac{3}{4}\frac{T}{\sqrt{T(T+4)}}\frac{2|\gamma_1|}{\sqrt{T}+\sqrt{T+4}} + \frac{2|\gamma_1|}{\sqrt{T+4}}$$

Therefore, on the event $\mathcal{E}(T)$, $-(n_{1,\frac{T}{4}}+1)\phi(T) \leq r_T$ and r_T is a real sequence converging to 0 as $T \to \infty$. This implies that on the joint event $\Big\{V_{2,\gamma_2}\left(\frac{T}{4}\right) < \lambda_T,\ V_{1,\gamma_1}\left(\frac{T}{4}\right) > \lambda_T,\ \mathcal{E}(T)\Big\}$,

$$\Big\{X_{\frac{T}{4}+1} > \lambda_T + r_T\Big\} \subseteq \Big\{X_{\frac{T}{4}+1} > \lambda_T - (n_{1,\frac{T}{4}}+1)\phi(T)\Big\}$$

Substituting back and applying the stability property (21), we finally obtain

$$\begin{aligned}
&\liminf_{T\to\infty} \mathrm{P}\Big(V_{2,\gamma_2}\left(\frac{T}{4}\right) < \lambda_T,\ V_{1,\gamma_1}\left(\frac{T}{4}\right) > \lambda_T,\ X_{\frac{T}{4}+1} > \lambda_T - (n_{1,\frac{T}{4}}+1)\phi(T),\ \mathcal{E}(T)\Big) \\
&\geq \liminf_{T\to\infty} \mathrm{P}\Big(V_{2,\gamma_2}\left(\frac{T}{4}\right) < \lambda_T,\ V_{1,\gamma_1}\left(\frac{T}{4}\right) > \lambda_T,\ X_{1\frac{T}{4}+1} > \lambda_T + r_T,\ \mathcal{E}(T)\Big) \\
&= \liminf_{T\to\infty} \mathrm{P}\Big(V_{2,\gamma_2}\left(\frac{T}{4}\right) < \lambda_T,\ V_{1,\gamma_1}\left(\frac{T}{4}\right) > \lambda_T,\ X_{\frac{T}{4}+1} > \lambda_T + r_T\Big) \\
&= \liminf_{T\to\infty} \mathrm{P}\Big(V_{2,\gamma_2}\left(\frac{T}{4}\right) < \lambda_T,\ V_{1,\gamma_1}\left(\frac{T}{4}\right) > \lambda_T\Big) \times \liminf_{T\to\infty} \mathrm{P}\Big(X_{\frac{T}{4}+1} > \lambda_T + o(1)\Big).
\end{aligned}$$

The last equality follows because $X_{1.\frac{T}{4}+1}$ is independent [1] of history up to time $\frac{T}{4}$. Let X be a identical copy of $X_{\frac{T}{4}}$. Then, from the last equality we conclude that :

$$\liminf_{T\to\infty} D_{2,T} \geq \Phi\Big(\frac{\gamma_1 - c}{2}\Big)\Phi\Big(\frac{c-\gamma_2}{2}\Big)\mathrm{P}\Big(X > 0\Big) \quad \text{as} \quad T \to \infty \tag{40}$$

The proof of the K armed case follows using an exactly same argument, where we analyze the event

$$\mathrm{P}\Big(V_{a,\gamma_2}\left(\frac{T}{4}\right) < \lambda_T\ \forall\ a \in \{2,\dots K\},\ V_{1,\gamma_1}\left(\frac{T}{4}\right) > \lambda_T,\ V_{1,\gamma_1}\left(\frac{T}{4}+1\right) > \lambda_T\Big)$$

Proof of claim (36b)

The proof follows by noting:

$$\begin{aligned}
&\Big\{V_{a,\gamma_2}\left(\frac{T}{2K}\right) < \lambda_T\ \forall a \in \{2,\dots,K\},\ V_{1,\gamma_1}\left(\frac{T}{2K}\right) > \lambda_T, V_{1,\gamma_1}\left(\frac{T}{2K}+1\right) > \lambda_T\Big\} \\
&\subseteq \Big\{V_{1,\gamma_1}\left(\frac{T}{2K}\right) > \lambda_T\Big\}
\end{aligned}$$

[1] See the argument in the proof of Lemma 5, equation 85

Proof of claim (36c)

We prove the claim for the two-arm bandit scenario. Extension to the K arm case follows via analogous argument. Now, to quantify the limiting behavior of $(D_{3,t})_{t\geq 1}$, consider the following :

$$D_{3,T} = \mathrm{P}\Big(\min_{\frac{T}{2K}+1\leq t\leq T} V_{1,\gamma_1}(t) > \lambda_T \Big)$$

$$= \mathrm{P}\Big(\min_{\frac{T}{2K}+1\leq t\leq T} V_{1,\gamma_1}(t) > \lambda_T,\ n_{1,\frac{T}{4}} \in \Big[\frac{T}{16}, \frac{3T}{16}\Big] \Big) + o(1)$$

In the two armed MAB problem, at step $t+1$ a reward is drawn from one of the arms where the decision to choose the reward distribution is based on the UCB strategy. Consider the reward generating processes for arm a be $(X_{a,t})_{t\geq 1}$. This process is a sequence of i.i.d random variables which are independent of the bandit algorithm. Now for arm a let,

$$\bar{\mu}_{a,t} := \frac{1}{t}\sum_{k=1}^{t} X_{a,k} \tag{41}$$

Before we continue, let us recall $g_\gamma(t) := \gamma/\sqrt{t}$ and define one additional entity,

$$V^*_{a,\gamma}(t) := \bar{\mu}_{a,t} + g_\gamma(t) \tag{42}$$

Therefore, from the definition of $\widehat{\mu}_{a,t}$ and equations (41) it follows that,

$$\widehat{\mu}_{a,t} = \bar{\mu}_{a,n_{a,t}} \tag{43}$$

Consequently if $n_{a,t} \in \Big[\frac{T}{16}, \frac{3T}{16}\Big]$, from equation (43) it follows that,

$$\widehat{\mu}_{a,t} \ \in \ \Big\{ \bar{\mu}_{a,t} \Big| t \in \Big[\frac{T}{16}, \frac{3T}{16}\Big] \Big\} \tag{44}$$

Therefore, from equations (42) and (44) it follows that:

$$\liminf_{T\to\infty} D_{3,T} \overset{(i)}{\geq} \liminf_{T\to\infty} \mathrm{P}\Big(\min_{\frac{T}{4}+1\leq t\leq T} V_{1,\gamma_1}(t) > \lambda_T, n_{1,\frac{T}{4}} \in \Big[\frac{T}{16}, \frac{3T}{16}\Big] \Big)$$

$$\overset{(ii)}{\geq} \liminf_{T\to\infty} \mathrm{P}\Big(\min_{\frac{T}{16}\leq t\leq T} V^\star_{1,\gamma_1}(t) > \lambda_T, n_{1,\frac{T}{4}} \in \Big[\frac{T}{16}, \frac{3T}{16}\Big] \Big)$$

$$\overset{(iii)}{=} \liminf_{T\to\infty} \mathrm{P}\Big(\min_{\frac{T}{16}\leq t\leq T} V^\star_{1,\gamma_1}(t) > \lambda_T \Big)$$

$$= 1 - \limsup_{T\to\infty} \mathrm{P}\Big(\min_{\frac{T}{16}\leq t\leq T} V^\star_{1,\gamma_1}(t) \leq \lambda_T \Big)$$

$$= 1 - \limsup_{T\to\infty} \mathrm{P}\Big(\max_{\frac{T}{16}\leq t\leq T} (-V^\star_{1,\gamma_1}(t)) \geq -\lambda_T \Big)$$

Inequality (i) and equality (iii) follow from stability (21). Inequality (ii) follows from equation (44). Now, the lemma stated below completes the proof.

Lemma 7. *Suppose $(X_{a,t})_{t\geq 1}$ are centered 1-subgaussian random variables and $-V^*_{1,t}(\beta_1)$ is as defined in equation (42). Then there exists an $\alpha \in (0,1)$ and a $c_\alpha > 0$ such that for $\lambda_T = -c_\alpha/\sqrt{T}$ we have :*

$$\limsup_{T\to\infty} \mathrm{P}\Big(\max_{\frac{T}{16}\leq t\leq T} (-V^\star_{1,\gamma_1}(t)) \geq -\lambda_T \Big) \leq \alpha$$

We note that the random process $(V^\star_{1,t}(\beta_1))_{t\geq 1}$ is neither a sub-martingale nor a super-martingale, and as consequence, standard maximal concentration inequalities like Doob's inequality cannot be applied. To prove this lemma we shall apply Theorem 1 and the proof is provided in the next section.

Proof of Lemma 7

Let us recall from equations (41) and (42) that

$$V^*_{a,\gamma}(t) := \bar{\mu}_{a,t} + g_\gamma(t) = \frac{1}{t}\sum_{k=1}^{t} X_{a,t} + g_\gamma(t)$$

where $(X_{a,t})_{t\geq 1}$ are centered 1-subgaussian random variables and $g_\gamma(t) := \gamma/\sqrt{t}$. To prove this lemma we shall use the following maximal inequality for un-centered sampled means which we state below.

Lemma 8. *Suppose $X_1, \ldots, X_n$ are i.i.d. 1-subgaussian random variables with mean 0 and variance σ_X. Let $g(n)$ be any real decreasing function and fix $\lambda > 0$. Then there exists absolute constant $\tilde{C} > 0$ for which we have,*

$$\lambda \mathrm{P}\Big(\max_{\alpha N\leq n\leq \beta N} \Big[\bar{X}_{n+1} + g(n+1)\Big] \geq \lambda \Big) \leq \frac{1}{\sqrt{N}}\left(\tilde{C}\sqrt{\frac{2}{\pi}} e^{-Ng(\beta N)^2/2} + 2\tilde{C} \right) + \Big(g(\alpha N) - g(\beta N) \Big)$$
$$+ g(\beta N)\Big(1 - 2\Phi\Big(-\frac{\sqrt{N} g(\beta N)}{\sigma_X} \Big)\Big) + \sum_{k=\alpha N}^{\beta N - 1} \frac{1}{(k+1)\sqrt{k}} \sqrt{\frac{2}{\pi}} \tag{45}$$

Therefore, for the choice of $\lambda_T = -c/\sqrt{T}$ we obtain the following :

$$-\frac{1}{\lambda_T\sqrt{T}}\left(\tilde{C}\sqrt{\frac{2}{\pi}} e^{-Tg_{\gamma_1}(T)^2/2} + 2\tilde{C} \right) = \frac{1}{c}\Big[\tilde{C}\sqrt{\frac{2}{\pi}} e^{-(\beta_1)^2/2} + 2\tilde{C} + \gamma\Big(1 - 2\Phi\Big(-\beta_1 \Big)\Big)\Big] \tag{46}$$

and,

$$\frac{\sqrt{T}}{c}\sqrt{\frac{2}{\pi}} \sum_{k=T/16}^{T-1} \frac{1}{(k+1)\sqrt{k}} \leq \frac{4}{c}\sqrt{\frac{2}{\pi}} \frac{16T}{T+16} \leq \frac{4}{c}\sqrt{\frac{2}{\pi}} \tag{47}$$

The second last inequality in the above equation holds because $k \geq T/16$. Finally, we have :

$$\frac{\sqrt{T}}{c}\Big(g\left(\frac{T}{16}\right) - g(T)\Big) = \frac{\sqrt{T}}{c}\Big(\frac{16}{\sqrt{T}} - \frac{1}{\sqrt{T-1}} \Big) \leq \frac{16}{c} \tag{48}$$

Therefore,for any choice of $\alpha \in (0,1)$, it follow from equations (46), (47) and (48) it follows that by taking c sufficiently large,

$$\mathrm{P}\Big(\max_{t\in[\frac{T}{16},T]}(-V^*_{a,\gamma_1}(t)) \geq -\lambda_T\Big) < \alpha$$

This completes our proof. Lemma 8 is derived in the next section, which requires application of a generalized Doob's inequality. Before we conclude this section here is an important remark. From equation (11) it follows that for *any* choice of γ in $\mathbb{R}$ it follows that,

$$\sqrt{\frac{2}{\pi}}e^{-\gamma^2/2} + \gamma\Big(1 - 2\Phi(-\gamma)\Big) > 0 \tag{49}$$

This is because equation (49) corresponds to equation (11) when $\mu = \gamma$ and $\sigma = 1$ as $\mathbf{E}[|V|] > 0$. This observation allows us to be flexible in the choice of constants γ_1 and γ_2 in Theorem 1.

B Maximal inequality for sample means

In this section we derive a maximal inequality for centered sample means where the reward distribution is sub Gaussian. As noted in the earlier section, the process $(\bar{X}_n)$ is not a martingale, sub-martingale or super-martingale. Consequently, standard maximal inequality of Doob and Kolmogorov cannot be applied. Here we first derive a maximal inequality for a general class of processes and then apply that result to derive our desired inequality.

B.1 Maximal inequality for a general class of processes

In this section we discuss a result for random processes of a specific kind which generalize the standard Doob's maximal inequality. Consider a stochastic process $(Q_t)_{t\geq 1}$ and let $(\mathcal{F}_t)_{t\geq 1}$ be the natural filtration w.r.t this process. Assume that a decomposition of the following kind holds:

$$\mathbf{E}\Big[Q_{t+1}\Big|\mathcal{F}_t\Big] = Q_t + R_t \tag{50}$$

where $R_t \in \mathcal{F}_t$.

We note that if $R_t \leq 0$ (almost surely), then $(Q_t, \mathcal{F}_t)_{t\geq 1}$ is a sub-maringale (analogously we have super-martingales). However, it may happen that R_t is a random variable which can take both positive and negative values then the process $(Q_t, \mathcal{F}_t)_{t\geq 1}$ is neither a sub-martingale nor a super-martingale. Hence, our setup is more general. For such processes, we have the following result.

Proposition 1. Fix $\lambda > 0$ and suppose that $\alpha < \beta$ are distinct, arbitrary constants in $(0, 1]$, and Q_k, R_k are as defined in Lemma 50 for each k. Then, we have the following maximal inequality :

$$\mathrm{P}\Big(\max_{\alpha T\leq t\leq \beta T} Q_t \geq \lambda\Big) \leq \frac{1}{\lambda}\Big(\mathbf{E}\Big[|Q_{\beta T}|\Big] + \sum_{k=\alpha T}^{\beta T-1}\mathbf{E}\Big[|R_k|\Big]\Big) \tag{51a}$$

where

We note that when $R_t = 0$ almost surely then this result reduces to Doob's maximal inequality. The proof of this proposition follows the standard approach. Let us begin by defining a stopping time τ, defined as follows :

$$\tau := \min\Big(\inf\Big\{k \geq \alpha T \;\Big|\; \max_{\alpha T\leq t\leq k} Q_t \geq \lambda\Big\}, \beta T\Big) \tag{52}$$

Now, consider the following lemma :

Lemma 9. *Consider the stopping time τ as defined in equation (52) and let $\mathcal{F}_\tau$ be the stopped σ-field. Then we have the following two results:*

a $\{\max_{\alpha T\leq t\leq \beta T} Q_t \geq \lambda\}$ *belongs in the stopped σ field $\mathcal{F}_\tau$.*

b $\{\max_{\alpha T\leq t\leq \beta T} Q_t \geq \lambda\} = \{Q_\tau \geq \lambda\}$

By applying Lemma 9, we obtain the following :

$$\lambda\mathrm{P}\Big(\max_{\alpha T\leq t\leq \beta T} Q_t \geq \lambda\Big) = \mathbf{E}\Big[\lambda\mathbb{I}_{\{\max_{\alpha T\leq t\leq \beta T} Q_t\geq\lambda\}}\Big] = \mathbf{E}\Big[\lambda\mathbb{I}_{\{Q_\tau\geq\lambda\}}\Big] \leq \mathbf{E}\Big[Q_\tau\mathbb{I}_{\{Q_\tau\geq\lambda\}}\Big] \tag{53}$$

Lemma 10. *Suppose τ_1, τ_2 are two stopping times such that $\mathrm{P}(\tau_1 \leq \tau_2 \leq n) = 1$ (where n is fixed). Then, for any event $A \in \mathcal{F}_{\tau_1}$, we have :*

$$\mathbf{E}\Big[Q_{\tau_1}\mathbb{I}_{\{A\}}\Big] = \mathbf{E}\Big[Q_{\tau_2}\mathbb{I}_{\{A\}} - \Big(\sum_{\tau_1}^{\tau_2-1} R_k\Big)\mathbb{I}_{\{A\}}\mathbb{I}_{\{\tau_2>\tau_1\}}\Big] \tag{54}$$

where $\mathcal{F}_{\tau_1}$ is the stopped σ-field (w.r.t τ_1) defined as $\{A : A \cap \{\tau_1 = k\} \in \mathcal{F}_k, \text{for all} \quad k \in \mathcal{N}\}$.

This lemma is a direct generalization of the standard OST result. Now, suppose this lemma holds. Then by replacing $\tau_1 = \tau$ and $\tau_2 = \beta T$ in Lemma 10 we have :

$$\mathbf{E}\Big[Q_{\tau}\mathbb{I}_{\{Q_\tau \geq \lambda\}}\Big] = \mathbf{E}\Big[Q_{\beta T}\mathbb{I}_{\{Q_\tau \geq \lambda\}} - \Big(\sum_{k=\tau}^{\beta T-1} R_{k-1}\Big)\mathbb{I}_{\{Q_\tau \geq \lambda\}}\mathbb{I}_{\{\beta T>\tau\}}\Big] \tag{55}$$

Therefore, from equations (53) and (55) we have the following chain of inequalities :

$$\begin{aligned}
\lambda \mathrm{P}\Big(\max_{\alpha T \leq t \leq \beta T} Q_t \geq \lambda\Big) &\leq \mathbf{E}\Big[Q_{\beta T}\mathbb{I}_{\{Q_\tau \geq \lambda\}} - \Big(\sum_{k=\tau}^{\beta T-1} R_{1,k}\Big)\mathbb{I}_{\{Q_\tau \geq \lambda\}}\mathbb{I}_{\{\beta T>\tau\}}\Big] \\
&\leq \mathbf{E}\Big[\Big|Q_{\beta T}\mathbb{I}_{\{Q_\tau \geq \lambda\}} - \Big(\sum_{k=\tau}^{\beta T-1} R_k\Big)\mathbb{I}_{\{Q_\tau \geq \lambda\}}\mathbb{I}_{\{\beta T>\tau\}}\Big] \\
&\leq \mathbf{E}\Big[|Q_{\beta T}|\mathbb{I}_{\{Q_\tau \geq \lambda\}} + \Big(\sum_{k=\tau}^{\beta T-1} |R_k|\Big)\mathbb{I}_{\{Q_\tau \geq \lambda\}}\mathbb{I}_{\{\beta T>\tau\}}\Big] \\
&\leq \mathbf{E}\Big[|Q_{\beta T}|\mathbb{I}_{\{Q_\tau \geq \lambda\}}\Big] + \mathbf{E}\Big[\Big(\sum_{k=\tau}^{\beta T-1} |R_k|\Big)\mathbb{I}_{\{Q_\tau \geq \lambda\}}\mathbb{I}_{\{\beta T>\tau\}}\Big] \\
&\leq \mathbf{E}\Big[|Q_{\beta T}|\Big] + \mathbf{E}\Big[\sum_{k=\tau}^{\beta T-1} |R_k|\Big]
\end{aligned}$$

This completes the proof of Lemma 1, once we prove Lemma 10. We note that as $\tau_1 \leq \tau_2 \leq n$, we have

$$\mathbf{E}\Big[\Big(Q_{\tau_2} - Q_{\tau_1}\Big)\mathbb{I}_{\{A\}}\Big] = \mathbf{E}\Big[\Big(Q_{\tau_2} - Q_{\tau_1}\Big)\mathbb{I}_{\{A\}}\mathbb{I}_{\{\tau_2>\tau_1\}}\Big]$$

This is true because on the event $\{\tau_1 = \tau_2\}$, $Q_{\tau_2} - Q_{\tau_1} = 0$. This observation leads us to the

following set of equalities.

$$\begin{aligned}
\mathbf{E}\Big[\Big(Q_{\tau_2} - Q_{\tau_1}\Big)\mathbb{I}_{\{A\}}\Big] &= \mathbf{E}\Big[\Big(Q_{\tau_2} - Q_{\tau_1}\Big)\mathbb{I}_{\{A\}}\mathbb{I}_{\{\tau_2 > \tau_1\}}\Big] \\
&= \mathbf{E}\Big[\sum_{\tau_1+1}^{\tau_2}\Big(Q_k - Q_{k-1}\Big)\mathbb{I}_{\{\tau_2>\tau_1, A\}}\Big] \\
&= \mathbf{E}\Big[\sum_{2}^{n}\Big(Q_k - Q_{k-1}\Big)\mathbb{I}_{\{\tau_1<k\leq\tau_2, A\}}\Big] \\
&= \sum_{2}^{n}\Big(\mathbf{E}\Big[Q_k\mathbb{I}_{\{\tau_1<k\leq\tau_2, A\}}\Big] - \mathbf{E}\Big[Q_{k-1}\mathbb{I}_{\{\tau_1<k\leq\tau_2, A\}}\Big]\Big)
\end{aligned}$$

Now, as τ_1 and τ_2 are both stopping times, $\{\tau_1 \leq k-1\}$ and $\{\tau_2 \leq k-1\}^c$ lie in the σ-field $\mathcal{F}_{k-1}$. Furthermore, as $A \in \mathcal{F}_{\tau_1}$, it implies that $\{\tau_1 \leq k-1\} \cap A$ measurable in $\mathcal{F}_{k-1}$. Hence, we conclude that $\{\tau_1 \leq k-1\} \cap A \cap \{\tau_2 \leq k-1\}^c$ belongs to $\mathcal{F}_{k-1}$. This is equivalent to the statement $A \cap \{\tau_1 < k \leq \tau_2\} \in \mathcal{F}_{k-1}$, for any $A \in \mathcal{F}_{\tau_1}$. Therefore, from equation (50) we have:

$$\mathbf{E}\Big[Q_k 1\left\{\tau_1 < k \leq \tau_2, A\right\}\Big] = \mathbf{E}\Big[Q_{k-1} 1\left\{\tau_1 < k \leq \tau_2, A\right\}\Big] + \mathbf{E}\Big[\Big(R_{k-1}\Big)1\{\tau_1 < k \leq \tau_2, A\}\Big] \tag{56}$$

Therefore, by applying equation (56) in the previous chain of inequalities, we obtain :

$$\begin{aligned}
\mathbf{E}\Big[\Big(Q_{\tau_2} - Q_{\tau_1}\Big)\mathbb{I}_{\{A\}}\Big] &= \sum_{2}^{n}\mathbf{E}\Big[\Big(R_{k-1}\Big)\mathbb{I}_{\{\tau_1<k\leq\tau_2, A\}}\Big] \\
&= \mathbf{E}\Big[\sum_{2}^{n}\Big(R_{k-1}\Big)\mathbb{I}_{\{\tau_1<k\leq\tau_2, A\}}\Big] \\
&= \mathbf{E}\Big[\sum_{\tau_1+1}^{\tau_2}\Big(R_{k-1}\Big)\mathbb{I}_{\{\tau_2>\tau_1\}}\mathbb{I}_{\{A\}}\Big]
\end{aligned}$$

Therefore, by rearranging the terms, we have our result.

$$\mathbf{E}\Big[Q_{\tau_1}\mathbb{I}_{\{A\}}\Big] = \mathbf{E}\Big[Q_{\tau_2}\mathbb{I}_{\{A\}} - \Big(\sum_{\tau_1}^{\tau_2-1} R_k\Big)\mathbb{I}_{\{A\}}\mathbb{I}_{\{\tau_2>\tau_1\}}\Big]$$

B.2 Application to sample means

Let $g(n)$ be any real valued function and suppose that $X_1, \ldots, X_n$ are iid standard gaussian random variables. We wish to derive a maximal inequality for the process $(\bar{X}_n + g(n))_{n\geq 1}$. Then we have Lemma 8 stated in the earlier section.

Lemma. Suppose $X_1, \ldots, X_n$ are i.i.d. 1 sub-Gaussian random variables with mean 0 and variance σ_X. Let $g(n)$ be any real decreasing function and fix $\lambda > 0$. Then there exists absolute constant

$\tilde{C} > 0$ for which we have,

$$\lambda \mathrm{P}\Big(\max_{\alpha N \le n \le \beta N} \Big[\bar{X}_{n+1} + g(n+1) \Big] \ge \lambda \Big) \le \frac{1}{\sqrt{N}} \left(\tilde{C} \sqrt{\frac{2}{\pi}} e^{-N g(\beta N)^2/2} + 2\tilde{C} \right) + \Big(g(\alpha N) - g(\beta N) \Big) + g(\beta N) \Big(1 - 2\Phi\Big(- \frac{\sqrt{N} g(\beta N)}{\sigma_X} \Big) + \sum_{k=\alpha N}^{\beta N - 1} \frac{1}{(k+1)\sqrt{k}} \sqrt{\frac{2}{\pi}} \tag{57}$$

The proof is a direct application of Proposition (1). We can apply this lemma due to the following observation.

$$\begin{aligned} \bar{X}_{n+1} + g(n+1) &= \frac{n}{n+1} \bar{X}_n + \frac{X_{n+1}}{n+1} + g(n+1) \\ &= \bar{X}_n + \frac{X_{n+1} - \bar{X}_n}{n+1} + g(n+1) \\ &= \Big[\bar{X}_n + g(n) \Big] + \frac{X_{n+1} - \bar{X}_n}{n+1} + [g(n+1) - g(n)] \end{aligned} \tag{58}$$

Let $\mathcal{F}_n$ be the σ-field generated by $X_1, \ldots, X_n$]]. Then it follows that $\bar{X}_n$ and $g(n)$ is measurable with respect to $\mathcal{F}_n$ and X_{n+1} is independent of $\mathcal{F}_n$. Therefore we have,

$$\mathbf{E}\Big[\bar{X}_{n+1} + g(n+1) \Big| \mathcal{F}_t \Big] = \Big[\bar{X}_n + g(n) \Big] + \left[- \frac{\bar{X}_n}{n+1} + (g(n+1) - g(n)) \right] \tag{59}$$

By comparing equations (59) and (50) and applying Proposition 1 we obtain the following maximal inequality:

$$\begin{aligned} &\mathrm{P}\Big(\max_{\alpha N \le n \le \beta N} \Big[\bar{X}_{n+1} + g(n+1) \Big] \ge \lambda \Big) \\ &\le \frac{1}{\lambda} \left(\mathbf{E}\Big[|\bar{X}_{\beta N} + g(\beta N)| \Big] + \sum_{k=\alpha N}^{\beta N - 1} \mathbf{E}\Big[\Big| - \frac{\bar{X}_k}{k+1} + (g(k+1) - g(k)) \Big| \Big] \right) \\ &\le \frac{1}{\lambda} \left(\mathbf{E}\Big[|\bar{X}_{\beta N} + g(\beta N)| \Big] + \sum_{k=\alpha N}^{\beta N - 1} \mathbf{E}\Big[\Big| \frac{\bar{X}_k}{k+1} \Big| \Big] + \sum_{k=\alpha N}^{\beta N - 1} \mathbf{E}\Big[\Big| (g(k+1) - g(k)) \Big| \Big] \right) \end{aligned} \tag{60}$$

If the random variables $X_1, \ldots, X_n$ are Gaussian random variables, we can write the first moment of the absolute values more explicitly. To do so we shall apply the following well known result (Leone et al. (1961)).

Lemma 11. *If $V \sim N(\mu, \sigma^2)$, then :*

$$\mathbf{E}[|V|] = \sigma \sqrt{\frac{2}{\pi}} e^{-\mu^2/2\sigma^2} + \mu \Big(1 - 2\Phi\Big(- \frac{\mu}{\sigma} \Big) \Big)$$

By using this lemma, we have:

$$\mathbf{E}\Big[|\bar{X}_{\beta N} + g(\beta N)|\Big] = \frac{1}{\sqrt{N}}\sqrt{\frac{2}{\pi}}e^{-Ng(\beta N)^2/2} + g(\beta N)\Big(1 - 2\Phi\Big(-\sqrt{T}g(\beta N)\Big)\Big)$$

and,

$$\mathbf{E}\Big[\Big|\frac{\bar{X}_k}{k+1}\Big|\Big] = \frac{1}{(k+1)\sqrt{k}}\sqrt{\frac{2}{\pi}}$$

Let us make a simplifying assumption that g is a decreasing function. Then the above calculations yield the explicit upper bound stated below.

$$\lambda \mathrm{P}\Big(\max_{\alpha N \le n \le \beta N}\Big[\bar{X}_{n+1} + g(n+1)\Big] \ge \lambda\Big)$$

$$\le \frac{1}{\sqrt{N}}\sqrt{\frac{2}{\pi}}e^{-Ng(\beta N)^2/2} + g(\beta N)\Big(1 - 2\Phi\Big(-\sqrt{T}g(\beta N)\Big) + \sum_{k=\alpha N}^{\beta N - 1}\frac{1}{(k+1)\sqrt{k}}\sqrt{\frac{2}{\pi}} + \Big(g(\alpha N) - g(\beta N)\Big)$$

This concludes the proof for the case when the reward distribution is Gaussian. The extension to the case of sub-gaussian rewards follows directly from the following lemma, which is proved in the next section.

Lemma 12. *Suppose $X_1, \ldots, X_n$ are i.i.d. 1-sub-Gaussian distribution with mean μ. Then we have :*

$$\mathbf{E}[|\bar{X}_n|] \le \frac{1}{\sqrt{n}}\Big(\tilde{C}\sqrt{\frac{2}{\pi}}e^{-\mu^2 n/2} + 2\tilde{C}\Big) + \mu\Big(1 - 2\Phi\Big(-\frac{\mu\sqrt{n}}{\sigma_X}\Big)\Big)$$

where σ_X is the variance of X_1 and $\tilde{C} > 0$ is a universal constant.

Proof of Lemma 12

Suppose $X_1, \ldots, X_n$ are sub-Gaussian random variables. Recall that a random variable X is called σ-sub-Gaussian, if the following inequality holds $\forall\ \lambda > 0$:

$$\mathbf{E}[e^{\lambda(X - \mathbf{E}[X])}] \le e^{\sigma^2\lambda^2/2}$$

Hence if X is a σ-sub-Gaussian random variable, then it is also $\tilde{\sigma}$-sub-Gaussian for any $\tilde{\sigma} > \sigma$. One can define a unique variance proxy for sub-gaussian random variables. This is called *optimal variance proxy* (Arbel et al. (2020); Vershynin (2018)) which we define below:

$$\mathrm{Var}_G(X) := \inf_{\sigma > 0}\Big\{\mathbf{E}[e^{\lambda(X - \mathbf{E}[X])}] \le e^{\sigma^2\lambda^2/2} \quad \forall \quad \lambda > 0\Big\} \tag{61}$$

It has several well documented properties (Vershynin (2018)), which we state below in form of a lemma.

Lemma 13. *Suppose $X_1, \ldots, X_n$ are centered sub-Gaussian random with variance $\sigma^2_{X_i}$. Then the optimal variance proxy satisfies the following properties:*

1. $\mathrm{Var}_G(cX_i) = c^2\,\mathrm{Var}_G(X_i)$.

2. $\sigma^2_{X_i} \leq \mathrm{Var}_G(X_i)$ *for all* $i \in [n]$.

3. $\mathrm{Var}_G(\sum_{i=1}^n X_i) \leq \sum_{i=1}^n \mathrm{Var}_G(X_i)$.

4. *There exists absolute constants* $c_1, c_2 > 0$ *such that* $c_1||X_i||_{\psi_2} \leq \sqrt{\mathrm{Var}_G(X_i)} \leq c_2||X_i||_{\psi_2}$.

In property 4 of $\mathrm{Var}_G(X_i)$, $||X_i||_{\psi_2}$ is the *Orlicz* norm of X with respect to the function $\psi_2(x) := e^{x^2} - 1$. Let us begin our proof. It follows that if we have i.i.d. sub- Gaussian random variables $X_1, \ldots, X_n$ it follows from Lemma 13 that

$$\mathrm{Var}_G(\bar{X}_n) \leq \frac{1}{n}\,\mathrm{Var}_G(X_1)$$

Furthermore, an alternate characterization for sub gaussian random variable states that if X is any sub-Gaussian random variable then we have

$$\mathbf{E}[|X|] \leq C||X||_{\psi_2} \tag{62}$$

Therefore, if X is a sub-Gaussian random variable then the following inequality follows from Equation (62) and Lemma 13 that there exists absolute constant $\tilde{C}$ such that

$$\mathbf{E}[|X|] \leq \tilde{C}\,\sqrt{\mathrm{Var}_G(X)} \tag{63}$$

Now we shall use this result to obtain an upper bound on $\mathbf{E}[|\bar{X}_n + \mu|]$. Let Z be a normal random variable with mean 0 and variance $\mathrm{Var}_G(\bar{X}_n)$. Then we have

$$\begin{aligned}\mathbf{E}[|\bar{X}_n + \mu|] &\leq \mathbf{E}[|Z + \mu|] + \mathbf{E}[|\bar{X}_n - Z|] \\ &= \sqrt{\mathrm{Var}_G(\bar{X}_n)}\sqrt{\frac{2}{\pi}}e^{-\mu^2/2\,\mathrm{Var}_G(\bar{X}_n))^2} + \mu\Big(1 - 2\Phi\Big(-\frac{\mu}{\sqrt{\mathrm{Var}_G(\bar{X}_n)}}\Big)\Big) + \mathbf{E}[|\bar{X}_n - Z|]\end{aligned}$$

We control the second term by applying Equation (62) and Lemma 13 as follows

$$\begin{aligned}\mathbf{E}[|\bar{X}_n - Z|] &\leq C||\bar{X}_n - Z||_{\psi_2} \\ &\leq \tilde{C}\sqrt{\mathrm{Var}_G(\bar{X}_n - Z)} \\ &\leq \tilde{C}\sqrt{\mathrm{Var}_G(\bar{X}_n) + \mathrm{Var}_G(Z)} \\ &\leq \tilde{C}\left[\sqrt{\mathrm{Var}_G(\bar{X}_n)} + \sqrt{\mathrm{Var}_G(Z)}\right] \\ &\leq 2\tilde{C}\frac{\sqrt{\mathrm{Var}_G(X_1)}}{\sqrt{n}}\end{aligned}$$

Hence we have,

$$\begin{aligned}\mathbf{E}[|\bar{X}_n - Z|] \leq \tilde{C}\frac{\sqrt{\mathrm{Var}_G(X_1)}}{\sqrt{n}}\sqrt{\frac{2}{\pi}}e^{-\mu^2/2\,\mathrm{Var}_G(\bar{X}_n))^2} + \mu\Big(1 - 2\Phi\Big(-\frac{\mu}{\sqrt{\mathrm{Var}_G(\bar{X}_n)}}\Big)\Big) \\ + \tilde{C}\frac{\sqrt{\mathrm{Var}_G(X_1)}}{\sqrt{n}}\end{aligned}$$

Now since $X_1, \ldots, X_n$ are 1-sub-Gaussian random variables, it follows from the definition of optimal variance proxy that $\mathrm{Var}_G(X_1) \leq 1$. Furthermore, as $\sigma_X^2/n = \mathrm{Var}(\bar{X}_n) \leq \mathrm{Var}_G(\bar{X}_n)$. Due to this we note that

$$\begin{aligned}
\mathbf{E}[|\bar{X}_n - Z|] &\leq \tilde{C}\frac{1}{\sqrt{n}}\sqrt{\frac{2}{\pi}}e^{-\mu^2 n/2} + \mu\Big(1 - 2\Phi\Big(-\frac{\mu\sqrt{n}}{\sigma_X}\Big)\Big) + 2\tilde{C}\frac{1}{\sqrt{n}} \\
&\leq \frac{1}{\sqrt{n}}\left[\tilde{C}\sqrt{\frac{2}{\pi}}e^{-\mu^2 n/2} + 2\tilde{C}\right] + \mu\Big(1 - 2\Phi\Big(-\frac{\mu\sqrt{n}}{\sigma_X}\Big)\Big)
\end{aligned}$$

C Instability and Non-normality of Batched Bandits

In this section we shall discuss three batched bandit strategies discussed in Zhang et al. (2020). These algorithms where shown to have non-normal limits as proved in Zhang et al. (2020). Here we prove that they are also not stable. Hence, these algorithms show that if the stability condition is not satisfied by a bandit algorithm, then it may not satisfy a central limit theorem.

Let us first discuss the setup. Here we consider a two-armed, batched bandits with two epochs. The first epoch consists of $T/2$ runs followed by the second epoch. Consider arm 1, and let us define the first epoch by a sequence of binary actions $\{\mathcal{A}_{1,i}, \mathcal{A}_{2,i}\}_{i=1}^{T/2}$. Here for $k \in \{1,2\}$, $\mathcal{A}_{k,i} = 1$ indicates that arm 1 is pulled at step i during the k^{th} epoch. Consequently, we note that $\mathcal{A}_{1,i} = 0$ indicates that arm 2 is pulled. Conditionally on the action, the observed reward is

$$X_i = \begin{cases} X_{1,i}^{(1)}, & \text{if } \mathcal{A}_{1,i} = 1, \\ X_{1,i}^{(2)}, & \text{if } \mathcal{A}_{1,i} = 0. \end{cases} \tag{64}$$

We assume that the actions are independent Bernoulli random variables:

$$\mathcal{A}_{1,i} \sim \text{Ber}(p_1), \quad i = 1, \ldots, T/2. \tag{65}$$

Let $\mathcal{F}_i := \{\mathcal{A}_{1,1}, X_1, \ldots, \mathcal{A}_{1,i}, X_i : i = 1, \ldots, T\}$ denote the history of the first epoch. In the second epoch, we select actions

$$\{\mathcal{A}_{2,i}\}_{i=1}^{T/2} \sim \text{Ber}(p_2(T)), \tag{66}$$

where the success probability $p_2(T)$ is adapted to the history of the first epoch. Hence, by changing the choice of p_1 and $p_2(T)$ we obtain different batched bandit strategies. Now as mentioned earlier we shall present three such strategies which were discussed in Zhang et al. (2020).

ϵ-Greedy ETC

Here we describe an ϵ greedy Explore then Commit (ETC) strategy for some fixed $\epsilon > 0$. Let $\alpha = T/2$ and define $\widehat{\mu}_{1,\alpha T}$ to be the mean rewards in the first epoch for arm a. Now, let $p_1 = 1/2$, and $p_2(T)$ is defined as follows:

$$p_2(T) = \begin{cases} 1 - \dfrac{\epsilon}{2}, & \text{if } \widehat{\mu}_{1,\frac{T}{4}} > \widehat{\mu}_{2,\frac{T}{4}}, \\ \dfrac{\epsilon}{2}, & \text{otherwise .} \end{cases} \tag{67}$$

Therefore, for the first $T/2$ rounds either the first or the second arm is pulled with equal probability $1/2$. Then the arm which had the higher mean reward in the first epoch is drawn (independently) with higher probability in the second epoch.

Thompson Sampling in batched setting

Suppose $p_1 = 1/2$. For each arm a set $n_{a,t}$ as the number of arm pulls at the end of first epoch. Similarly, let $\widehat{\mu}_{a,\frac{T}{2}}$ be the mean reward of arm a at the end of first epoch. Assume we have β_1, β_2

as i.i.d priors $N(0,1)$. Furthermore, let $\tilde{\beta}_1,\ \tilde{\beta}_2$ be the sampled posterior means of arms 1 and 2 respectively. Finally, fix $\pi_{\min}$ and $\pi_{\max}$ such that $0 < \pi_{\min} \leq \pi_{\max} < 1$. Now we define $p_2(T)$ as follows:

$$p_2(T) = \begin{cases} \pi_{\max}, & \text{if } \mathrm{P}(\tilde{\beta}_1 > \tilde{\beta}_2 | \mathcal{H}_{T/2}) \geq \pi_{\max}, \\ \mathrm{P}(\tilde{\beta}_1 > \tilde{\beta}_2 | \mathcal{H}_{T/2}), & \text{if } \pi_{\min} < \mathrm{P}(\tilde{\beta}_1 > \tilde{\beta}_2 | \mathcal{H}_{T/2}) \leq \pi_{\max}, \\ \pi_{\min}, & \text{otherwise .} \end{cases} \tag{68}$$

Therefore, for the first $T/2$ rounds either the first or the second arm is pulled with equal probability $1/2$. Then the arm which has higher probability of posterior mean in the first epoch is drawn (independently) with higher probability in the second epoch.

UCB in batched setting

Here we describe a batch Upper Confidence Bound (UCB) strategy for some fixed $\pi_{\max} > 1/2$. Let $\alpha = T/2$ and define the upper confidence bound for arm a as follows:

$$U_a\left(\frac{T}{4}\right) := \widehat{\mu}_a(t) + \sqrt{\frac{\log \frac{T}{4}}{n_{a,\frac{T}{4}}}} \tag{69}$$

Suppose $p_1 = 1/2$. Now we define $p_2(T)$ as follows:

$$p_2(T) = \begin{cases} \pi_{\max}, & \text{if } U_1\left(\frac{T}{4}\right) > U_2\left(\frac{T}{4}\right), \\ 1 - \pi_{\max}, & \text{otherwise .} \end{cases} \tag{70}$$

Therefore, for the first $T/2$ rounds each arm is pulled with equal probability $1/2$. Then the arm which had the highest UCB index in the first epoch is drawn (independently) with higher probability in the second epoch. In Appendix C of Zhang et al. (2020), the authors show that for the aforementioned algorithms, $\sqrt{n_{a_1,T}(\mathcal{A})}(\bar{\mu}_{a_1,T} - \mu_{a_1})$ converges to a non-normal limit. We state it in the form of a lemma below.

Lemma 14. *Suppose we run either the ϵ-greedy ETC algorithm or the UBC algorithm (in a batched setting). Then, we have:*

$$\sqrt{n_{a_1,T}(\mathcal{A})}(\hat{\mu}_{a_1,T} - \mu_{a_1}) \xrightarrow{d} Y \tag{71}$$

where

$$Y = \sqrt{\frac{1}{3-\epsilon}}\left(Z_1 - \sqrt{2-\epsilon}Z_3\right)\mathbb{I}_{Z_1 > Z_2} + \sqrt{\frac{1}{1+\epsilon}}\left(Z_1 - \sqrt{\epsilon}Z_3\right)\mathbb{I}_{Z_1 < Z_2}, \qquad \textit{for ETC}$$

$$Y = \left(\sqrt{\frac{\frac{1}{2}}{\frac{1}{2} + \pi_{max}}} Z_1 + \sqrt{\frac{\pi_{max}}{\frac{1}{2} + \pi_{max}}} Z_3 \right) \mathbb{I}_{(Z_1 > Z_2)} + \left(\sqrt{\frac{\frac{1}{2}}{\frac{3}{2} - \pi_{max}}} Z_1 + \sqrt{\frac{1 - \pi_{max}}{\frac{3}{2} - \pi_{max}}} Z_3 \right) \mathbb{I}_{(Z_1 < Z_2)}$$

for UCB, and

$$Y = \sqrt{\frac{1/2 + 1 - \pi_*}{1 + 2\pi_*}} \left(\sqrt{1/2}\ Z_1 + \sqrt{\pi_*} Z_3 \right) - \sqrt{\frac{1/2 + \pi_*}{3 - 2\pi_*}} \left(\sqrt{1/2}\ Z_2 + \sqrt{1 - \pi_*} Z_4 \right), \quad \textit{for Thompson}$$

for Z_1, Z_2, Z_3 are i.i.d $N(0,1)$ random variables and $\pi_* \in (0,1)$.

We note that even though stability implies asymptotic normality, it has not been proved yet whether stability is a necessary condition. Consequently, non-normal limiting distribution does not imply that the algorithm is necessarily unstable. However, in the next lemma we show that these algorithms are not stable.

Lemma 15. *The batched ϵ-greedy ETC algorithm, batched UCB algorithm and batched Thompson algorithm are unstable.*

Lemma 15 demonstrates that instability in batched bandit algorithms gives rise to non-normal limits. Since lemma 1 already establishes that stability is a sufficient condition for asymptotic normality, these examples highlight the central role of stability: its violation can naturally lead to non-normal asymptotic behavior.

Proof of Lemma 15

Let us first prove it for the ϵ-greedy ETC bandit strategy. We know that for the first $\lfloor T/2 \rfloor$ times, both the arms are pulled with equal probability. For arm 1 we have:

$$n_{1,T} = \sum_{i=1}^{T/2} \mathcal{A}_{1,i} + \sum_{j=1}^{T/2} \mathcal{A}_{2,j} \tag{72}$$

where $\mathcal{A}_{1,i}$ are i.i.d $\text{Ber}(p_1)$ and $\mathcal{A}_{2,i}$ are i.i.d $\text{Ber}(p_2(T))$. We recall that,

$$p_2(T) = \begin{cases} 1 - \dfrac{\epsilon}{2}, & \text{if } \widehat{\mu}_{1,\frac{T}{2}} > \widehat{\mu}_{1,\frac{T}{2}}, \\ \dfrac{\epsilon}{2}, & \text{otherwise .} \end{cases}$$

Now, we define two independent triangular sequences of i.i.d random variables $(\mathcal{B}_{T,i})_{1 \le i \le T/2}$ and $(\mathcal{C}_{T,i})_{1 \le i \le T/2}$ such that $\mathcal{B}_{T,i} \sim \text{Ber}\left(1 - \dfrac{\epsilon}{2}\right)$ and $\mathcal{C}_{T,i} \sim \text{Ber}\left(\dfrac{\epsilon}{2}\right)$. These random variables are *independent* of the bandit strategy and act as potential outcomes for the random variables $\mathcal{A}_{2,i}$. This means the following

$$\mathcal{A}_{2,i} = \begin{cases} \mathcal{B}_{T,i}, & \text{if } \widehat{\mu}_{1,\frac{T}{2}} > \widehat{\mu}_{1,\frac{T}{2}}, \\ \mathcal{C}_{T,i}, & \text{otherwise .} \end{cases} \tag{73}$$

From equation (72) it follows that,

$$\frac{n_{1,T}}{T} = \frac{1}{2}\left(\frac{2}{T}\sum_{i=1}^{T/2}\mathcal{A}_{1,i}\right) + \frac{1}{2}\left(\frac{2}{T}\sum_{j=1}^{T/2}\mathcal{A}_{2,j}\right)$$

Fix some arbitrary $\delta > 0$. We claim the following:

$$\lim_{T\to\infty} \mathrm{P}\left(\left|\frac{n_{1,T}}{T} - \left(\frac{1}{4} + \frac{\epsilon}{2}\right)\right| > \delta\right) = \frac{1}{2}$$

Therefore, by applying Lemma 2 the above result implies instability of the algorithm. Now, by applying the WLLN,

$$\frac{2}{T}\sum_{i=1}^{T/2}\mathcal{A}_{1,i} \xrightarrow{\mathrm{P}} \frac{1}{2}$$

Hence, it follows that it is enough to prove

$$\lim_{T\to\infty} \mathrm{P}\left(\left|\frac{1}{2}\left(\frac{2}{T}\sum_{j=1}^{T/2}\mathcal{A}_{2,j}\right) - \frac{\epsilon}{2}\right| > \frac{\delta}{2}\right) = \frac{1}{2}$$

Equation (73) shows that as $T \to \infty$ the above holds if and only if $\mathcal{A}_{2,i} = \mathcal{C}_{T,i}$. We prove this rigorously below.

$$\begin{aligned}
&\mathrm{P}\left(\left|\frac{1}{2}\left(\frac{2}{T}\sum_{j=1}^{T/2}\mathcal{A}_{2,j}\right) - \frac{\epsilon}{2}\right| > \frac{\delta}{2}\right) \\
&= \mathrm{P}\left(\widehat{\mu}_{2,\frac{T}{2}} > \widehat{\mu}_{1,\frac{T}{2}},\ \left|\frac{1}{2}\left(\frac{2}{T}\sum_{j=1}^{T/2}\mathcal{C}_{T,i}\right) - \frac{\epsilon}{2}\right| > \frac{\delta}{2}\right) + o(1) \\
&= \mathrm{P}\left(\widehat{\mu}_{2,\frac{T}{2}} > \widehat{\mu}_{1,\frac{T}{2}}\right) \times \mathrm{P}\left(\left|\frac{1}{2}\left(\frac{2}{T}\sum_{j=1}^{T/2}\mathcal{C}_{T,i}\right) - \frac{\epsilon}{2}\right| > \frac{\delta}{2}\right) + o(1) \\
&= \frac{1}{2} \times \mathrm{P}\left(\left|\frac{1}{2}\left(\frac{2}{T}\sum_{j=1}^{T/2}\mathcal{C}_{T,i}\right) - \frac{\epsilon}{2}\right| > \frac{\delta}{2}\right) + o(1)
\end{aligned}$$

The last equality holds true because $\mathcal{C}_{T,i}$ are independent of the bandit strategy and the reward distributions for both the arms are identical (which are chosen with equal probability). This proves that the ϵ-greedy algorithm is unstable.

Now, the proof of instability of the batched UCB is analogous to that of the ETC. By replacing $1 - \pi_{\max}$ in place of $\epsilon/2$ we observe that we only need to prove,

$$\lim_{T\to\infty} \mathrm{P}\left(\left|\frac{n_{1,T}}{T} - \left(\frac{1}{4} + 1 - \pi_{\max}\right)\right| > \delta\right) > 0$$

A careful look at the proof of the ETC indicates that our objective is to prove,

$$\lim_{T\to\infty} \mathrm{P}\left(U_2\left(\frac{T}{2}\right) > U_1\left(\frac{T}{2}\right)\right) > 0$$

If possible, suppose stability holds. Then for any $\epsilon > 0$, by applying lemma 2 we know that $\mathrm{P}(n_{1,T/2}/n_{2,T/2} \in (1-\epsilon, 1+\epsilon))$ tends to 1 as $T \to \infty$. By using this result we have the following chain of equalities:

$$\begin{aligned}
&\lim_{T\to\infty} \mathrm{P}\left(U_2\left(\frac{T}{2}\right) > U_1\left(\frac{T}{2}\right)\right) \\
&= \lim_{T\to\infty} \mathrm{P}\left(\widehat{\mu}_{2,\frac{T}{2}} + \sqrt{\frac{2\log 1/\gamma}{n_{2,T}}} > \widehat{\mu}_{1,\frac{T}{2}} + \sqrt{\frac{2\log 1/\gamma}{n_{1,T}}}\right) \\
&= \lim_{T\to\infty} \mathrm{P}\left(\sqrt{n_{2,T}}\,\widehat{\mu}_{2,\frac{T}{2}} + \sqrt{2\log 1/\gamma} > \frac{\sqrt{n_{2,T}}}{\sqrt{n_{1,T}}}\,\sqrt{n_{2,T}}\,\widehat{\mu}_{1,\frac{T}{2}} + \frac{\sqrt{n_{2,T}}}{\sqrt{n_{1,T}}}\,\sqrt{2\log 1/\gamma}\right) \\
&= \lim_{T\to\infty} \mathrm{P}\left(\sqrt{n_{2,T}}\,\widehat{\mu}_{2,\frac{T}{2}} - \frac{\sqrt{n_{2,T}}}{\sqrt{n_{1,T}}}\,\sqrt{n_{2,T}}\,\widehat{\mu}_{1,\frac{T}{2}} > \frac{\sqrt{n_{2,T}}}{\sqrt{n_{1,T}}}\,\sqrt{2\log 1/\gamma} - \sqrt{2\log 1/\gamma}\right)
\end{aligned}$$

From lemma 1 it follows that $\sqrt{n_{2,T}}\,\widehat{\mu}_{2,\frac{T}{2}} - \sqrt{1-\epsilon}\,\sqrt{n_{2,T}}\,\widehat{\mu}_{1,\frac{T}{2}} \xrightarrow{d} N(0, 2-\epsilon)$. Therefore, from the last equality we have,

$$\lim_{T\to\infty} \mathrm{P}\left(U_2\left(\frac{T}{2}\right) > U_1\left(\frac{T}{2}\right)\right) = \Phi\left(-\frac{\sqrt{1+\epsilon}\,\sqrt{2\log 1/\gamma} - \sqrt{2\log 1/\gamma}}{\sqrt{2-\epsilon}}\right) > 0$$

This proves that the batched UCB algorithm is also unstable. Now, we shall prove that the batch Thompson sampling algorithm is also unstable. To prove this, we shall invoke proposition 1 in Appendix C of paper (Zhang et al. (2020)):

Lemma 16. *For identical two armed batch Thompson sampling, we have:*

$$\mathrm{P}(\tilde{\beta}_1 > \tilde{\beta}_2 | \mathcal{H}_{T/2}) \xrightarrow{d} U(0,1) \quad as \quad T \to \infty$$

As a consequence, we have:

$$\mathrm{P}(\mathrm{P}(\tilde{\beta}_1 > \tilde{\beta}_2 | \mathcal{H}_{T/2}) \geq \pi_{\max}) \to 1 - \pi_{\max}$$

We observe that $(\mathcal{A}_{2,i})_{1\leq i\leq T/2}$ is a mixture of i.i.d Bernoulli random variables indexed by the random variable $p_2(T)$. Therefore, there exists a triangular sequence of i.i.d random variables $(\mathcal{B}_{T,i})_{1\leq i\leq T/2}$ (independent of the bandit stategy) such that $\mathcal{B}_{T,i} \sim \mathrm{Ber}(\pi_{\max})$ and,

$$\mathcal{A}_{2,i} = \mathcal{B}_{T,i} \quad , \text{if} \quad p_2(T) = \pi_{\max}$$

Now, by applying similar arguments, we observe the following chain of inequalities:

$$\begin{aligned}
&\lim_{T\to\infty} \mathrm{P}\left(\left|\frac{1}{2}\left(\frac{2}{T}\sum_{j=1}^{T/2}\mathcal{A}_{2,j}\right) - \pi_{\max}\right| > \frac{\delta}{2}\right) \\
&= \mathrm{P}\left(\mathrm{P}(\tilde{\beta}_1 > \tilde{\beta}_2|\mathcal{H}_{T/2}) \geq \pi_{\max},\ \left|\frac{1}{2}\left(\frac{2}{T}\sum_{j=1}^{T/2}\mathcal{B}_{T,j}\right) - \frac{\epsilon}{2}\right| > \frac{\delta}{2}\right) \\
&= 1 - \pi_{\max} > 0
\end{aligned}$$

This concludes the proof that the batched Thompson Sampling strategy is unstable.

D Proofs of Auxiliary Results

In this section, we shall first outline the proof of instability for Anytime-MOSS, followed by the proof for KL-UCB-SWITCH. This is followed by proofs of Lemmas 2,6,9 and 17

Instability of Anytime-MOSS

The proof of instability of Anytime MOSS follows along similar lines of argument presented in Section 5 with some slight modification. Let $U_a^A(t), U_a^V(t)$ be the UCB indices of Anytime MOSS and Vanilla Moss (see Table 1). Note that as $t \leq T \leq TK$, the following holds:

$$U_a^A(t) = \widehat{\mu}_a(t) + \sqrt{\frac{\max\{0,\ \log(\frac{t}{Kn_{a,t}})\}}{n_{a,t}}} \leq \widehat{\mu}_a(t) + \sqrt{\frac{\log(\frac{T}{n_{a,t}})}{n_{a,t}}} = U_a^V(t) \tag{74}$$

Now let us consider event $\mathcal{E}_0^V(T)$ for defined as follows

$$\mathcal{E}_0^{A,V}(T) := \left\{ U_a^V\left(\frac{T}{2K}\right) < \lambda_T \ \forall a \in \{2, \ldots, K\},\ \min_{\frac{T}{2K} \leq t \leq T} U_1^A(t) > \lambda_T \right\}. \tag{75}$$

We note that $U_a^A(T/2K) \leq U_a^V(T/2K)$ and on the event $\{U_a^V(T/2K) < \lambda_T \ \ \forall a \geq 2\} \cap \{U_1^A(T/2K) > \lambda_T\}$, we have arm 1 is pulled at time $T/(2K) + 1$. Consequently, we have for all arms $a \geq 2$,

$$n_{a,\frac{T}{2K}+1}(\mathcal{A}) = n_{a,\frac{T}{2K}}(\mathcal{A}), \qquad \text{and} \qquad \widehat{\mu}_{a,\frac{T}{2K}+1} = \widehat{\mu}_{a,\frac{T}{2K}}.$$

Hence, it follows from Assumption (A2) that

$$U_a^V\Big(\frac{T}{2K} + 1\Big) \leq U_a^V\Big(\frac{T}{2K}\Big) < \lambda_T.$$

By applying this argument iteratively it follows that on event $\mathcal{E}_0^{A,V}(T)$ arm 1 is pulled at every step from $T/(2K)$ to T. Hence,

$$\mathcal{E}_0^{A,V}(T) \subset \left\{ \frac{n_{1,T}(\mathcal{A})}{T} \geq \frac{K+1}{2K},\ \frac{n_{a,T}(\mathcal{A})}{T} \leq \frac{1}{2K}\ \forall a \geq 2 \right\}. \tag{76}$$

where algorithm $\mathcal{A}$ is Anytime-MOSS. Therefore, as discussed in Section 5, once we prove

$$\liminf_{T\to\infty} \mathrm{P}(\mathcal{E}_0^{A,V}(T)) > 0$$

then instability of Anytime MOSS will follow. Furthermore, an inspection of the algorithm shows that it satisfies Assumption (A3). Now suppose that Anytime MOSS is *stable*. By Lemma 2 we have

$$\frac{n_{a,T/2K}}{T/2K^2} \xrightarrow{\mathrm{P}} 1$$

Consequently, we have $\mathrm{P}(\mathcal{E}(T)) \to 1$ as $T \to \infty$, and for any sequence of events $(A_T)_{T\geq 1}$,

$$\liminf_{T\to\infty} \mathrm{P}(A_T) = \liminf_{T\to\infty} \mathrm{P}(A_T \cap \mathcal{E}(T)). \tag{77}$$

In particular, for the choice of $A_T = \mathcal{E}_0^{A,V}(T)$ we obtain,

$$\liminf_{T\to\infty} \mathrm{P}(\mathcal{E}_0^{A,V}(T)) = \liminf_{T\to\infty} \mathrm{P}(\mathcal{E}_0^{A,V}(T) \cap \mathcal{E}(T)). \tag{78}$$

On the event $\mathcal{E}_0^{A,V}(T) \cap \mathcal{E}(T)$, Assumption (A3) implies that the upper confidence bounds $U_a^A(t)$ for all $a \in [K]$ lie between $\widetilde{U}_{a,\beta_1}^A(t)$ and $\widetilde{U}_{a,\beta_2}^A(t)$ where,

$$\widetilde{U}_{a,\beta}^A(t) = \widehat{\mu}_{a,t} + \frac{\beta}{\sqrt{n_{a,t}}}$$

Moreover, on $\mathcal{E}(T)$, there exist constants γ_1, γ_2 (depending on β_1, β_2) such that for each arm $a \in [K]$,

$$\widehat{\mu}_{a,t} + \frac{\gamma_1}{\sqrt{t}} \leq \widetilde{U}_{a,\beta_1}^A(t) \leq U_a^A(t) \leq \widetilde{U}_{a,\beta_2}^A(t) \leq \widehat{\mu}_{a,t} + \frac{\gamma_2}{\sqrt{t}}. \tag{79}$$

As done in Section 5, we define the *pseudo upper confidence bound* below

$$V_{a,\gamma}^A(t) := \widehat{\mu}_{a,t} + \frac{\gamma}{\sqrt{t}}.$$

We note that the rest of the proof is exactly identical to Section 5 and hence we conclude our proof.

Instability of Anytime KL-UCB-SWITCH

The proof of instability of Anytime-KL-UCB-SWITCH borrows ideas from the proof of Anytime MOSS. Let $U_a(t)$ be the UCB for arm $a \in [K]$. As $t \leq T \leq TK$, it follows that

$$U_a(t) \leq \begin{cases} \sup\Big\{q \in [0,1] :\ KL(\widehat{\mu}_a(t), q) \leq \dfrac{\phi\left(\frac{T}{n_{a,t}}\right)}{n_{a,t}}\Big\}, & \text{if } n_{a,t} \leq f(t,K), \\ \widehat{\mu}_a(t) + \sqrt{\dfrac{1}{2n_{a,t}}\,\phi\left(\frac{T}{n_{a,t}}\right)}, & \text{if } n_{a,t} > f(t,K), \end{cases} \tag{80}$$

where $\phi(x) := \log^+(x)(1 + (\log^+ x)^2)$ and $f(t,K) = [(t/K)^{1/5}]$.

Let us define two UCB indices for arm a as follows

$$U_a^{(1)}(t) := \widehat{\mu}_a(t) + \sqrt{\frac{1}{n_{a,t}}\,\phi\left(\tfrac{T}{n_{a,t}}\right)}, \quad U_a^{(2)}(t) := \widehat{\mu}_a(t) + \sqrt{\frac{1}{2n_{a,t}}\,\phi\left(\tfrac{T}{n_{a,t}}\right)}$$

Then due to inequality 16, the UCB index $U_a(t) \leq U_a^{(1)}(t)$. Motivated by this observation, we define the following event

$$\mathcal{E}_0^{(1)}(T) := \left\{ U_a^{(1)}\left(\frac{T}{2K}\right) < \lambda_T \ \forall a \in \{2,\dots,K\},\ \min_{\frac{T}{2K}\leq t\leq T} U_1(t) > \lambda_T \right\}. \tag{81}$$

We note the event $\mathcal{E}_0^{A,V}(T)$ in equation (75) is similar to $\mathcal{E}_0^{(1)}(T)$ in equation (81). Now analogous argument as the one provided in the proof of Anytime MOSS completes the proof.

Proof of Lemma 2

We begin the proof by observing that since the reward distributions of both the arms are identical, the following is true due to symmetry.

$$n_{1,T}(\mathcal{A}) \overset{\mathrm{d}}{=} n_{a,T}(\mathcal{A}) \quad \text{for all } a \in \{2, \ldots, K\}$$

Now, suppose stability holds. This implies that there exists real sequences $(n^*_{a,T}(\mathcal{A}))$ such that :

$$\frac{n_{a,T}(\mathcal{A}))}{n^*_{a,T}(\mathcal{A})} \xrightarrow{\mathrm{P}} 1 \quad \text{as} \quad T \to \infty$$

This implies that for any arm $a \in \{2, \ldots, K\}$,

$$\frac{n_{1,T}(\mathcal{A})}{n^*_{1,T}(\mathcal{A})} \overset{\mathrm{d}}{=} \frac{n_{a,T}(\mathcal{A})}{n^*_{1,T}(\mathcal{A})} \tag{82}$$

We note that stability implies that $n_{1,T}(\mathcal{A})/n^\star_{1,T}(\mathcal{A})$ converges to 1 in probability, which further implies in distribution convergence. Therefore from equation (82) it follows that

$$\frac{n_{a,T}(\mathcal{A})}{n^*_{1,T}(\mathcal{A})} \xrightarrow{d} 1 \implies \frac{n_{a,T}(\mathcal{A})}{n^*_{1,T}(\mathcal{A})} \xrightarrow{\mathrm{P}} 1 \tag{83}$$

Again, from stability we know that for any arm $a \in [K]$,

$$\frac{n_{a,T}(\mathcal{A})}{n^*_{a,T}(\mathcal{A})} \xrightarrow{\mathrm{P}} 1 \tag{84}$$

Therefore from equations (83) and (84) we obtain,

$$\frac{n^\star_{1,T}(\mathcal{A})}{n^*_{a,T}(\mathcal{A})} \to 1$$

This implies that we can assume that for each arm a, $n^*_{1,T}(\mathcal{A}) = n^*_{a,T}(\mathcal{A})$. This leads us to the following observation.

$$K = \lim_{T\to\infty} \sum_{a=1}^{K} \frac{n_{a,T}(\mathcal{A}))}{n^*_{a,T}(\mathcal{A})} = \lim_{T\to\infty} \sum_{a=1}^{K} \frac{n_{a,T}(\mathcal{A}))}{n^*_{1,T}(\mathcal{A})} = \lim_{T\to\infty} \frac{T}{n^*_{1,T}(\mathcal{A})}$$

Therefore, as $T \to \infty$, $n^*_{1,T}(\mathcal{A})/T$ converges to T/K.

Proof of Lemma 5

We shall prove this result by proving a similar result for Let us a sub-vector $\mathbf{L}_t$ defined below.

$$\mathbf{L}_t := \begin{bmatrix} \widehat{\mu}_{1,t} \\ \widehat{\mu}_{2,t} \\ n_{1,t} \\ n_{2,t} \end{bmatrix}$$

Therefore, ignoring the order in which the elements appear in the definition of $\mathbf{Z}_t$ it follows that:

$$\mathbf{Z}_t = \begin{bmatrix} V_{1,\gamma_1}(t) \\ V_{2,\gamma_2}(t) \\ \mathbf{L}_t \end{bmatrix}$$

Now, we recall that,

$$V_{a,\gamma}(t) := \widehat{\mu}_{a,t} + \gamma \sqrt{\frac{\log(T/2n_{a,t})}{n_{a,t}}}$$

Now, according to our algorithm structure, when we have two arms a_1 and a_2 we have the following:

$$n_{a_1,t+1} = n_{a_1,t} + 1\left\{U_{a_1}(t) > U_{a_2}(t)\right\},$$

$$\widehat{\mu}_{a_1,t+1} = \frac{n_{a,t}}{n_{a_1,t+1}}\widehat{\mu}_{a_1,t} + \frac{X_{t+1}}{n_{a_1,t+1}} 1\left\{U_{a_1}(t) > U_{a_2}(t)\right\}.$$

Now, let $X_{a,t} := \mu_a + \xi_{a,t}$ be the potential reward process for arm a. We note that for each a, process $(X_{a,t})_{t\geq 1}$ is independent of the algorithm. Furthermore, if arm a is chosen at time $t+1$ then from the definitions it follows that $X_{t+1} = X_{a,n_{a,t}+1}$. We claim that $X_{a_1,n_{a_1,t+1}}$ is independent of the history till time t. This may seem counter-intuitive as from the definition it looks like it is dependent on $n_{a,t+1}$. To see this we observe that,

$$X_{a,n_{a,t+1}} = \sum_{i=1}^{\infty} X_{a,i} 1\left\{n_{a,t+1} = i\right\}$$

This implies that,

$$\mathrm{P}(X_{a,n_{a,t+1}} \in B \mid n_{a,t+1} = k) = \mathrm{P}(X_{a,k} \in B \mid n_{a,t+1} = k) = \mathrm{P}(X_{a,k} \in B) = \mathrm{P}(X_{a,1} \in B)$$

Furthermore,

$$\begin{aligned}
&\mathrm{P}(X_{a,n_{a,t+1}} \in B) \\
&= \sum_{i=1}^{\infty} \mathrm{P}(X_{a,n_{a,t+1}} \in B \mid n_{a,t+1} = i) \times \mathrm{P}(n_{a,t+1} = i) \\
&= \sum_{i=1}^{\infty} \mathrm{P}(X_{a,1} \in B) \times \mathrm{P}(n_{a,t+1} = i) \\
&= \mathrm{P}(X_{a,1} \in B) \qquad (85)
\end{aligned}$$

Hence, for any measurable subset B and k we have:

$$\mathrm{P}(X_{a,n_{a,t+1}} \in B \mid n_{a,t+1} = k) = \mathrm{P}(X_{a_1,1} \in B)$$

Therefore, we have proved that $X_{a_1,n_{a_1,t+1}}$ is independent of $n_{a,t+1}$. Consequently, we can rewrite the adaptive mean of arm a_1 as follows:

$$\widehat{\mu}_{a_1,t+1} = f(\widehat{\mu}_{a_1,t}, n_{a_1,t}, U_{a_1}(t), X_{a_1,n_{a_1,t}}) \qquad (86)$$

where f is a deterministic measurable function. Furthermore,we also assume that the UCB indices $U_{a_1}(t)$ are also deterministic functions of the sample means and arm pulls (Assumption (A1)) as follows:

$$U_{a_1}(t) = h_{a_1}(\widehat{\mu}_{1,t}, \widehat{\mu}_{2,t}, n_{1,t}, n_{2,t}) \tag{87}$$

Therefore, from equations (86) and (87) it follows that there exists a real function $\psi : \mathbb{R}^2 \to \mathbb{R}$ such that:

$$\mathbf{Z}_t = \begin{bmatrix} \psi(\mathbf{L}_t) \\ \mathbf{L}_t \end{bmatrix}$$

Now, consider the following lemma.

Lemma 17. *Assume that $(Q_t)_{t\geq 1}$ is a finite dimensional discrete time Markov chain of dimension d. Also, suppose that ψ is any real valued function such that $\psi : \mathbb{R}^d \to \mathbb{R}^k$, where $k \in \mathrm{N}$. Define a new process M_t as follows:*

$$M_t = \begin{bmatrix} \psi(Q_t) \\ Q_t \end{bmatrix}$$

Then the process $(M_t)_{t\geq 1}$ is also a Markov chain.

Therefore, if we can show that $(\mathbf{L}_t)_{t\geq 1}$ is a Markov chain we are done. We note that from equations (87) and (86) it follows that there exists a real measurable function η such that,

$$\mathbf{L}_{t+1} = \eta(\mathbf{L}_t, X_{1,t}, X_{2,t})$$

where $X_{1,t}, X_{2,t}$ are the respective reward processes independent of $\mathbf{L}_t$. Consequently we deduce that,

$$\begin{aligned}
&\mathrm{P}(\mathbf{L}_{t+1} = \mathbf{s}_{t+1} \mid \mathbf{L}_m = \mathbf{s}_m,\ m = 1, 2, \ldots, t) \\
&= \mathrm{P}(\eta(\mathbf{L}_t, X_{1,t}, X_{2,t}) = \mathbf{s}_{t+1} \mid \mathbf{L}_m = \mathbf{s}_m,\ m = 1, 2, \ldots, t) \\
&= \mathrm{P}(\eta(\mathbf{s}_t, X_{1,t}, X_{2,t}) = \mathbf{s}_{t+1} \mid \mathbf{L}_m = \mathbf{s}_m,\ m = 1, 2, \ldots, t) \\
&= \mathrm{P}(\eta(\mathbf{s}_t, X_{1,t}, X_{2,t}) = \mathbf{s}_{t+1}) \\
&= \mathrm{P}(\eta(\mathbf{s}_t, X_{1,t}, X_{2,t}) = \mathbf{s}_{t+1} \mid \mathbf{L}_t = \mathbf{s}_t) \\
&= \mathrm{P}(\eta(\mathbf{L}_t, X_{1,t}, X_{2,t}) = \mathbf{s}_{t+1} \mid \mathbf{L}_t = \mathbf{s}_t) \\
&= \mathrm{P}(\mathbf{L}_{t+1} = \mathbf{s}_{t+1} \mid \mathbf{L}_t = \mathbf{s}_t)
\end{aligned}$$

The chain of equalities above imply that $(\mathbf{L}_{t+1})_{t\geq 1}$ is a Markov chain and hence, we are done.

Proof of Lemma 6

We know that $(\mathbf{Q}_t)_{t\geq 1}$ is a finite - dimensional, discrete time Markov Chain. Let A, B and C be arbitrary sets comprising of states in the state space. We are interested in the probability of the event $\{\mathbf{Q}_{t-1} \in A, \mathbf{Q}_{t+1} \in C\}$, *given* that event $\{\mathbf{Q}_t \in B\}$ holds. Now, the following chain of equalities follow :

$$\begin{aligned}
&\mathrm{P}(\mathbf{Q}_{t-1} \in A, \mathbf{Q}_{t+1} \in C | \mathbf{Q}_t \in B) \\
&= \frac{\mathrm{P}(\mathbf{Q}_{t-1} \in A, \mathbf{Q}_t \in B, \mathbf{Q}_{t+1} \in C)}{\mathrm{P}(\mathbf{Q}_t \in B)} \\
&= \mathrm{P}(\mathbf{Q}_{t+1} \in C | \mathbf{Q}_t \in B, \mathbf{Q}_{t-1} \in A) \left[\frac{\mathrm{P}(\mathbf{Q}_t \in B, \mathbf{Q}_{t-1} \in A)}{\mathrm{P}(\mathbf{Q}_t \in B)} \right] \\
&= \mathrm{P}(\mathbf{Q}_{t+1} \in C | \mathbf{Q}_t \in B) \times \mathrm{P}(\mathbf{Q}_{t-1} \in A | \mathbf{Q}_t \in B)
\end{aligned}$$

The last equality follows from the definition of Markov Chain due to the reason highlighted below.

$$\begin{aligned}
&\mathrm{P}(\mathbf{Q}_{t+1} \in C | \mathbf{Q}_t \in B, \mathbf{Q}_{t-1} \in A) \\
&= \frac{\mathrm{P}(\mathbf{Q}_{t+1} \in C, \mathbf{Q}_t \in B, \mathbf{Q}_{t-1} \in A)}{\mathrm{P}(\mathbf{Q}_t \in B, \mathbf{Q}_{t-1} \in A)} \\
&= \frac{\mathrm{P}(\mathbf{Q}_{t+1} \in C, \mathbf{Q}_t \in B, \mathbf{Q}_{t-1} \in A, \mathbf{Q}_k \in \mathbb{R}, k = 1, \ldots, t-2)}{\mathrm{P}(\mathbf{Q}_t \in B, \mathbf{Q}_{t-1} \in A)} \\
&= \frac{\mathrm{P}(\mathbf{Q}_{t+1} \in C | \mathbf{Q}_t \in B) \times \mathrm{P}(\mathbf{Q}_t \in B, \mathbf{Q}_{t-1} \in A, \mathbf{Q}_k \in \mathbb{R}, k = 1, \ldots, t-2)}{\mathrm{P}(\mathbf{Q}_t \in B, \mathbf{Q}_{t-1} \in A)} \\
&= \frac{\mathrm{P}(\mathbf{Q}_{t+1} \in C | \mathbf{Q}_t \in B) \times \mathrm{P}(\mathbf{Q}_t \in B, \mathbf{Q}_{t-1} \in A)}{\mathrm{P}(\mathbf{Q}_t \in B, \mathbf{Q}_{t-1} \in A)} \\
&= \mathrm{P}(\mathbf{Q}_{t+1} \in C | \mathbf{Q}_t \in B)
\end{aligned}$$

This proves our result when $m = 1$. The proof for the general case follows iteratively.

Proof of Lemma 9

Define $M(\alpha T, \beta T) := \max_{\alpha T \leq t \leq \beta T} Q_t$. We wish to prove that $\{M(\alpha T, \beta T) \geq \lambda\} \bigcap \{\tau \leq k\}$ belongs in the sigma field $\mathcal{F}_k$, for each natural number k. Note that $M(r_1, r_2) := \max_{t \in [r_1, r_2]}(-X^*_{1,t})$ and $\mathcal{F}_{1,t} := \sigma(\xi_{1,1}, \ldots, \xi_{1,t})$. Now, recall that the definition of τ is as follows :

$$\tau := \min\Big(\inf\{k \geq \alpha T | M(\alpha T, k) \geq \lambda\}, \beta T \Big)$$

Fix an arbitrary natural number k. We split the proof in the following cases:

1. Suppose $k < \alpha T$. In this case, $\{\tau \leq k\} = \phi$. Therefore, $\{M(\alpha T, \beta T) \geq \lambda\} \bigcap \{\tau \leq k\} = \phi$ which belongs in $\mathcal{F}_k$.

2. Suppose $k > \beta T$. Then $\{\tau \leq k\} = \Omega$ and hence $\{M(\alpha T, \beta T) \geq \lambda\} \bigcap \{\tau \leq k\} = \{M(\alpha T, \beta T) \geq \lambda\}$ which belongs in $\mathcal{F}_{\beta T}$ which lies in $\mathcal{F}_k$ as $k > \beta T$.

3. Suppose $\alpha T \leq k \leq \beta T$. In this case, we note that $\{\tau \leq k\} = \{M(\alpha T, k) \geq \lambda\}$ and vice-versa. As a consequence, $\{\tau \leq k\} = \{M(\alpha T, k) \geq \lambda\} \subseteq \{M(\alpha T, \beta T) \geq \lambda\}$. Therefore, $\{M(\alpha T, \beta T) \geq \lambda\} \bigcap \{\tau \leq k\} = \{M(\alpha T, k) \geq \lambda\}$ which depends only on random variables $\xi_{1,\alpha T}, \ldots, \xi_{1,k}$ and hence belongs in $\mathcal{F}_k$

Therefore, we conclude that for any $k \in \mathcal{N}$, $\{M(\alpha T, \beta T) \geq \lambda\} \bigcap \{\tau \leq k\}$ belongs in the sigma field $\mathcal{F}_k$.

Proof of Lemma 17

We assume that Q_t is a d-dimensional discrete time Markov chain. We recall that,

$$M_t = \begin{bmatrix} \psi(Q_t) \\ Q_t \end{bmatrix}$$

where ψ is a real-valued function. Fix an event A in the extended state space of $(M_t)_{t \geq 1}$. Let $\mathbf{v}$ be an arbitrary element of A such that,

$$\mathbf{v} = \begin{bmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \end{bmatrix}$$

where $\mathbf{v}_1, \mathbf{v}_2$ are d-dimensional and k dimensional vectors respectively. This implies that,

$$M_t = \mathbf{v} \text{ if and only if } \quad \psi(Q_t) = \mathbf{v}_1 \ , \ Q_t = \mathbf{v}_2$$

This implies that,

$$\left\{ \begin{bmatrix} \psi(Q_t) \\ Q_t \end{bmatrix} = \begin{bmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \end{bmatrix} \right\} = \begin{cases} \{Q_t = \mathbf{v}_2\}, & \text{if } \mathbf{v}_2 \in \psi^{-1}(\mathbf{v}_1) \\ \phi, & \text{otherwise} \end{cases}$$

Motivated by this observation, for set A we define set B_A as follows:

$$B_A := \left\{ \mathbf{v}_2 : \begin{bmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \end{bmatrix} \in A \text{ and, } \mathbf{v}_2 \in \psi^{-1}(\mathbf{v}_1) \right\}$$

Therefore,

$$\left\{ \begin{bmatrix} \psi(Q_t) \\ Q_t \end{bmatrix} \in A \right\} = \{Q_t \in B_A\}$$

Now, let $A_1, A_2, \ldots, A_{t+1}$ be arbitrary sets in the extended state space. Now, suppose for some i, $M_i \in A_i$. This implies that there exists sets B_{A_i} such that:

$$\{M_i \in A_i\} = \{Q_i \in B_{A_i}\}$$

As Q_t is a Markov chain, from the equation above we have:

$$\begin{aligned}
\mathrm{P}\Big(M_{t+1} \in A_{t+1} \Big| M_m \in A_m, m = 1, 2, \ldots, t\Big) &= \mathrm{P}\Big(Q_{t+1} \in B_{A_{t+1}} \Big| Q_m \in B_{A_m}\ , m = 1, 2, \ldots, t\Big) \\
&= \mathrm{P}\Big(Q_{t+1} \in B_{A_{t+1}} \Big| Q_t \in B_{A_t}\Big) \\
&= \mathrm{P}\Big(M_{t+1} \in A_{t+1} \Big| M_t \in A_t\Big)
\end{aligned}$$

Hence, we conclude that $(M_t)_{\geq 1}$ is also a Markov chain.