Asymptotical properties of sequences of Multi-armed bandits under UCB algorithms*

Yifei Ouyang, Yuqiang Li and Xianyi Wu†
School of Statistics, KLATASDS-MOE, East China Normal University,
Shanghai 200062, PR China

March 25, 2025

Abstract: In this paper, we theoretically analyze the sampling behaviours for a sequence of two-armed bandits under conventional UCB algorithms. Assume the sequence of two-armed bandits has horizon-dependent gaps Δ_n. It is proved under certain conditions that the sequence of selected numbers of the optimal arm $N^*(n)$ has the property that $N^*(n)/n$ converges almost surely to 1, meanwhile the the sequence of selected numbers of the sub-optimal arm $N_*(n)$ has the property that $N_*(n)\Delta_n^2/\log n$ converges in probability to a constant dependent on the algorithm. As a result, we get the asymptotical limit of regrets and regret processes as well. To assess theoretical performances, several comparable experiments are conducted and discussed.

Keywords: Multi-armed bandit, UCB algorithm, Sub-optimal arm, Diffusion limit

1 Introduction

In sequential decision making problems, there exist quintessential exploration vs. exploitation trade-offs. This is a balance between staying with the option that gave highest payoffs in the past and exploring new options that might give higher payoffs in the future. Multi-armed bandit problems are the most basic examples derived from such dilemmas. This concept originated from a clinical trial study in 1933 [18] which introduced one of the earliest heuristic algorithms Thompson Sampling for multi-armed bandits [1]. In the classical stochastic k-armed bandit problem, a learner pulls one of k arms sequentially at each time $t \in \{1, 2, \ldots\}$, and the environment, according to the corresponding arm-dependent distribution of that arm, reveals a reward X_t. The most common objective of the

*This research is supported by National Key R&D Program of China (Nos. 2021YFA1000100 and 2021YFA1000101) and Natural Science Foundation of China (No. 72371103)

†Corresponding authors: Yuqiang Li at yqli@stat.ecnu.edu.cn and Xianyi Wu at xywu@stat.ecnu.edu.cn.

learner is to choose actions that lead to the largest possible cumulative reward over all n rounds, which is $\sum_{t=1}^{n} X_t$. Initially ignorant of the environment, the learner must navigate the balance between exploring new arms and exploiting the best arm played by far. Today, the MAB manifests itself in various forms and finds applications across a wide spectrum, including advertising placement, dynamic pricing, online auctions, e-commerce and matching markets among others [6].

In its most basic formulation, i.e. the stochastic stationary multi-armed bandit model, a Multi-armed bandit problem with K-arms is defined by a sequence of distributions P_i, $1 \leq i \leq K$, where each i is the index of an action (i.e., the arm of a bandit) and the distributions are unknown (in general, we call $\mathcal{P} := (P_1, P_2, \cdots, P_k)$) an environment). Successive plays of machine i yield rewards $X_{i,1}, X_{i,2}, \cdots$, which are independent and identically distributed according to P_i with unknown expectation μ_i. Independence also holds for rewards across machines; i.e., $X_{i,m}$ and $X_{j,m}$ are independent (and usually not identically distributed) for each $1 \leq i < j \leq K$ and each $m, n \geq 1$. During the interaction at round t, the learner selects an action (arm) $A_t \in \mathcal{A}$ according to a policy or an algorithm π, and then the environment responds to this action by providing the reward $X_t = X_{A_t, N_{A_t}(t)}$, where $N_i(t)$ is the number of times arm i was chosen after the end of round t, i.e., $N_i(t) := \sum_{s=1}^{t} \mathbb{1}\{A_s = i\}$. For convenience, we denote the MAB problems as $(K, \mathcal{P})$. If it has an end round T, i.e. after the T-th round, the interaction ends, we call T the horizon and denote the MAB problems as $(K, \mathcal{P}, H)$. The arm with the highest reward mean is referred to as the *optimal arm* and the highest reward mean is denoted by μ^*. The remaining arms, collectively referred to as *suboptimal arm*s.

The difference between the mean reward of the optimal arm and that of any suboptimal arm j is termed the sub-optimality gap, or simply the *gap*, of arm j, denoted as $\Delta_j := \mu^* - \mu_j$. Natural measures of the performance of an algorithm π after t rounds are the quantities below:

$$\hat{R}_\pi(t) := \sum_{k=1}^{K} \Delta_k N_k(t), \qquad R_\pi(t) = \sum_{k=1}^{K} \Delta_k \mathbb{E}(N_k(t)),$$

where the expectation is with respect to the randomness in the environment and policy and for notational simplicity their dependence on the unknown distributions is usually suppressed. The functions $\hat{R}_\pi(n), R_\pi(n)$ in the literature have been called pseudo-regret and regret, respectively.

There are much literature in the field of Multi-armed bandits. Most of them are about designing policies for different kinds of environment classes and estimating their optimality via finite time bounds and/or asymptotes of the regrets. The upper confidence bound (UCB) algorithm is the most commonly used algorithm in MAB problems and is increasingly used in reinforcement learning (see, for example, [4], [10] and the reference therein), which is based on the principle of optimism in

the face of uncertainty and states that one should act as if the environment is as nice as plausibly possible. Lai and Robbins [15] first used the confidence bounds and the idea of optimism, analysed the asymptotes for various parametric bandit problems, showed that the order of the smallest achievable regret is logarithmic in the horizon with constants related to the gaps. The first version of UCB is by Lai [14]. Auer et al [3] proposed a simple but widely used version of UCB under the name of UCB1 and proved that the optimal logarithmic regret is also achievable uniformly over any finite horizon for all reward distributions with bounded support. Moreover, the UCB-V algorithm of Audibert et al[2] takes into account the variance of the distributions and later, Garivier and Cappé [11] and Maillard et al. [17] independently proposed the KL-UCB algorithm which is shown to attain the optimal rate $\log T$ for any finite horizon. In addition, it is also proved that no algorithm can achieve an expected regret smaller than $c\sqrt{T}$ for any given horizon T (the constant hides dependence on the number of arms) uniformly over all problem instances (also called minimax regret). Theoretically, focusing purely on expected regret minimization is not enough to characterize the fluctuation of algorithms and evaluate their behaviours. Audibert et al [2] provides high-probability bounds on pseudo-regrets under a parametric family of UCB1 algorithms. Fan and Glynn [9] developed approximations to the regret distribution and found that Thompson sampling and UCB satisfy the same Strong laws of large number (SLLN) and Central limit theorem (CLT), with the asymptotes of both the SLLN and the (mean) centering sequence in the CLT matching the asymptotes of expected regret. Both the mean and variance in the CLT grow at $\log(T)$ rates. For more knowledge and literature on MAB please refer to the comprehensive book by Lattimore and Szepesvari [16].

The celebrated result of Lai and Robbins [15] shows that the regret minimization in the stochastic MAB problem is governed by the reciprocal of the minimal gap Δ and therefore, in the most of literature discussing the instance-dependent bounds of regrets, the gaps are assumed as positive constants independent of the horizon. However, algorithms become more complexity when the gap varies as horizon t, especially, the gap goes to zero as the horizon go to infinity. For example, when we discuss the uniformly upper bound of UCB regrets, a borderline is $\Delta \asymp \sqrt{\frac{\log t}{t}}$ and when we discuss the minimax bound of regrets, in the worst case, the gap are set as $\Delta \asymp 1/\sqrt{t}$ (see Lattimore and Szepesvari [16, Chapter 7 and Chapter 15]). Note that when it is emphasized that the gaps are functions of horizon, one is interested in sequences of MAB problems in some extend.

Naturally, from a theoretical perspective, one may wonder how the approximating-to-zero gaps effect the finite-time behaviours of algorithms, which differs from the asymptotical analyses in existing literature where the gap is fixed when horizon goes increasingly and are more complicated because we need explore at the same time the finite-time behaviours and the asymptotical rules as the horizon goes increasingly. These questions also come from some interesting statistical inferences and tests.

For example, for an unknown distribution Q, its statistical inference $\tilde{Q}$ can be obtained from sampling by a method and hence depends on the sample size n and the method. Considering an interesting thought experiment, if we test or compare the two distributions Q and $\tilde{Q}$ in a sequential adaptive test, what will happen or what can we get from the test's performance? Intuitively, answers will depend on the difference between Q and $\tilde{Q}$, which generally is related to the sample size n, as well as the duration H of sequential testing.

However, in existing literature there is rarely convincing quantitative and theoretical research on this kind of problems. A unique paper is related to Kalvit and Zeevi [12] where they investigated the arm-sampling behaviors under the assumption that the gaps vary on the horizon. For the UCB1 algorithm, considering a sequence of MABs $(2, \mathcal{P}_n, n)_{n\geq 1}$, they proved that if the gap of the n-th MAB $(2, \mathcal{P}_n, n)$ is $\Delta_n = \omega(\sqrt{\frac{\ln n}{n}})$ (Large gap), then for the n-th MAB, the number of times the optimal arm is chosen in the entire n rounds $N^*(n)$ satisfies that $N^*(n)/n \to 1$ in probability, otherwise, $\lim_{n\to\infty} N^*(n)/n \in (1/2, 1)$ or $= 1/2$ in probability depends on $\Delta_n \sim \sqrt{\frac{\theta \ln n}{n}}$ (Moderate gap, where θ is a positive constant) or $o(\sqrt{\frac{\ln n}{n}})$ (Small gap), respectively. They also provided the first complete process-level characterization of the MAB problem under UCB in the conventional diffusion scaling among other results. By the way, this paper justified the necessary of discussing small gaps by more numerical experiments and potential applications.

We remark that Kalvit and Zeevi [12] focused on the optimal arm's sampling characteristics and get accurate asymptotical speeds of pseudo-regrets in the cases of moderate or small gaps (Theorem 4 and 5 therein). But they left the case of large gaps. The one of main obstacles, in our opinion, is that the sub-optimal arm's sampling characteristics are not accurate by far in the case of large gaps, meanwhile the major driver of the regret performance of an algorithm is its sub-optimal arm sampling. In this paper, we try to fill the blank.

For simplicity, we will focus on the behaviours of a sequence of 2-armed bandit problems with large gaps under a UCB algorithm (see Section 2 below). Heuristically, we can image that when the gap Δ_n decrease from a constant to $\sqrt{\frac{\ln n}{n}}$, the sub-optimal arm is chosen more and more frequently. To find some concise and proper conclusions with an universal method is the main difficulty. Our results show that under certain conditions, the sub-optimal arm sampling $N(n)$ satisfies that $N(n)\Delta_n^2/\log n \to \rho$ in probability, where ρ is a constant dependent on the algorithm. When Δ_n is degenerated to a positive constant, this results is consistent with the existing asymptotical one which holds for the single MAB (see [9]). Based on this results, we also get the asymptotical results of regret.

Our contributions in this paper have two folds. We conduct a more detailed investigation on the behavior of a sequence of MAB problems under the so-called large gap conditions. An interesting quantitative relationship is identified between the number of samples allocated to the suboptimal

arm by the algorithm and the gap itself. We also present the characteristics of regret under these conditions using diffusion limits. These new results also hold for the single MAB problem and hence extend the existing asymptotical theory. Our study deeps the understanding of a sequence of MAB problems under UCB algorithms and provide a new angle to discuss behaviours of algorithms, especially in a sequence of environments. In methodology, as that Cowan and Katehakis [7] observed, the Law of the iterated logarithm (LIL) for reward sequences is the critical factor to get the growth rate of the regret, see also for example Fan and Glynn [9] and Kalvit and Zeevi [12]. However, when we consider a sequence of MABs, the reward sequences in essence are consisted of triangular arrays, which can not assure the LIL. To overcome this difficulty, we adopt a different methods by combining Donsker's invariant principle, continuous mapping theorem and Slutsky Lemma, which provides a new scheme of methods in some sense. For more details please refer to our main theorems and the corresponding Remarks in Section 3.

This paper is divided into 5 sections. The introduction, Section 1, primarily delineates the research background, motivations, and the contributions of this paper. The specifics of the model, algorithm and notation are detailed in Section 2. Sections 3 is the theoretical core of the paper, where we report the main theoretical results. Some complicated proofs are deferred to Appendices. The subsequent Section presents numerical simulation results. Finally, concluding remarks and open problems are presented in Section 5.

2 Notation, model and algorithm

For any positive natural number n, let $[n] = \{1, 2, 3, \cdots, n\}$. We say $a_n = o(b_n)$ or $b_n = \omega(a_n)$ if $\{a_n\}_{n\geq 1}$ and $\{b_n\}_{n\geq 1}$ are two sequences of real numbers and $\lim_{n\to\infty} \frac{a_n}{b_n} = 0$. Similarly, $a_n = O(b_n)$ or $b_n = \Omega(a_n)$ if $\limsup_{n\to\infty} \left|\frac{a_n}{b_n}\right| \leq C$ for some constant C. If $a_n = O(b_n)$ and $a_n = \Omega(b_n)$ hold simultaneously, we say $a_n = \Theta(b_n)$, and we write $a_n \asymp b_n$ and $a_n \sim b_n$ in the special cases where $\lim_{n\to\infty} \frac{a_n}{b_n} = c \in (0, +\infty)$ and $\lim_{n\to\infty} \frac{a_n}{b_n} = 1$, respectively. If either sequence $\{a_n\}_{n\in\mathbb{N}_+}$ or $\{b_n\}_{n\in\mathbb{N}_+}$ is random, and one of the aforementioned ratio conditions holds in probability, we use the subscript p with the corresponding Landau symbol. For example, $a_n = o_p(b_n)$ if $a_n/b_n \xrightarrow{p} 0$ as $n \to \infty$. Besides, the notation "$\xrightarrow{d}$" signifies convergence in distribution, while "$\Rightarrow$" denotes weak convergence. Lastly, the notations "$\lceil\cdot\rceil$" and "$\lfloor\cdot\rfloor$" are used herein to represent the ceiling function (rounding up) and the floor function (rounding down), respectively.

We call a probability distribution P σ-sub-Gaussian, if there exists a positive number σ such that for any $\lambda \in \mathbb{R}$, it holds that $\mathbb{E}(\exp(\lambda(X - \mathbb{E}(X)))) \leq \exp(\lambda^2\sigma^2/2)$, where X is a random variable following the distribution P. In a bandit, if for any arm $i \in \mathcal{A}$, the reward distribution P_i is

σ-sub-Gaussian, we say the bandit model is a *σ-sub gaussian bandit*.

For simplicity, in this paper we focus on a sequence of stochastic 2-armed sub gaussian bandits $(2, \mathcal{P}_n, n)$. For each n, the bandit problem $(2, \mathcal{P}_n, n)$ has the arm set $\mathcal{A} = \{1, 2\}$, the horizon n and the environment $\mathcal{P}_n = \{P_{n,1}, P_{n,2}\}$ where $P_{n,1}$, $P_{n,2}$ are sub-Gaussian with (unknown) parameter σ and unknown means $\mu_{n,1}$, $\mu_{n,2}$, respectively. We always assume that $\mu_{n,1} \neq \mu_{n,2}$ and denote the gap between the two arms by Δ_n, i.e. $\Delta_n = |\mu_{n,1} - \mu_{n,2}|$.

During the interaction at round t ($t \in \{1, 2, \ldots, n\}$), the learner selects an action (arm) $A_{n,t} \in \mathcal{A}$ and receives a reward $X_{n,t}$ where the action is chosen according to a policy or algorithm $\pi = (\pi_1, \cdots, \pi_n)$ which is a sequence of probability on $\mathcal{A}$ such that the conditional distribution of $A_{n,t}$ given $A_{n,1}, X_{n,1}, \ldots, A_{n,t-1}, X_{n,t-1}$ is $\pi_t(\cdot | A_{n,1}, X_{n,1}, \ldots, A_{n,t-1}, X_{n,t-1})$. Denote the sequence of rewards associated with the pulls of arm i by $(X_{n,i,j})_{j=1,2,\ldots}$, with $X_{n,i,j}$ representing the reward received the j-th time arm i is sampled. $(X_{n,i,j})_{i,j}$ are independent each other.

Let $N_{n,i}(t)$ be the number of times that the arm i was selected after the end of round $t \leq n$, i.e. $N_{n,i}(t) = \sum_{s=1}^{t} \mathbf{1}_{\{A_{n,t}=i\}}$. Then $X_{n,t} = X_{n,A_{n,t},N_{A_{n,t}}(n,t)}$ and the corresponding sample mean of the rewards

$$\bar{\mu}_{n,i}(t) := \frac{\sum_{j=1}^{N_{n,i}(t)} X_{n,i,j}}{N_{n,i}(t)}.$$

The following algorithm 1 is a classical and relatively simple version of UCB algorithm. where

Algorithm 1 UCB(δ) algorithms for k-armed bandits

1: **Input:** Confident level δ.
2: At $t = 1, 2, \ldots, k$, play each arm $i \in \{1, 2, \ldots, k\}$ once.
3: **for** $t \in \{k+1, k+2, \ldots, \}$ **do**
4: Play arm $A_t \in \underset{i \in \{1,2,\ldots,k\}}{\operatorname{argmax}} \left(\bar{\mu}_i(t-1) + \sqrt{\frac{2 \log 1/\delta}{N_i(t-1)}} \right)$ (Ties are broken uniformly)

$\bar{\mu}_i(t)$ is the sample mean of the rewards of the i-th arm after the round t. In general, we call

$$UCB_i(t) := \bar{\mu}_i(t-1) + \sqrt{\frac{2 \log 1/\delta}{N_i(t-1)}}$$

the UCB index of arm i at round t, and call

$$B_i(t) = \sqrt{\frac{2 \log 1/\delta}{N_i(t-1)}}$$

the exploration bonus or confidence width of arm i at round t. Lattimore and Szepesvári introduced this algorithm in [16, Chapter 7]) and analysed the regret's upper bound for 1-sub-Gaussion MAB under this algorithm with $\delta = n^2$ where n is the horizon. In this paper, we choose the δ in the form

of $1/n^{\rho/2}$. In other words, we introduce a coefficient $\rho \in \mathbb{R}_+$ such that

$$B_i(t) = \sqrt{\frac{\rho \log n}{N_i(t-1)}}, \qquad UCB_i(t) = \bar{\mu}_i(t-1) + B_i(t)$$

for all $t \leq n$. We denote this UCB algorithm by UCB(ρ), where the parameter ρ may adjust the arm-exploring rate. The version of UCB discussed here is most similar to that analysed by Auer et al.[3] under the name UCB1, but that algorithm used t rather than n in the exploration bonus.

We denote the regret of $(2, \mathcal{P}_n, n)$ after round $t \leq n$ under the algorithm UCB(ρ) by

$$R_\rho(n,t) = \Delta_n \mathbb{E}(N_{n,j}(t))$$

where $j = \operatorname{argmin}_{i=1,2} \mu_{n,i}$ is the sub-optimal arm. Similarly, we denote the pseudo-regret by

$$\hat{R}_\rho(n,t) := \Delta_n N_{n,j}(t).$$

In this paper, we also discuss the random regret

$$\tilde{R}_\rho(n,t) = \sum_{s=1}^{t} \left(\max\{\mu_{n,1}, \mu_{n,2}\} - X_{n,s}\right).$$

To simplify the notations, we write $N_{n,i}(n)$, $R_\rho(n,n)$, $\hat{R}_\rho(n,n)$ and $\tilde{R}_\rho(n,n)$ as $N_i(n)$, $R_\rho(n)$, $\hat{R}_\rho(n)$ and $\tilde{R}_\rho(n)$, respectively.

3 Theoretical results

First of all, we note that the regret is essentially dependent on the gap and the sub-optimal arm's selected numbers. Kalvit and Zeevi [12] obtained the asymptotes of the optimal arm's selected numbers under the UCB1 algorithm. However, in the large gaps, their result is not enough to get accurate asymptotes of the sub-optimal arm's selected numbers. In this section, we will provide theoretical answers for the following questions: in the large gap case, as $n \to \infty$,

(1) what is the asymptotical characteristic of the sub-optimal arm's selected numbers?

(2) what is the asymptotical limit of regret sequence $R_\rho(n)$ or pseudo regret sequence $\hat{R}_\rho(n)$?

(3) what is the asymptotical limit of random-regret sequence $\tilde{R}_\rho(n) = n\mu_1 - \sum_{s=1}^{n} X_{n,t}$?

Without loss of generality, in the sequel, we always assume the arm 1 is the optimal arm. The following assumption essentially say that Δ_n is in the large gap case.

Assumption 1. *Assume that there exists a constant* $\alpha \in [0, 1/2]$ *such that* $h(n) := n^{\alpha}\Delta_n$ *is* $o(n^{\epsilon})$ *for any* $\epsilon > 0$. *In addition, if* $\alpha = 1/2$, $h(n) = \omega(\sqrt{\ln n})$.

Furthermore, to fertilize the theoretical analyses, we need the following technical conditions on the function $h(n)$.

Assumption 2. *The sequence* $\{h(n)\}_{n\geq 1}$ *can be extended to a positive function* $h(\cdot)$ *on* $(0, +\infty)$, *which satisfies the following conditions.*

(1) $h'(t) \geq 0$ *for sufficiently large* t,

(2) $\alpha h(t) \geq t h'(t)$ *for sufficiently large* t,

(3) $\lim_{t\to\infty} h(t)/t^{\beta} = 0$ *for any* $\beta > 0$.

We get the asymptotical characteristic of the sub-optimal arm's selected numbers as follows.

Theorem 1. *Apply the algorithm UCB(*ρ*) as shown in Algorithm 1 on a sequence of two-armed* σ*-subgaussian bandit problem* $(2, \mathcal{P}_n, n)$. *Suppose assumptions 1 and 2 hold. Then for any* $\rho > 4\sigma^2$, *as* $n \to \infty$,

$$\frac{N_1(n)}{n} \to 1 \quad a.s., \qquad \frac{N_2(n)\Delta_n^2}{\log n} \xrightarrow{p} \rho.$$

The proof of Theorem 1 is deferred to Appendix B.

Since distributions on bounded intervals are also sub-gaussian distributions, the aforementioned theorem can be applicable to bandits where the reward distributions of all arms are bounded. This leads to the following corollary:

Corollary 1. *In a two-armed bandit, if the reward distributions for both arms lie within the interval* $[a, b]$, *using the UCB(*ρ*) algorithm with* $\rho > (b-a)^2$, *the same conclusion as in Theorem 1 can be deduced.*

Proof: It is well known that probability distributions within the interval $[a, b]$ are $\frac{b-a}{2}$-subgaussian distributions (see, Lattimore and Szepesvári[16, Chapter 5]). Consequently, the condition $\rho > (b-a)^2$ indeed fulfills the prerequisites of Theorem 1. □

Below are some remarks on these results.

Remark 1. *Kalvit and Zeevi [12] studied the asymptotical limits of a sequence of MABs with rewards in* $[0, 1]$. *The UCB type algorithm in their paper is not UCB(*ρ*) but UCB1, which, for comparisons, is stated below and denoted by UCB1(*ρ*). When* $\rho > 1$, *[12] proved that, among other results, if the gap is large,* $N_1(n)/n \to 1$ *in probability. Meanwhile, Fan and Glynn [9] proved that if a two-armed*

Algorithm 2 UCB1(ρ) for k-armed bandits

Input: Exploration coefficient $\rho \in \mathbb{R}_+$.
At $t = 1, 2, \ldots, k$, play each arm $i \in \{1, 2, \ldots, k\}$ once.
for $t \in \{k+1, k+2, \ldots, H\}$ **do**
    Play arm $A_t \in \underset{i \in \{1,2,\ldots,k\}}{\operatorname{argmax}} \left(\bar{\mu}_i(t-1) + \sqrt{\frac{\rho \log t}{N_i(t-1)}} \right)$ (Ties are broken uniformly)

MAB has Gaussian rewards with variance 1*, denoting the gap constant by* Δ*, under the algorithm* $UCB1(2)$*, the number of pulling the sub-optimal arm* $N_2(t)$ *almost surely has an asymptotical rate* $2 \log n / \Delta_n$*. Namely, as the round* t *goes to infinity,* $N_2(t)\Delta^2 / \log t \to 2$ *almost surely. Our results strengthen Kalvit and Zeevi's result on* $N_1(n)$ *from convergence in probability to convergence almost surely, and generalize the Fan and Glynn's result from Gaussian rewards to sub-Gaussian rewards (our results are valid on the single MAB problem because a single MAB problem with infinite horizon can be looked as a sequence of MAB problems with increasing horizons and constant gaps fulfill our assumptions), though we take a slightly different UCB algorithm. Our result shows that even we apply the UCB(*δ*) algorithm on the MAB problems with varied environments and varied horizons, it can exhibit stability as same as applied to a fixed environment.*

Remark 2. *The reason why we take UCB(*ρ*) rather than UCB1(*ρ*) mainly comes from technology and methodology. Intuitively, because what we concern is universal natures of MABs exhibited in their finite horizon, whose relationship we have not made any assumptions about except the longer horizons and smaller gaps, to find some meaningful results their regular behaviours shall be recognized in an understandable manner as soon as possible. While we believe for long horizon* n *and after large* $1 << t \leq n$ *rounds, UCB(*ρ*) and UCB1(*ρ*) are very similar in behaviours, the relatively small exploration in UCB1(*ρ*) bring us extra obstacles, especially when* $t << n$*, since as the gap becomes smaller, the small exploration is harder to bring us quickly out the trip of sub-optimal arm being mistakenly evaluated as the optimal arm.*

Remark 3. *In order to get the asymptotical results on* $N_2(t)$*, it is required that* $\rho > 4\sigma^2$*. When the rewards are bounded in* $[0, 1]$*, it is* $\rho > 1$*, which is consistent with the setting in Kalvit and Zeevi [12]. There is a gap compared to the setting when studying the asymptotical rate of regret for a single MAB problem (see, for example, Fan and Glynn [9] with* $\rho = 2\sigma^2$*). For a single MAB problem, how* ρ *effects the corresponding regret was discussed in [2], where they reported that given the UCB1(*ρ*) algorithm, the greater the* ρ*, the thinner the tail on the pseudo-regret.*

Remark 4. *Cowan and Katehakis [7] observed that the Law of the iterated logarithm (LIL) for reward sequences is only really required for the derivation of the regret remainder term bounds when the asymptotes of algorithms' regret are considered. Fan and Glynn [9] successfully applied the LIL to*

get their almost surely asymptotical rate of $N_2(t)$. Kalvit and Zeevi [12] also directly borrowed this tool to their work on a sequence of MABs, however, it maybe need more explanations. In fact, when we consider a sequence of MABs $(2, \mathcal{P}_n, n)$, the reward sequences in essence are consisted of triangular arrays—for a given n, the possible reward sequences are $X_{n,1,i}, X_{n,2,i}$ where $1 \leq i \leq n$. However, the LIL is not always true for triangular arrays. A simple counterexample can be constructed as follows. If $\{X_{n,i},\ \ n \geq 1, 1 \leq i \leq n\}$ is a set of independent random variables with the same standard normal distribution, then

$$\liminf_{n\to\infty} \frac{\sum_{i=1}^n X_{n,i}}{\sqrt{n \log\log n}} = +\infty. \tag{1}$$

In fact, by some direct computations, we have that

$$\mathbb{P}\left(\frac{\sum_{i=1}^n X_{n,i}}{\sqrt{n\log n}} \geq \sqrt{2}\right) = \int_{\sqrt{2\log n}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} dx \geq \frac{1}{2\sqrt{2\pi \log n}} \frac{1}{n}(1 - \frac{1}{n}),$$

and hence

$$\sum_{n=1}^{\infty} \mathbb{P}\left(\frac{\sum_{i=1}^n X_{n,i}}{\sqrt{n\log n}} > \sqrt{2}\right) = +\infty,$$

which plus Borel-Cantalli Lemma shows that

$$\liminf_{n\to\infty} \frac{\sum_{i=1}^n X_{n,i}}{\sqrt{n\log n}} \geq \sqrt{2}.$$

In this paper we abandon the LIL tool. Instead, we use a new scheme—a combinator of Donsker's invariant principle, continuous mapping theorem and Slutsky Lemma, to get the asymptotics of $N_2(n)$. For details please see the proof of Theorem 1 in Appendix B.

Furthermore, from Theorem 1, we can get the asymptotical rate of the regret $R_\rho(n)$ and the pseudo regret $\hat{R}_\rho(n)$.

Theorem 2. *Under the assumption of Theorem 1, $\hat{R}_\rho(n) \sim_p \rho \log n/\Delta_n$ and $R_\rho(n) \sim \rho \log n/\Delta_n$.*

Proof: Pseudo-regret $\hat{R}_\rho(n) \sim_p \rho \log n/\Delta_n$ is a plain corollary of Theorem 1. In addition, from Theorem 1 we have that

$$\liminf_{n\to\infty} \frac{R_\rho(n)\Delta_n}{\log n} = \liminf_{n\to\infty} \Delta_n^2 \mathbb{E}\left(\frac{N_2(n)}{\log n}\right) \geq \rho.$$

On the other hand, for any $l \in (0,1)$, from (22) in Appendix A,

$$R_\rho(n) = \Delta_n \mathbb{E}(N_2(n)) \leq \Delta_n \left(\frac{\rho \log n}{\Delta_n^2 (1-2l)^2} + 1 + \mathbb{E}(G_{n,2}) + \mathbb{E}(G_{n,3})\right).$$

Furthermore,

$$\mathbb{E}(G_{n,2}) \leq \sum_{m=s_1(n)}^{n} \mathbb{P}\left(\bar{\mu}_{n,2}(m) > \mu_{n,2} + \delta\right) \leq \sum_{m=s_1(n)}^{n} \exp\left\{-\frac{m\delta^2}{2\sigma^2}\right\}$$
$$\leq \frac{1}{1-e^{-\frac{\delta^2}{2\sigma^2}}} \exp\left\{-\frac{s_1(n)\delta^2}{2\sigma^2}\right\} \leq \frac{2\sigma^2}{l^2\Delta_n^2} n^{-\frac{\rho l^2}{2\sigma^2}},$$

and

$$\mathbb{E}(G_{n,3}) \leq \sum_{k=s_1(n)}^{n} \sum_{t=k+1}^{n} \mathbb{P}\left(\bar{\mu}_{n,1}(t-k) < \mu_{n,1} - \sqrt{\frac{\rho \log n}{t-k}}\right)$$
$$\leq n^2 \exp\left\{\frac{-\rho \log n}{2\sigma^2}\right\} = n^{2-\frac{\rho}{2\sigma^2}}.$$

Hence,

$$\limsup_{n\to\infty} \frac{R_\rho(n)\Delta_n}{\log n} \leq \frac{\rho}{(1-2l)^2} + \limsup_{n\to\infty} \left(\frac{2\sigma^2}{l^2 \log n} n^{-\frac{\rho l^2}{2\sigma^2}} + \frac{\Delta_n^2 n^{2-\frac{\rho}{2\sigma^2}}}{\log n}\right) = \rho.$$

Combining the two sides, and letting $l \to 0$, we get the conclusion of Theorem 2. □

Diffusion scaling serves as a fundamental method for assessing the performance of stochastic systems, extensively utilized within the operations research domain. However, the diffusion limit behavior of bandit algorithms is still inadequately understood and remains predominantly unexplored. [8] analyzed the diffusion limit of Thompson sampling under Gaussian priors. Kuang and Wager[13] studied the diffusion limits for a class of sequentially randomized experiments. Their work is not applicable to the UCB algorithm. Kalvit and Zeeve [12] analysed the UCB algorithm in the conventional diffusion scaling in the cases of moderate and small gaps. Below, we present the diffusion limit under the large gap case.

We have the following theorem:

Theorem 3. *Consider UCB(ρ) as shown in Algorithm 1 on a two-armed σ-subgaussian bandit problem. Let the variances of the two reward distributions be σ_1^2 and σ_2^2. Under the assumptions of Theorem 1, as $n \to \infty$,*

(1) when $\alpha = 0.5$ and $h(n) = \omega(\log n)$ or $\alpha < 0.5$,

$$\frac{\tilde{R}_\rho(n, \lfloor nt \rfloor)}{\sqrt{n}} \Rightarrow \sigma_1 B(t);$$

(2) when $\alpha = 0.5$ *and* $\exists \theta_2 > 0$ *s.t.* $h(n) \sim \theta \log n$,

$$\frac{\tilde{R}_\rho(n)}{\sqrt{n}} \to \frac{\rho}{\theta} + \sigma_1 B(1), \qquad \textit{in distribution};$$

(3) when $\alpha = 0.5$ *and* $h(n) = o(\log n)$,

$$\frac{h(n)\tilde{R}_\rho(n)}{\sqrt{n}\log n} \to \rho, \qquad \textit{in probability.}$$

where the symbol " $\Rightarrow$ " denotes the weak convergence in Skorokhod Space $\mathcal{D}[0,1]$, $t \in [0,1]$ *and* $B(t)$ *is a standard Brownian motion in* $\mathbb{R}$.

Proof: Let $\mathcal{C}$ be the space of continuous functions $[0,1] \mapsto \mathbb{R}$, endowed with the uniform metric that defines the distance between two continuous functions $x(\cdot)$ and $y(\cdot)$ on $[0,1]$ as $\rho(x,y) := \sup_{t\in[0,1]} |x(t) - y(t)|$. Let $\mathcal{D}$ be the space of right-continuous functions with left limits, mapping $[0,1] \mapsto \mathbb{R}^2$, and endowed with the Skorohod metric (see [5], Chapters 2 and Chapters 3, for an overview). Let D_0 be the set of elements of D of the form (ϕ_1, ϕ_2) where ϕ_i is a non-decreasing real-valued function satisfying $0 \le \phi_i(t) \le 1$ for $i \in \{1,2\}$ and $t \in [0,1]$.

For $t \in [0,1]$, define

$$\psi_{1,n}(t) := \frac{\sum_{j=1}^{\lfloor nt \rfloor} X_{n,1,j} - \mu_n nt}{\sqrt{n}}, \ \psi_{2,n}(t) := \frac{\sum_{j=1}^{\lfloor a_n t \rfloor} X_{n,2,j} - \left(\mu_n - \frac{h(n)}{n^\alpha}\right) a_n t}{\sqrt{a_n}},$$

where $a_n := \frac{2\rho n^{2\alpha}(\log n)^2}{(h(n))^2}$ and $\mu_n = \mathbb{E}(X_{n,1,1})$. Define

$$W_1(t) := \sigma_1 B(t), \qquad W_2(t) := \sigma_2 W(t),$$

where $B(t)$ and $W(t)$ are independent standard Brownian motions in $\mathbb{R}$. Then $(\psi_{1,n}, \psi_{2,n}) \in \mathcal{D}$, and for any $i \in \{1,2\}$, $\mathbb{P}(W_i \in \mathcal{C}) = 1$. Since for fixed $i \in \{1,2\}$, $(X_{n,i,j})_{j\in\mathbb{N}}$ are independent and identically random variables, we know from the generalized Donsker's Theorem (See [5], Section 10, for details) that as $n \to \infty$,

$$(\psi_{1,n}, \psi_{2,n}) \Rightarrow (W_1, W_2) \text{ in } \mathcal{D}.$$

For $t \in [0,1]$, let $g_1(t) := t$, $g_2(t) := 0$,

$$\phi_{1,n}(t) := \frac{N_{1,n}(\lfloor nt \rfloor)}{n}, \quad \phi_{2,n}(t) := \begin{cases} \frac{N_{2,n}(\lfloor nt \rfloor)}{a_n}, & \frac{N_{2,n}(\lfloor nt \rfloor))}{a_n} \le 1, \\ 1, & \frac{N_{2,n}(n)}{a_n} > 1, \end{cases}$$

then $(\phi_{1,n}, \phi_{2,n}) \in \mathcal{D}_0$. Consequently, by Theorem 1, we obtain as $n \to \infty$,

$$(\phi_{1,n}, \phi_{2,n}) \Rightarrow (g_1, g_2) \text{ in } \mathcal{D}_0.$$

Thus, we have convergence in the product space (see [5], Theorem 3.9), i.e., as $n \to \infty$,

$$(\psi_{1,n}, \psi_{2,n}, \phi_{1,n}, \phi_{2,n}) \Rightarrow (W_1, W_2, g_1, g_2) \text{ in } \mathcal{D} \times \mathcal{D}_0.$$

For $i \in \{1, 2\}$ and $i \in \{1, 2\}$, define the composition $(\psi_{i,n} \circ \phi_{i,n})(t) := \psi_{i,n}(\phi_{i,n}(t))$, $(W_1 \circ g_1)(t) := W_1(g_1(t)) = W_1(t)$, $(W_2 \circ g_2)(t) := W_2(g_2(t)) = W_2(0)$. Since $W_1, W_2, g_1, g_2 \in \mathcal{C}$ w.p. 1, it follows from the continuous mapping theorem that as $n \to \infty$,

$$(\psi_{1,n} \circ \phi_{1,n}, \psi_{2,n} \circ \phi_{2,n}) \Rightarrow (W_1 \circ g_1, W_2 \circ g_2) \text{ in } \mathcal{D}.$$

Note that

$$\begin{aligned}
\tilde{R}_\rho(n, \lfloor nt \rfloor) &= \sum_{j=1}^{N_{1,n}(\lfloor nt \rfloor)} (\mu_n - X_{n,1,j}) + \sum_{j=1}^{N_{2,n}(\lfloor nt \rfloor)} (\mu_n - X_{n,2,j}) \\
&= \sum_{j=1}^{N_{1,n}(\lfloor nt \rfloor)} (\mu_n - X_{n,1,j}) + \sum_{j=1}^{N_{2,n}(\lfloor nt \rfloor)} (\mu_n - \Delta_n - X_{n,2,j}) + \Delta_n N_{2,n}(\lfloor nt \rfloor).
\end{aligned}$$

Therefore,

$$\begin{aligned}
\frac{\tilde{R}_\rho(n, \lfloor nt \rfloor)}{\sqrt{n}} &= -\frac{\sum_{j=1}^{N_{1,n}(\lfloor nt \rfloor)} (X_{n,1,j} - \mu_n)}{\sqrt{n}} - \frac{\sum_{j=1}^{N_{2,n}(\lfloor nt \rfloor)} (X_{n,2,j} - \mu_n + \Delta_n)}{\sqrt{n}} + \frac{\Delta_n N_{2,n}(\lfloor nt \rfloor)}{\sqrt{n}} \\
&= \frac{\Delta_n N_{2,n}(\lfloor nt \rfloor)}{\sqrt{n}} - \psi_{1,n} \circ \phi_{1,n}(t) - \sqrt{\frac{a_n}{n}} \psi_{2,n} \circ \phi_{2,n}(t).
\end{aligned}$$

If $\alpha = 0.5$ and $h(n) = \omega(\log n)$ or $\alpha < 0.5$, as $n \to \infty$, from Theorem 1 and the continuous mapping theorem, we have that

$$\frac{\tilde{R}_\rho(n, \lfloor nt \rfloor)}{\sqrt{n}} \Rightarrow W_1 \circ g_1 = W_1(t),$$

in $\mathcal{D}[0, 1]$.

By the same arguments, if $\alpha = 0.5$ and $\exists \theta_2 > 0$ s.t. $h(n) \sim \theta_2 \log n$, then as $n \to \infty$,

$$\begin{aligned}\frac{\tilde{R}_\rho(n)}{\sqrt{n}} = \frac{\tilde{R}_\rho(n,n)}{\sqrt{n}} &= -\frac{\sum_{j=1}^{N_{1,n}(n)} (X_{n,1,j} - \mu_n)}{\sqrt{n}} - \frac{\sum_{j=1}^{N_{2,n}(n)} (X_{n,2,j} - \mu_n + \Delta_n)}{\sqrt{n}} + \frac{\Delta_n N_{2,n}(n)}{\sqrt{n}} \\ &= -\psi_{1,n} \circ \phi_{1,n}(1) - \sqrt{\frac{a_n}{n}} \psi_{2,n} \circ \phi_{2,n}(1) + \frac{\Delta_n N_2(n)}{\sqrt{n}} \\ &\Rightarrow W_1(1) + \frac{\rho}{\theta_2};\end{aligned}$$

if $\alpha = 0.5$, $h(n) = o(\log n)$ and $h(n) = \omega\left(\sqrt{\log n}\right)$, then as $n \to \infty$,

$$\frac{\Delta_n \tilde{R}_\rho(n)}{\log n} = -\frac{h(n)}{\log n} \psi_{1,n} \circ \phi_{1,n}(1) - \frac{h(n)}{\log n} \sqrt{\frac{a_n}{n}} \psi_{2,n} \circ \phi_{2,n}(1) + \frac{h(n) \Delta_n N_2(n)}{\sqrt{n} \log n} \to \rho,$$

in probability. From above results, we can readily get the desired conclusions. □

4 Numerical Simulations

In existing literature, there is very little work of reporting numerical experiments on the performance of a sequence of MABs. In this section, we provide some concrete and comparable simulation results which can not only help us to assess and verify the performance of Theorem 1, exhibit the stochastic natures of MABs, but also inspire us to make other possible discovery. Our numerical experiments are aimed to evaluate the impact of the length of horizon, the parameter ρ in UCB(ρ) algorithms and the distributions of rewards. We also simulate the behaviour in the UCB1(ρ) algorithm in order to explore the difference between the two algorithms. We design serval groups of 2-armed bandit numerical experiments. In all experiments, Arm 1 is set to be the optimal arm with an expected value 1/2.

In the first group experiments, we let the reward distributions be Bernoulli, set the varied gaps $\Delta_n = 1/\ln(n)$, and let the horizon n varies in the set$\{10^3, 10^4, 10^5, 10^6\}$ i.e. when we simulate the n-horizon two-armed bandits, the reward distribution of suboptimal arm is $Ber(1/2 - 1/\ln(n))$. In addition we conduct the experiments under the UCB(ρ) algorithm with $\rho = 1/2$ and $\rho = 2$ respectively. As a comparison, we also conduct experiments under the UCB1(ρ). Every experiment is repeated 200 times. We observe the numbers of the suboptimal arm (namely, Arm 2) is selected and compute the ratios of this number over $\log n/\Delta_n^2$. The box-plot diagram of the ratios is presented as follows.

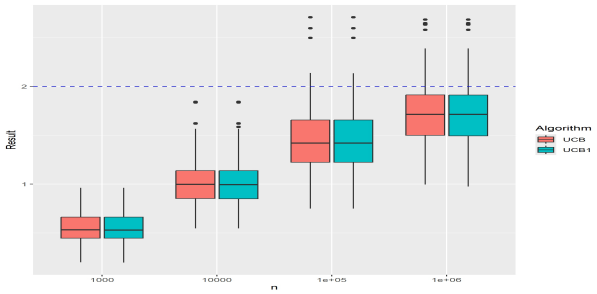

Figure 1: Box-Plot of Result= $N_2(n)\Delta_n^2/\ln(n)$ under Bernoulli distributions of rewards with $\Delta_n = 1/\ln(n)$ and $\rho = 1$ after 200 simulations .

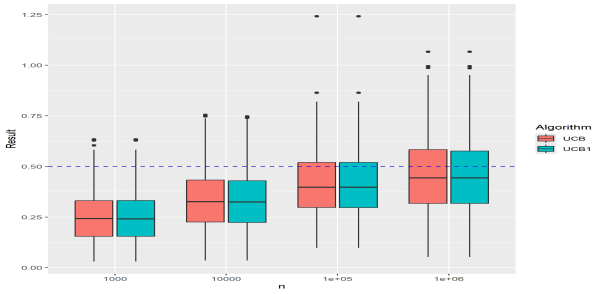

Figure 2: Box-Plot of Result= $N_2(n)\Delta_n^2/\ln(n)$ under Bernoulli distributions of rewards with $\Delta_n = 1/\ln(n)$ and $\rho = 4$ after 200 simulations .

In order to explore the possible differences arising from the distribution of rewards, we make the second group of experiments, where except that the distributions are replaced by the normal distributions with variance parameter $\sigma^2 = 1/4$, other setting are same as those in the first group. The box-plot diagram of the ratios is presented as follows.

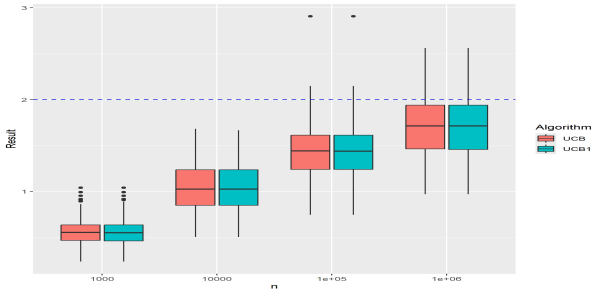

Figure 3: Box-Plot of Result= $N_2(n)\Delta_n^2/\ln(n)$ under Normal distributions of rewards with $\Delta_n = 1/\ln(n)$ and $\rho = 4$ after 200 simulations.

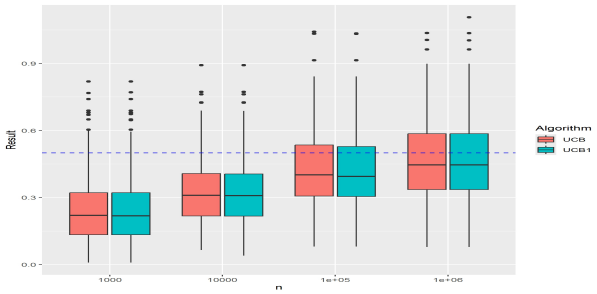

Figure 4: Box-Plot of Result= $N_2(n)\Delta_n^2/\ln(n)$ under Normal distributions of rewards with $\Delta_n = 1/\ln(n)$ and $\rho = 1$ after 200 simulations.

From the above figures we have the following observations.

(1) In all figures 1-4, the ratio $N_2(n)\Delta_n^2/\ln(n)$ exhibits a trend of getting closer to the parameter ρ. But when the horizon becomes increasingly longer, we cannot observe the phenomenon that the ratio $N_2(n)\Delta_n^2/\ln(n)$ becomes more and more concentrated. This shows that even if Theorem 1 can promise the concentration in theory (in these cases, the conditions of Theorem 1 hold), the timescales are too long. Cowan and Katehakis [7] also explicitly pointed out the similar observations.

(2) In Figure 2 and Figure 4, $\rho = 1/2$ and $\sigma^2 = 1/4$, which is out the work range of Theorem 1. We also observe more outliers in Figure 4 than that in Figures 3 even when the horizon n is large. These evidences show the difficulty when we try to prove Theorem 1 for small ρ.

(3) From the figures 1-4, we can readily see that the statistical behaviours of UCB1(ρ) is almost same as those of UCB(ρ). We guess all theoretical results in Section 3 shall be true for the algorithm UCB1(ρ), while the rigorous proof is still open for us.

As we remarked in Remark 3, Fan and Glynn [9] got the asymptotical result of $N_2(n)\Delta^2/\ln n$ in the case where a fixed two-armed bandit has rewards of normal distributions with variance $\sigma^2 = 1$ by taking the algorithm UCB1(2) ($\rho = 2\sigma^2$). While, in this paper, we only get the asymptotical result for a sequence of two-armed bandits in the setting $\rho > 4\sigma^2$. To explore the possible difficulties, we design two groups of experiments. In the first group of experiments we let the reward distributions be Bernoulli, set the varied gaps $\Delta_n \equiv 0.1$, and let the horizon n varies in the set $\{10^3, 10^4, 10^5, 10^6\}$. The algorithm for a fixed two-armed bandit is UCB1(1/2) while the one for a sequence of two-armed bandits is UCB(1/2). The setting of the second group is same except that the reward distributions is replaced by Normal with variance $\sigma^2 = 1/4$. In these experiments, for the given two-armed bandit, we simulate the increasing horizons by adding i.i.d rewards. Alternatively, to simulate a sequence of

two-armed bandits, for each different n, we independently produce an n-horizon two-armed bandit. For simplicity, we call the former as a fixed environment and the latter as a varied environment, respectively. Every experiment is repeated 100 times. The box-plot diagram of the ratios is presented as follows.

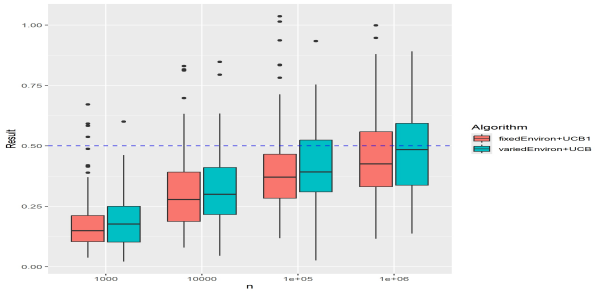

Figure 5: Box-Plot of Result= $N_2(n)\Delta_n^2/\ln(n)$ under Bernoulli distributions of rewards with $\Delta_n = 0.1$ and $\rho = 1/2$ after 100 simulations.

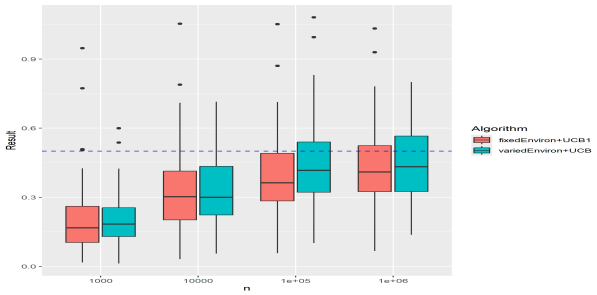

Figure 6: Box-Plot of Result= $N_2(n)\Delta_n^2/\ln(n)$ under normal distributions of rewards with $\Delta_n = 0.1$ and $\rho = 1/2$ after 100 simulations.

From Figure 5 and Figure 6, we observe that in our experimental setting which imitates the situation $\rho = 2\sigma^2$, the statistical performances of the ratios $N_2(n)\Delta_n^2/\ln(n)$ of the sequences of two-armed bandits is at least not worse than those of the single two-armed bandit. However, it is hard to assert that the asymptotical result of $N_2(n)\Delta_n^2/\ln n$ can be extended to the situation $\rho \geq 2\sigma^2$ when the rewards have σ-sub-Gaussian distributions, because our theoretical results are based on convergence of probability and/or almost surely convergence which is essentially not to be verified by the Box-Plot figures.

5 Conclusion and open problems

In this paper we study the sampling behaviours of a sequence of two-armed bandits $\{(2, \mathcal{P}_n, n)\}_n$ under the algorithm UCB(ρ). Assume the two-armed bandit $(2, \mathcal{P}_n, n)$ has σ-sub-Gaussion distributed rewards and the gap $\Delta_n = \omega\left(\sqrt{\frac{\ln n}{n}}\right)$, i.e. by the terminology in [12], $\{(2, \mathcal{P}_n, n)\}_n$ have large gaps. In theory, we proved that when $\rho > 4\sigma^2$ the number $N^*(n)$ of the optimal arm selected has the property that $N^*(n)/n$ converges almost surely to 1 and the selected numbers of the sub-optimal arm $N_*(n)$ has the property that $N_*(n)\Delta_n^2/\ln n$ converges in probability to the constant ρ. Based on these results, we get the asymptotical limit of regrets and regret processes as well. These theoretical results generalize and improve the existing ones in [12] and [9]. In numerical experiments, we simulate the behaviors of two-armed bandits under different setting including algorithms, sizes of gap, the parameter ρ, the horizon n and the distribution of rewards, which provide more evidences and intuitions to understand the bandit problems, especially, when the gaps are asymptotically degenerative.

Although in theory we have obtained the accurate asymptotical speed of arms' selected numbers, there are still many unsolved problems. For example, in this paper, we only get the results under the algorithm UCB(ρ), while the algorithm discussed in [12] and [9] is UCB1(ρ). We believe that the performances of UCB(ρ) and UCB1(ρ) are similar, which are also verified by our simulation. However, it is open for us by far to prove the same conclusions under the algorithm UCB1(ρ). In addition, in this paper, we only consider the relatively simple case of two-armed bandit problems. It is more interesting and complicated to explore the same problems for the bandits with more than 2 arms, which we will discuss in other place.

Appendix A: Auxiliary results

We will use the following inequality in our proofs, which can be readily get from the property of sub-Gaussian distribution.

Lemma 1 (**Chernoff-Hoeffding bound**). *Suppose that* $\{Y_{i,j} : i \in \{1, 2\}, j \in \mathbb{N}_+\}$ *is a collection of independent,* σ*-subgaussian random variables. Then, for any* $m_1, m_2 \in \mathbb{N}_+$ *and* $x > 0$,

$$\mathbb{P}\left(\frac{\sum_{j=1}^{m_1} Y_{1,j}}{m_1} - \frac{\sum_{j'=1}^{m_2} Y_{2,j'}}{m_2} \geq x\right) \leq \exp\left(\frac{-x^2 m_1 m_2}{2\sigma^2(m_1 + m_2)}\right).$$

Proof: By the properties of σ-subgaussian random variables (see [16], Lemma 5.4), random variable

$\frac{\sum_{j=1}^{m_1} Y_{1,j}}{m_1} - \frac{\sum_{j'=1}^{m_2} Y_{2,j'}}{m_2}$ is $\sqrt{\frac{1}{m_1}+\frac{1}{m_2}}\sigma$-subgaussian. Therefore,

$$\mathbb{P}\left(\frac{\sum_{j=1}^{m_1} Y_{1,j}}{m_1} - \frac{\sum_{j'=1}^{m_2} Y_{2,j'}}{m_2} \geq x\right) \leq \exp\left(\frac{-x^2}{2\sigma^2}\left(\frac{1}{m_1}+\frac{1}{m_2}\right)^{-1}\right),$$

which leads to the desired result. □

Before proving Theorem 1, we prove the following lemma.

Lemma 2. *Under the same conditions as Theorem 1, there exist $\epsilon_1(\rho,\sigma,\alpha) \in (0,\rho)$ (dependent only on ρ,σ,α), such that*

$$0 < \rho - \epsilon_1(\rho,\sigma,\alpha) < \liminf_{n\to\infty} \frac{N_{i'}(n)\Delta_n^2}{\ln n} < \limsup_{n\to\infty} \frac{N_{i'}(n)\Delta_n^2}{\ln n} \leq \frac{\rho}{\left(1-\frac{1}{2}\vee\sqrt{\frac{4\sigma^2}{\rho}}\right)^2} \quad a.s. \tag{2}$$

Proof: Without loss of generality, suppose that arm 1 is optimal, i.e., $\mu_1 > \mu_2$. We divide the proof into two parts.

(I) In this part, we show the left inequality of (2), i.e.

$$\liminf_{n\to\infty} \frac{N_2(n)\Delta_n^2}{\ln n} \geq \rho - \epsilon_1(\rho,\sigma,\alpha) > 0 \quad \text{a.s.}$$

Our argument is enlightened by the method used in Appendix D1 of [12].

For any $\epsilon \in (0,\rho)$, we define

$$v(t) := t - (\rho-\epsilon)\frac{t^{2\alpha}\ln t}{(h(t))^2}.$$

Observed that

$$v'(t) = 1 + \frac{\rho-\epsilon}{h^2(t)}\left[\frac{2h'(t)t^{2\alpha}\ln t}{h(t)} - \frac{2\alpha\ln t+1}{t^{1-2\alpha}}\right].$$

and that

$$[t^{-\alpha}h(t)]' = \frac{th'(t)-\alpha h(t)}{t^{\alpha+1}}.$$

Based on the assumptions of $h(t)$, for sufficiently large t we have that $v(t)$ is positive and increasing and that $t^{-\alpha}h(t)$ is decreasing. Therefore, for sufficiently large n, we have that $u(n) := \lceil v(n)\rceil > 0$

and for $t \in [v(n), n]$, $v(t)$ is increasing and $t^{-\alpha}h(t)$ is decreasing. Note that

$$\begin{aligned} N_1(n) \leq & u(n) + \sum_{t=u(n)}^{n-1} \mathbb{1}\left\{A_{n,t+1} = 1, N_{n,1}(t) \geq u(n)\right\} \\ \leq & u(n) + Z_1(n), \end{aligned} \tag{3}$$

where

$$Z_1(n) = \sum_{t=u(n)}^{n-1} \mathbb{1}\left\{\bar{Y}_{n,1}(t) - \bar{Y}_{n,2}(t) \geq \sqrt{\rho \ln n}\left(\frac{1}{\sqrt{N_{n,2}(t)}} - \frac{1}{\sqrt{N_{n,1}(t)}}\right) - \Delta_n, N_{n,1}(t) \geq u(t)\right\},$$

and

$$\bar{Y}_{n,i}(t) := \frac{\sum_{j=1}^{N_{n,i}(t)} Y_{n,i,j}}{N_{n,i}(t)}, \quad Y_{n,i,j} := X_{n,i,j} - \mu_{n,i},\ i \in \{1,2\},\ j \in \mathbb{N}_+.$$

It is easy to check that for sufficiently large n, $t - 1 \geq N_{n,1}(t) \geq u(t)$ and

$$\frac{1}{\sqrt{N_{n,2}(t)}} - \frac{1}{\sqrt{N_{n,1}(t)}} = \frac{1}{\sqrt{t - N_{n,1}(t)}} - \frac{1}{\sqrt{N_{n,1}(t)}} \geq 0.$$

Therefore

$$\begin{aligned} \mathbb{E}\left[Z_1(n)\right] \leq & \sum_{t=u(n)}^{n-1} \mathbb{P}\left\{\bar{Y}_{n,1}(t) - \bar{Y}_{n,2}(t) \geq \sqrt{\rho \ln t}\left(\frac{1}{\sqrt{N_{n,2}(t)}} - \frac{1}{\sqrt{N_{n,1}(t)}}\right) - \frac{h(t)}{t^\alpha}, N_{n,1}(t) \geq u(t)\right\} \\ = & \sum_{t=u(n)}^{n-1} \mathbb{P}\left\{\bar{Y}_{n,1}(t) - \bar{Y}_{n,2}(t) \geq r_1\left(N_{n,2}(t), N_{n,1}(t)\right)\sqrt{\rho \ln t}, N_{n,1}(t) \geq u(t)\right\} \\ = & \sum_{t=u(n)}^{n-1} \sum_{m=u(t)}^{t-1} \mathbb{P}\left\{\bar{Y}_{n,1}(t) - \bar{Y}_{n,2}(t) \geq r_1\left(N_{n,2}(t), N_{n,1}(t)\right)\sqrt{\rho \ln t}, N_{n,1}(t) = m\right\} \\ \leq & \sum_{t=u(n)}^{n-1} \sum_{m=u(t)}^{t-1} \mathbb{P}\left\{\frac{\sum_{j=1}^{m} Y_{n,1,j}}{m} - \frac{\sum_{j'=1}^{t-m} Y_{n,2,j'}}{t-m} \geq r_1\left(t-m, m\right)\sqrt{\rho \ln t}\right\}, \end{aligned} \tag{4}$$

where

$$r_1(x, y) := \frac{1}{\sqrt{x}} - \frac{1}{\sqrt{y}} - \frac{h(t)}{t^\alpha \sqrt{\rho \ln t}}.$$

Since $m \in [u(t), t-1] \subset [v(t), t-1]$, for t large enough,

$$r_1\left(t - m, m\right) \geq r_1\left(t - v(t), v(t)\right) = \frac{r_2(t, \epsilon)h(t)}{t^\alpha \sqrt{\ln t}} > 0. \tag{5}$$

where

$$r_2(t,\epsilon) := \frac{1}{\sqrt{\rho-\epsilon}} - \frac{1}{\sqrt{t^{1-2\alpha}(h(t))^2(\ln t)^{-1}-\rho+\epsilon}} - \frac{1}{\sqrt{\rho}}.$$

We can therefore apply the Chernoff-Hoeffding bound (Lemma 1) to (4) to conclude that

$$\begin{aligned}\mathbb{E}\left[Z_1(n)\right] &\le \sum_{t=u(n)}^{n-1}\sum_{m=u(t)}^{t} \exp\left[\left(\frac{1}{\sqrt{t-m}} - \frac{1}{\sqrt{m}} - \frac{h(t)}{t^\alpha\sqrt{\rho\ln t}}\right)^2 \frac{-\rho m(t-m)\ln t}{2\sigma^2 t}\right] \\ &= \sum_{t=u(n)}^{n-1}\sum_{m=u(t)}^{t} \exp\left[\frac{-\rho\ln t}{2\sigma^2 t}\left(f_1(t,m)\right)^2\right], \end{aligned} \tag{6}$$

where

$$f_1(t,x) := r_1(t-x,x)\sqrt{x(t-x)} = \sqrt{x} - \sqrt{t-x} - \frac{h(t)\sqrt{x(t-x)}}{t^\alpha\sqrt{\rho\ln t}}.$$

Notice that for t large enough, $f_1(t,x)$ is non-decreasing on $x \in [v(t), t-1]$. Thus, we have in (6) that

$$f_1^2(t,m) \ge f_1^2(t,v(t)) = v(t)\big(t-v(t)\big)r_1\left(t-v(t),v(t)\right) = \left(t - \frac{t^{2\alpha}(\rho-\epsilon)\ln t}{(h(t))^2}\right)(\rho-\epsilon)\left(r_2(t,\epsilon)\right)^2.$$

Using these facts in (6), we conclude

$$\begin{aligned}\mathbb{E}\left[Z_1(n)\right] &\le \sum_{t=u(n)}^{n-1}\sum_{m=u(t)}^{t} \exp\left[\frac{-\rho\ln t}{2\sigma^2}\left(1 - \frac{(\rho-\epsilon)\ln t}{t^{1-2\alpha}(h(t))^2}\right)(\rho-\epsilon)\left(r_2(t,\epsilon)\right)^2\right] \\ &\le \sum_{t=u(n)}^{n-1} \frac{\rho\ln t}{(h(t))^2} \exp\left\{\left[2\alpha - \frac{\rho}{2\sigma^2}\left(1 - \frac{(\rho-\epsilon)\ln t}{t^{1-2\alpha}(h(t))^2}\right)(\rho-\epsilon)\left(r_2(t,\epsilon)\right)^2\right]\ln t\right\}. \end{aligned} \tag{7}$$

For any $\delta > 0$, by Markov's inequality, we then have

$$\mathbb{P}\left(N_1(n) - u(n) \ge \frac{\delta n^{2\alpha}\ln n}{(h(n))^2}\right) \le \mathbb{P}\left(Z_1(n) \ge \frac{\delta n^{2\alpha}\ln n}{(h(n))^2}\right) \le \frac{\mathbb{E}\left(Z_1(n)\right)(h(n))^2}{\delta n^{2\alpha}\ln n}.$$

Consequently, by (7),

$$\begin{aligned}&\mathbb{P}\left(N_1(n) \ge u(n) + \frac{\delta n^{2\alpha}\ln n}{(h(n))^2}\right) \\ \le &\frac{(h(n))^2}{\delta n^{2\alpha}\ln n}\sum_{t=u(n)}^{n-1} \frac{\rho\ln t}{(h(t))^2} \exp\left\{\left[2\alpha - \frac{\rho}{2\sigma^2}\left(1 - \frac{(\rho-\epsilon)\ln t}{t^{1-2\alpha}(h(t))^2}\right)(\rho-\epsilon)\left(r_2(t,\epsilon)\right)^2\right]\ln t\right\}.\end{aligned}$$

Note that $\lim_{t\to\infty} \frac{\ln t}{t^{1-2\alpha}(h(t))^2} = 0$. For n large enough, there exist a constant C_1 such that

$$\mathbb{P}\left(N_1(n) \geq u(n) + \frac{\delta n^{2\alpha}\ln n}{(h(n))^2}\right) \leq \frac{C_1 \ln n}{\delta(h(u(n)))^2} \exp\left\{\left[2\alpha - \frac{\rho}{2\sigma^2}\left(1-\sqrt{\frac{\rho-\epsilon}{\rho}}\right)^2\right]\ln n\right\}. \quad (8)$$

Let $g(\epsilon) := 2\alpha - \frac{\rho}{2\sigma^2}\left(1-\sqrt{\frac{\rho-\epsilon}{\rho}}\right)^2$. Note that $\rho > 4\sigma^2 \geq 2\sigma^2(2\alpha+1)$. Since $g(\rho) < -1$ and $g(0) = 2\alpha \geq 0$. Therefore, there exist $\epsilon_1(\rho,\sigma,\alpha) \in (0,\rho)$ such that $\forall \epsilon \geq \epsilon_1(\rho,\sigma,\alpha)$, $g(\epsilon) < -1$. Finally since $\delta > 0$ is arbitrary, we conclude from (8) using the Borel-Cantelli Lemma that

$$\begin{aligned}
&-\liminf_{n\to\infty}\left(\frac{N_2(n)(h(n))^2}{n^{2\alpha}\ln n} - \rho + \epsilon_1(\rho,\sigma,\alpha)\right) \\
&= \limsup_{n\to\infty} \rho - \epsilon_1(\rho,\sigma,\alpha) - \frac{N_2(n)(h(n))^2}{n^{2\alpha}\ln n} \\
&= \limsup_{n\to\infty} \frac{N_1(n) - \left[n - (\rho-\epsilon_1(\rho,\sigma,\alpha))\, n^{2\alpha}(h(n))^{-2}\ln n\right]}{n^{2\alpha}(h(n))^{-2}\ln n} \leq 0 \quad \text{a.s.},
\end{aligned} \quad (9)$$

i.e.,

$$\liminf_{n\to\infty} \frac{N_2(n)\Delta^2}{\ln n} \geq \rho - \epsilon_1(\rho,\sigma,\alpha) > 0 \quad \text{a.s.} \quad (10)$$

(II) In this part, we show the right inequality of (2). It is sufficient to prove that

$$\limsup_{n\to\infty} \frac{N_2(n)\Delta_n^2}{\ln n} \leq \frac{\rho}{(1-2l)^2}, \qquad \text{a.s.}$$

for any $l \in \left(1/4 \vee \sqrt{\frac{\sigma^2}{\rho}}, 1/2\right)$.

Let $s(n) := \left\lceil \frac{4\rho\ln n}{\Delta_n^2} \right\rceil$. Since $\rho > 4\sigma^2$, for any $l \in \left(1/4 \vee \sqrt{\frac{\sigma^2}{\rho}}, 1/2\right)$ satisfying that $l^2\rho/\sigma^2 > 1$ and $\frac{1}{(1-2l)^2} > 4$. Let $\delta = l\Delta_n$. For the arm 2, we have

$$N_2(n) \leq s(n) + \sum_{t=s(n)}^{n-1} \mathbb{1}\left\{A_{n,t+1} = 2, N_{n,2}(t) \geq s(n)\right\} = s(n) + G_{n,1} + G_{n,2} + G_{n,3}, \quad (11)$$

where

$$\begin{aligned}
G_{n,1} &= \sum_{t=s(n)}^{n-1} \mathbb{1}\left\{A_{n,t+1} = 2, \bar{\mu}_{n,2}(t) + \sqrt{\frac{\rho\ln n}{N_{n,2}(t)}} \geq \mu_{n,1} - \delta, \bar{\mu}_{n,2}(t) \leq \mu_{n,2} + \delta, N_{n,2}(t) \geq s(n)\right\}, \\
G_{n,2} &= \sum_{t=s(n)}^{n-1} \mathbb{1}\left\{A_{n,t+1} = 2, \bar{\mu}_{n,2}(t) + \sqrt{\frac{\rho\ln n}{N_{n,2}(t)}} \geq \mu_{n,1} - \delta, \bar{\mu}_{n,2}(t) > \mu_{n,2} + \delta, N_{n,2}(t) \geq s(n)\right\}, \\
G_{n,3} &= \sum_{t=s(n)}^{n-1} \mathbb{1}\left\{A_{n,t+1} = 2, \bar{\mu}_{n,2}(t) + \sqrt{\frac{\rho\ln n}{N_{n,2}(t)}} < \mu_{n,1} - \delta, N_{n,2}(t) \geq s(n)\right\}.
\end{aligned}$$

$G_{n,1}$ is upper bounded via

$$\begin{aligned} G_{n,1} &\leq \sum_{t=s(n)}^{n-1} \mathbb{1}\left\{A_{n,t+1}=2, (\Delta_n - 2\delta)^2 \leq \frac{\rho \ln n}{N_{n,2}(t)}, N_{n,2}(t) \geq s(n)\right\} \\ &\leq \sum_{t=s(n)}^{n-1} \mathbb{1}\left\{A_{n,t+1}=2, s(n) \leq N_{n,2}(t) \leq \frac{\rho \ln n}{(\Delta_n - 2\delta)^2}\right\} \\ &\leq \frac{\rho \ln n}{(\Delta_n - 2\delta)^2} - s(n) + 1. \end{aligned} \tag{12}$$

For G_2, we have that

$$\begin{aligned} G_{n,2} &\leq \sum_{t=s(n)}^{n-1} \mathbb{1}\left\{A_{n,t+1}=2, \bar{\mu}_{n,2}(t) > \mu_{n,2} + \delta, N_{n,2}(t) \geq s(n)\right\} \\ &\leq \sum_{m=s(n)}^{N_2(n)} \mathbb{1}\left(\bar{\mu}_{n,2}(m) > \mu_{n,2} + \delta\right). \end{aligned} \tag{13}$$

Therefore,

$$\begin{aligned} \mathbb{P}(G_{n,2} > \varepsilon s(n)) &\leq \frac{1}{\varepsilon s(n)} \sum_{m=s(n)}^{n} \mathbb{P}\left(\bar{\mu}_{n,2}(m) > \mu_{n,2} + \delta\right) \leq \frac{1}{\varepsilon s(n)} \sum_{m=s(n)}^{n} \exp\left\{\frac{-m\delta^2}{2\sigma^2}\right\} \\ &\leq \frac{1}{\varepsilon s(n)} \frac{1}{1 - e^{-\frac{\delta^2}{2\sigma^2}}} \exp\left\{\frac{-s(n)\delta^2}{2\sigma^2}\right\} = \frac{1}{\varepsilon s(n)} \frac{2\sigma^2}{\delta^2} \exp\left\{\frac{-s(n)\delta^2}{2\sigma^2}\right\}, \end{aligned}$$

and hence, using the fact $s(n)\delta^2/\sigma^2 > \frac{2\rho l^2}{\sigma^2} \ln n$, we have that

$$\mathbb{P}(G_{n,2} > \varepsilon s(n)) \leq \frac{\sigma^2}{2\rho l^2 \varepsilon \ln n} n^{-\frac{2\rho l^2}{\sigma^2}}.$$

Since $\frac{\rho l^2}{\sigma^2} > 1$, by Borel-Cantelli Lemma and the $G_{n,2}/s(n) \leq \varepsilon$ almost surely.

For $G_{n,3}$, we have that

$$\begin{aligned} G_{n,3} &= \sum_{t=s(n)}^{n-1} \mathbb{1}\left\{A_{n,t+1}=2, \bar{\mu}_{n,1}(t) + \sqrt{\frac{\rho \ln n}{N_{n,1}(t)}} \leq \bar{\mu}_{n,2}(t) + \sqrt{\frac{\rho \ln n}{N_{n,2}(t)}} < \mu_{n,1} - \delta, N_{n,2}(t) \geq s(n)\right\} \\ &\leq \sum_{t=s(n)}^{n-1} \mathbb{1}\left\{A_{n,t+1}=2, \bar{\mu}_{n,1}(t) < \mu_{n,1} - \delta - \sqrt{\frac{\rho \ln n}{N_{n,1}(t)}}, N_{n,2}(t) \geq s(n)\right\} \\ &\leq \sum_{k=s(n)}^{N_2(n)} \sum_{t=k+1}^{n} \mathbb{1}\left\{\bar{\mu}_{n,1}(t-k) < \mu_{n,1} - \delta - \sqrt{\frac{\rho \ln n}{t-k}}\right\}. \end{aligned} \tag{14}$$

Therefore, for any $\varepsilon > 0$, thanks to $\rho > 4\sigma^2$, we have that

$$\begin{aligned}
\mathbb{P}(G_{n,3} > \varepsilon s(n)) &\le \frac{1}{\varepsilon s(n)} \mathbb{E}(G_{n,3}) \\
&\le \frac{1}{\varepsilon s(n)} \sum_{k=s(n)}^{n} \sum_{t=k+1}^{n} \mathbb{P}\left(\bar{\mu}_{n,1}(t-k) < \mu_{n,1} - \delta - \sqrt{\frac{\rho \ln n}{t-k}} \right) \\
&\le \frac{1}{\varepsilon s(n)} \sum_{k=s(n)}^{n} \sum_{t=k+1}^{k+s(n)} \mathbb{P}\left\{ \bar{\mu}_{n,1}(t-k) < \mu_{n,1} - \sqrt{\frac{\rho \ln n}{t-k}} \right\} \\
&\qquad + \frac{1}{\varepsilon s(n)} \sum_{k=s(n)}^{n} \sum_{t=k+s(n)}^{n} \mathbb{P}\left\{ \bar{\mu}_{n,1}(t-k) < \mu_{n,1} - \delta \right\} \\
&\le \frac{1}{\varepsilon s(n)} n \Big[s(n) \exp\{-\frac{\rho \ln n}{2\sigma^2}\} + \frac{1}{1 - e^{-\delta^2/2\sigma^2}} \exp\{-\frac{s(n)\delta^2}{2\sigma^2}\} \Big] \\
&\le \frac{1}{\varepsilon} \Big[n^{1 - \frac{\rho}{2\sigma^2}} + \frac{2\sigma^2 n}{s(n)\delta^2} \exp\{-\frac{s(n)\delta^2}{2\sigma^2}\} \Big].
\end{aligned}$$

The same arguments as those used to $G_{n,2}$ implies that $G_{n,3} \le \varepsilon s(n)$ almost surely.

Summing up, we get that

$$N_2(n) \le \frac{\rho \ln n}{(\Delta_n - 2\delta)^2} + 2\varepsilon s(n)$$

almost surely. Then letting $\varepsilon \to 0$ yields that

$$\frac{N_2(n)\Delta_n^2}{\ln n} \le \frac{\rho}{(1-2l)^2} \qquad a.s.$$

which completes the proof of the second step. □

By the same argument with some necessary modifications, we can readily get the following result.

Corollary 2. *For any constant $\gamma \in (0,1]$, there exist constants $0 < C_1 < +\infty$,*

$$C_1 < \liminf_{n\to\infty} \frac{N_{n,i'}(\gamma n)(h(\gamma n))^2}{(\gamma n)^{2\alpha} \ln n}. \quad a.s. \tag{15}$$

Proof: Note that following the same argument in Part 1, we have that

$$N_{n,1}(\gamma n) \le u(\gamma n) + Z_1(n),$$

where

$$\mathbb{E}\left[Z_1(n)\right] \le \sum_{t=u(\gamma n)}^{\gamma n - 1} \sum_{m=u(t)}^{t-1} \mathbb{P}\left\{ \frac{\sum_{j=1}^{m} Y_{n,1,j}}{m} - \frac{\sum_{j'=1}^{t-m} Y_{n,2,j'}}{t-m} \ge r_1\left(t-m, m\right) \sqrt{\rho \ln t} \right\},$$

and that

$$\mathbb{P}\left(N_{n,1}(\gamma n) \geq u(\gamma n) + \frac{\delta(\gamma n)^{2\alpha}\ln(\gamma n)}{(h(\gamma n))^2}\right)$$
$$\leq \frac{C_1 \ln(\gamma n)}{\delta(h(u(\gamma n)))^2} \exp\left\{\left[2\alpha - \frac{\rho}{2\sigma^2}\left(1-\sqrt{\frac{\rho-\epsilon}{\rho}}\right)^2\right]\ln n\right\}, \tag{16}$$

which implies the left inequality of (15) by Borel-Cantelli Lemma. □

Appendix B: Proof of Theorem 1

Proof: From the conclusion of Lemma 2 in Appendix A and the equation $n = N_1(n) + N_2(n)$, we can derive the first part of the theorem's conclusion, i.e., as $n \to \infty$,

$$\frac{N_{i^*}(n)}{n} = \frac{N_1(n)}{n} = 1 - \frac{N_2(n)}{n} \to 1 \quad \text{a.s.} \tag{17}$$

Next, we will prove the second part of the conclusion. We divide the proof into two steps.

Step 1: Proving that for any $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{N_2(n)\Delta_n^2}{\ln n} \leq \rho - \varepsilon\right) = 0. \tag{18}$$

By (4), we have that for any given $\epsilon \in (0, \varepsilon/2)$

$$\mathbb{E}\left[Z_1(n)\right] \leq \sum_{t=u(n)}^{n-1} \mathbb{P}\left\{\bar{Y}_{n,1}(t) - \bar{Y}_{n,2}(t) \geq r_1\left(N_{n,2}(t), N_{n,1}(t)\right)\sqrt{\rho \ln t}, N_{n,1}(t) \geq u(t)\right\}$$
$$\leq \sum_{t=u(n)}^{n-1} \mathbb{P}\left\{\bar{Y}_{n,1}(t) - \bar{Y}_{n,2}(t) \geq \frac{\sqrt{\rho}h(t)r_2(t,\epsilon)}{t^\alpha}, N_{n,1}(t) \geq u(t)\right\}$$
$$= \sum_{t=u(n)}^{n-1} \mathbb{P}\left(W_{n,1}(t) \geq r_2(t,\epsilon), N_{n,1}(t) \geq u(t)\right),$$

where

$$W_{n,1}(t) := \frac{t^\alpha}{\sqrt{\rho}h(t)}\left(\bar{Y}_{n,1}(t) - \bar{Y}_{n,2}(t)\right).$$

We already know that when t is large enough, $r_2(t,\epsilon) > 0$. For any $i \in \{1,2\}$, let

$$Q_{n,1}(t) = \left|\frac{\sum_{j=1}^{N_{n,1}(\lfloor nt \rfloor)} Y_{n,1,j}}{\sqrt{n}}\right|, Q_{n,2}(t) = \left|\frac{\sum_{j=1}^{N_{n,2}(\lfloor nt \rfloor)} Y_{n,2,j}}{\sqrt{n^{2\alpha}(\ln n)^{1+c}/h^2(n)}}\right|$$

where $c \in (0,1)$ is a constant. We have that

$$\begin{aligned}
\max_{u(n)\le t\le n-1} |W_{n,1}(t)| &\le \max_{u(n)\le t\le n-1} \frac{t^\alpha}{\sqrt{\rho}h(t)} \left(\left| \frac{\sum_{j=1}^{N_{n,1}(t)} Y_{n,1,j}}{N_{n,1}(t)} \right| + \left| \frac{\sum_{j'=1}^{N_{n,2}(t)} Y_{n,2,j'}}{N_{n,2}(t)} \right| \right) \\
&\le \max_{u(n)\le t\le n-1} \left\{ \frac{\sqrt{2}n^\alpha}{\sqrt{\rho}h(n)} \left(\frac{\sqrt{n}}{N_{n,1}(t)} Q_{n,1}(t) + \frac{\sqrt{n^{2\alpha}(\ln n)^{1+c}/h^2(n)}}{N_{n,2}(t)} Q_{n,2}(t) \right) \right\} \\
&\le \max_{u(n)\le t\le n-1} \frac{\sqrt{2}n^{\alpha+1/2}}{\sqrt{\rho}h(n)N_{n,1}(t)} \times \max_{u(n)\le t\le n-1} Q_{n,1}(t) \\
&\quad + \max_{u(n)\le t\le n-1} \frac{n^{2\alpha}\sqrt{2(\ln n)^{1+c}}}{\sqrt{\rho}h^2(n)N_{n,2}(t)} \times \max_{u(n)\le t\le n-1} Q_{n,2}(t). \qquad (19)
\end{aligned}$$

By using Corollary 2, it is easy to see that as $n \to \infty$, almost surely,

$$\max_{u(n)\le t\le n-1} \frac{\sqrt{2}n^{\alpha+1/2}}{\sqrt{\rho}h(n)N_{n,1}(t)} \to 0, \quad \max_{u(n)\le t\le n-1} \frac{n^{2\alpha}\sqrt{2(\ln n)^{1+c}}}{\sqrt{\rho}h^2(n)N_{n,2}(t)} \to 0.$$

In addition, since $Y_{n,i,j} := X_{n,i,j} - \mu_{n,i}$, the collection $\{Y_{n,i,j} : j \in \mathbb{N}\}$ is a series of independent and identically distributed random variables, characterized by $\mathbb{E}(Y_{n,i,1}) = 0$ and $\mathrm{Var}(Y_{n,i,1}) = \mathrm{Var}(X_{i,1}) \le \sigma^2$ (the properties of σ-subgaussian random variables). By the generalized Donsker's Theorem ([5, P.77]),

$$\frac{\sum_{j=1}^{\lceil nt \rceil} Y_{n,1,j}}{\sqrt{n}} \quad \text{and} \quad \frac{\sum_{j=1}^{\lceil n^{2\alpha}(\ln n)^{1+c}t/h^2(n) \rceil} Y_{n,2,j}}{\sqrt{n^{2\alpha}(\ln n)^{1+c}/h^2(n)}}$$

converge weakly in Skorohod Space to a Brownian motion $\{W(t); t \ge 0\}$, respectively. Using Corollary 2 again, we have that almost surely,

$$\frac{N_{n,1}(t)}{n} \to 1 \quad \text{and} \quad \frac{N_{n,2}(t)}{n^{2\alpha}(\ln n)^{1+c}/h^2(n)} \to 0.$$

Therefore, by the continuous mapping theorem,

$$\max_{u(n)\le t\le n-1} Q_{n,1}(t) \quad \text{and} \quad \max_{u(n)\le t\le n-1} Q_{n,2}(t)$$

converge in distribution to $W(1)$ and $W(0)$, respectively. Consequently, the Slutsky's Lemma plus (19) implies that in probability,

$$\max_{u(n)\le t\le n-1} |W_{n,1}(t)| \to 0, \qquad (20)$$

as $n \to \infty$. Taking $\delta = \varepsilon/2 - \epsilon > 0$, we have that as $n \to \infty$,

$$\begin{aligned}
&\mathbb{P}\left(N_1(n) - u(n) \geq \frac{\delta n^{2\alpha} \ln n}{(h(n))^2}\right) \leq \mathbb{P}\left(Z_1(n) \geq \frac{\delta n^{2\alpha} \ln n}{(h(n))^2}\right) \\
&\leq \frac{\mathbb{E}\left(Z_1(n)\right)(h(n))^2}{\delta n^{2\alpha} \ln n} \leq \frac{(h(n))^2}{\delta n^{2\alpha} \ln n} \sum_{t=u(n)}^{n-1} \mathbb{P}(W_{n,1}(t) \geq r_2(t,\epsilon)) \\
&\leq \frac{\rho+\epsilon}{\delta} \mathbb{P}\left(\max_{u(n)\leq t\leq n-1} W_1(t) \geq \min_{u(n)\leq t\leq n-1} r_2(t,\epsilon)\right) \to 0,
\end{aligned}$$

where the last limit follows from (20) and the fact

$$\lim_{n\to\infty} \min_{u(n)\leq t\leq n-1} r_2(t,\epsilon) > 0.$$

As a result,

$$\begin{aligned}
&\limsup_{n\to\infty} \mathbb{P}\left(\frac{N_1(n) - n + \rho n^{2\alpha}(h(n))^{-2}\ln n}{n^{2\alpha}(h(n))^{-2}\ln n} \geq 2(\epsilon+\delta)\right) \\
&\leq \limsup_{n\to\infty} \mathbb{P}\left(\frac{N_1(n) - n + \rho n^{2\alpha}(h(n))^{-2}\ln n}{n^{2\alpha}(h(n))^{-2}\ln n} \geq \epsilon + \delta + \frac{(h(n))^2}{n^{2\alpha}\ln n}\right) \\
&= \limsup_{n\to\infty} \mathbb{P}\left(N_1(n) - \left(n - (\rho-\epsilon)n^{2\alpha}(h(n))^{-2}\ln n\right) \geq \delta n^{2\alpha}(h(n))^{-2}\ln n + 1\right) \\
&\leq \limsup_{n\to\infty} \mathbb{P}\left(N_1(n) - u(n) \geq \delta n^{2\alpha}(h(n))^{-2}\ln n\right) = 0.
\end{aligned}$$

Consequently,

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{N_1(n) - n + \rho n^{2\alpha}(h(n))^{-2}\ln n}{n^{2\alpha}(h(n))^{-2}\ln n} \geq \varepsilon\right) = 0,$$

from which we can readily get (21) since $N_2(n) = n - N_1(n)$.

Step 2 Proving that for any $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{N_2(n)\Delta_n^2}{\ln n} \geq \rho + \varepsilon\right) = 0. \tag{21}$$

This part is similar to and simpler than Part 2 in the proof of Lemma 2.

Let $s_1(n) := \left\lceil \frac{\rho \ln n}{\Delta_n^2} \right\rceil = \left\lceil \frac{\rho n^{2\alpha} \ln n}{(h(n))^2} \right\rceil$, $l \in (0, 0.5)$ and $\delta = l\frac{h(n)}{n^\alpha} = l\Delta_n$. For the arm 2, we have

$$N_2(n) \leq s_1(n) + \sum_{t=s(n)}^{n-1} \mathbb{1}\left\{A_{n,t+1} = 2, N_{n,2}(t) \geq s(n)\right\} = s_1(n) + G_{n,1} + G_{n,2} + G_{n,3}, \tag{22}$$

where

$$G_{n,1} = \sum_{t=s_1(n)}^{n-1} \mathbb{1}\left\{A_{n,t+1}=2, \bar{\mu}_{n,2}(t)+\sqrt{\frac{\rho \ln n}{N_{n,2}(t)}} \geq \mu_{n,1}-\delta, \bar{\mu}_{n,2}(t) \leq \mu_{n,2}+\delta, N_{n,2}(t) \geq s_1(n)\right\},$$

$$G_{n,2} = \sum_{t=s_1(n)}^{n-1} \mathbb{1}\left\{A_{n,t+1}=2, \bar{\mu}_{n,2}(t)+\sqrt{\frac{\rho \ln n}{N_{n,2}(t)}} \geq \mu_{n,1}-\delta, \bar{\mu}_{n,2}(t) > \mu_{n,2}+\delta, N_{n,2}(t) \geq s_1(n)\right\},$$

$$G_{n,3} = \sum_{t=s_1(n)}^{n-1} \mathbb{1}\left\{A_{n,t+1}=2, \bar{\mu}_{n,2}(t)+\sqrt{\frac{\rho \ln n}{N_{n,2}(t)}} < \mu_{n,1}-\delta, N_{n,2}(t) \geq s_1(n)\right\}.$$

Obviously, $G_{n,1}$ is upper bounded via

$$G_{n,1} \leq \frac{\rho \ln n}{(\Delta_n - 2\delta)^2} - s_1(n) + 1.$$

For $G_{n,2}$, since

$$G_{n,2} \leq \sum_{m=s_1(n)}^{N_2(n)} \mathbb{1}\left(\bar{\mu}_{n,2}(m) > \mu_{n,2}+\delta\right).$$

due to Lemma 2, for any $\varepsilon > 0$ we have that

$$\begin{aligned}
\limsup_{n\to\infty} \mathbb{P}(G_{n,2} > \varepsilon s_1(n)) &\leq \limsup_{n\to\infty} \mathbb{P}\left(\sum_{m=s_1(n)}^{g(\rho)s_1(n)} \mathbb{1}\left(\bar{\mu}_{n,2}(m) > \mu_{n,2}+\delta\right) > \varepsilon s_1(n)\right) \\
&\leq \limsup_{n\to\infty} \frac{1}{\varepsilon s_1(n)} \sum_{m=s_1(n)}^{g(\rho)s_1(n)} \mathbb{P}\left(\bar{\mu}_{n,2}(m) > \mu_{n,2}+\delta\right) \\
&\leq \limsup_{n\to\infty} \frac{g(\rho)-1}{\varepsilon} \exp\left\{\frac{-s_1(n)\delta^2}{2\sigma^2}\right\},
\end{aligned}$$

where $g(\rho) = \left(1 - \frac{1}{2} \vee \sqrt{\frac{4\sigma^2}{\rho}}\right)^{-2}$, and hence

$$\lim_{n\to\infty} \mathbb{P}(G_{n,2} > \varepsilon s_1(n)) \leq \lim_{n\to\infty} \frac{g(\rho)-1}{\varepsilon} n^{-\frac{\rho l^2}{2\sigma^2}} = 0.$$

Therefore $G_{n,2}/s_1(n) \to 0$ in probability.

By the same discussions, since

$$G_{n,3} \leq \sum_{k=s_1(n)}^{N_2(n)} \sum_{t=k+1}^{n} \mathbb{1}\left\{\bar{\mu}_{n,1}(t-k) < \mu_{n,1} - \delta - \sqrt{\frac{\rho \ln t}{t-k}}\right\}.$$

for any $\varepsilon > 0$, we have that

$$\begin{aligned}\limsup_{n\to\infty} \mathbb{P}(G_{n,3} > \varepsilon s(n)) &\leq \limsup_{n\to\infty} \mathbb{P}\left(\sum_{k=s_1(n)}^{g(\rho)s(n)} \sum_{t=k+1}^{n} \mathbb{1}\left\{\bar{\mu}_{n,1}(t-k) < \mu_{n,1} - \delta - \sqrt{\frac{\rho \ln t}{t-k}}\right\} > \varepsilon s(n)\right) \\ &\leq \limsup_{n\to\infty} \frac{1}{\varepsilon s_1(n)} \sum_{k=s_1(n)}^{g(\rho)s_1(n)} \sum_{t=k+1}^{n} \mathbb{P}\left(\bar{\mu}_{n,1}(t-k) < \mu_{n,1} - \delta - \sqrt{\frac{\rho \ln n}{t-k}}\right) \\ &\leq \limsup_{n\to\infty} \frac{g(\rho)-1}{\varepsilon} n \exp\left\{\frac{-\rho \ln n}{2\sigma^2}\right\} = 0,\end{aligned}$$

which implies that $G_{n,3}/s_1(n) \to 0$ in probability as $n \to \infty$.

Summing up, we obtain that for any $l \in (0,1)$,

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{N_2(n)}{s_1(n)} < \frac{1}{(1-l)}\right) = 1.$$

Letting l to approach zero, we get the desired results (21). □

References

[1] Shipra Agrawal and Navin Goyal. "Analysis of Thompson sampling for the multi-armed bandit problem". In: *Proceedings of the 25th Annual Conference on Learning Theory*. Vol. 23. PMLR, 2012, pp. 39.1–39.26.

[2] Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. "Exploration–exploitation tradeoff using variance estimates in multi-armed bandits". In: *Theoretical Computer Science* 410.19 (2009), pp. 1876–1902.

[3] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. "Finite-time analysis of the multi-armed bandit problem". In: *Machine Learning* 47 (2002), pp. 235–256.

[4] Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. "Minimax Regret Bounds for Reinforcement Learning". In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by Doina Precup and Yee Whye Teh. Vol. 70. Proceedings of Machine Learning Research. PMLR, 2017, pp. 263–272. URL: https://proceedings.mlr.press/v70/azar17a.html.

[5] Patrick Billingsley. *Convergence of probability measures*. John Wiley & Sons, 1968.

[6] Sébastien Bubeck and Nicoló Cesa-Bianchi. "Regret analysis of stochastic and nonstochastic multi-armed bandit problems". In: *Foundations and Trends® in Machine Learning* 5.1 (2012), pp. 1–122.

[7] Wesley Cowan and Michael N. Katehakis. "Exploration-exploitation polices with almost sure, arbitrarily slow growing asymptotic regret". In: *Probability in the Engineering and Informational Sciences* 34 (2020), pp. 406–428.

[8] Lin Fan and Peter W Glynn. "Diffusion approximations for Thompson sampling". In: *arXiv preprint arXiv:2105.09232* (2021).

[9] Lin Fan and Peter W Glynn. "The typical behavior of bandit algorithms". In: *arXiv preprint arXiv:2210.05660* (2022).

[10] Ronan Fruit et al. "Efficient Bias-Span-Constrained Exploration-Exploitation in Reinforcement Learning". In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by Jennifer Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, 2018, pp. 1578–1586. URL: https://proceedings.mlr.press/v80/fruit18a.html.

[11] 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*. Vol. 19. PMLR, 2011, pp. 359–376.

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

[13] Xueheng Kuang and Stefan Wager. "Weak signal asymptotics for sequentially randomized experiments". In: *Management Science* (2021).

[14] Tze Leung Lai. "Adaptive treatment allocation and the multi-armed bandit problem". In: *The Annals of Statistics* 15 (1987), pp. 1091–1114.

[15] Tze Leung Lai and Herbert Robbins. "Asymptotically efficient adaptive allocation rules". In: *Ad-vances in Applied Mathematics* 6.1 (1985), pp. 4–22.

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

[17] R. Maillard O. A. Munos and G. Stoltz. "A finite-time analysis of multi armed bandits problems with Kullback-Leibler divergences". In: *COLT 2011 - The 24th Annual Conference on Learning Theory*. Vol. 11. 2011, pp. 497–514.

[18] William R. Thompson. "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples". In: *Biometrika* 25.3-4 (1933), pp. 285–294.