UCB for Large-Scale Pure Exploration: Beyond Sub-Gaussianity

Zaile Li
Technology and Operations Management Area, INSEAD, Fontainebleau, France
zaile.li@insead.edu

Weiwei Fan
Advanced Institute of Business and School of Economics and Management, Tongji University, Shanghai, China
wfan@tongji.edu.cn

L. Jeff Hong
Department of Industrial and Systems Engineering, University of Minnesota, Minneapolis, Minnesota
lhong@umn.edu

Abstract. Selecting the best alternative from a finite set represents a broad class of pure exploration problems. Traditional approaches to pure exploration have predominantly relied on Gaussian or sub-Gaussian assumptions on the performance distributions of all alternatives, which limit their applicability to non-sub-Gaussian—especially heavy-tailed—problems. The need to move beyond sub-Gaussianity may become even more critical in large-scale problems, which tend to be especially sensitive to distributional specifications. In this paper, motivated by the widespread use of upper confidence bound (UCB) algorithms in pure exploration and beyond, we investigate their performance in the large-scale, non-sub-Gaussian settings. We consider the simplest category of UCB algorithms, where the UCB value for each alternative is defined as the sample mean plus an exploration bonus that depends only on its own sample size. We abstract this into a meta-UCB algorithm and propose letting it select the alternative with the largest sample size as the best upon stopping. For this meta-UCB algorithm, we first derive a distribution-free lower bound on the probability of correct selection. Building on this bound, we analyze two general non-sub-Gaussian scenarios: (1) all alternatives follow a common location-scale structure and have bounded variance; and (2) when such a structure does not hold, each alternative has a bounded absolute moment of order $q > 3$. In both settings, we show that the meta-UCB algorithm—and therefore a broad class of UCB algorithms—can achieve the sample optimality. These results demonstrate the applicability of UCB algorithms for solving large-scale pure exploration problems with non-sub-Gaussian distributions. Numerical experiments support our results and provide additional insights into the comparative behaviors of UCB algorithms within and beyond our meta-UCB framework.

Key words: Upper confidence bound; pure exploration; large-scale; sample optimality; heavy tails

1. Introduction

Pure exploration refers to an important class of selection problems where the main objective is to identify the best alternative—the one with the highest mean performance—among a finite set. The true means of the alternatives are unknown, but the decision-maker is allowed to collect noisy observations through stochastic simulation or experimentation to estimate their values, so that an informed selection decision can be made. This best-seeking problem under uncertainty has been extensively studied in (at least) two parallel literature: ranking and selection (Bechhofer 1954, Kim and Nelson 2006) in stochastic simulation, and best arm identification (Audibert et al.

2010) in machine learning. Despite using different terminology, both frameworks share the same fundamental goal: to design efficient algorithms that adaptively allocate exploration effort across the alternatives to effectively leverage the sample information gathered during exploration, so as to maximize selection accuracy while minimizing the total sampling cost.

A wide range of algorithmic approaches have been developed for efficient pure exploration. Prevalent examples include pairwise comparison (Kim and Nelson 2006), optimal computing budget allocation (Glynn and Juneja 2004, Chen and Lee 2011), Bayesian expected value of information (Frazier et al. 2008), sequential elimination (Karnin et al. 2013), and Thompson sampling (Russo 2020, Qin and You 2025), among others. We refer readers to Hong et al. (2021) and Lattimore and Szepesvári (2020) for comprehensive reviews. Among the many approaches, one particularly interesting idea is the use of upper confidence bound (UCB) algorithms (Lai 1987, Auer et al. 2002)—a cornerstone of the multi-armed bandit literature. While UCB was originally designed for cumulative reward maximization in online learning, its structural simplicity—making sampling decisions based solely on per-alternative UCB indices—has made it widely applicable to many learning tasks (see, e.g., Agrawal 2019), with pure exploration being no exception (see, e.g., Audibert et al. 2010 and Jamieson and Nowak 2014). These works have made significant theoretical progress and have found important applications in areas such as Monte Carlo tree search (Fu 2016), drug discovery (Terayama et al. 2018), and prompt engineering for large language models (Pryzant et al. 2023).

A predominant assumption in these pure exploration works is that the performance distributions of alternatives are Gaussian or, more generally, sub-Gaussian. While this assumption offers significant analytical convenience and often forms the foundation for algorithm design, it also limits the scope of applicability. In ranking and selection, assuming Gaussianity is a longstanding convention, typically justified by central limit theorem arguments—the batch mean of simulation outputs is often approximately normal (Kim and Nelson 2006). This normality assumption supports a range of analytical tools, such as Brownian motion approximations and Bayesian conjugacy (see, e.g., Fan et al. 2016 and Chick and Inoue 2001). In best arm identification, the assumption of sub-Gaussianity is more prevalent. Sub-Gaussian distributions include many commonly encountered families, such as Gaussian, Bernoulli, and all bounded distributions, making them broadly accepted in the design of pure exploration algorithms (see, e.g., Jamieson et al. 2014).

However, non-sub-Gaussian problems frequently arise in practice. In light-tailed problems, the alternative distributions exhibit exponential tail decay but are slower than sub-Gaussian ones; the exponential distribution is a canonical example, commonly seen in the analysis of queueing systems (Gao and Gao 2016, Zhang et al. 2020). In heavy-tailed problems, the distributions have tails that decay more slowly than exponentially, for example, sub-exponentially in the log-normal family, or polynomially in the Pareto family. Heavy-tailed distributions are particularly relevant in contexts where extreme values are more likely, including reliability engineering (John and Chen 2006, Chen 2011), healthcare operations (Strum et al. 2000), financial portfolio management (Panahi 2016), and network control (Abry et al. 2010). Non-sub-Gaussianity introduces significant challenges for pure exploration, as many tools and theoretical guarantees built on sub-Gaussianity are no longer valid. These challenges have recently motivated growing interest in developing methodologies for handling general, especially heavy-tailed, distributions (see, e.g., Glynn and Juneja 2018 and Agrawal et al. 2020). A detailed review of recent progress is provided in Section 1.1.

The importance of carefully addressing non-sub-Gaussianity is particularly pronounced when the number of alternatives k is large. In practice, the true distributional form of each alternative is often unknown, and a pragmatic strategy is to apply prominent sub-Gaussian algorithms heuristically. Interestingly, despite the lack of theoretical guarantees, this approach has sometimes been shown to perform well in small-scale problems, where k is relatively small (e.g., fewer than 20), even under heavy-tailed distributions (Nelson et al. 2001, Shin et al. 2018). One intuition may come from the *high-confidence asymptotic regime*, where k is fixed and the probability of correct selection (PCS) is driven close to one (Dong and Zhu 2016). In this regime, the total sampling budget B grows large, and each alternative is allocated a growing number of observations. As B increases, the sample averages converge, and by the central limit theorem, they may approximate normality. However, as the number of alternatives k increases, the reliability of this strategy deteriorates. Nelson et al. (2001) and Shin et al. (2018) demonstrate that Gaussian-specific algorithms can suffer substantial performance loss under heavy-tailed distributions when k grows large (e.g., $k \approx 500$). This degradation likely stems from the accumulation of misspecification errors introduced by extreme values across alternatives—errors that may be negligible in small-scale problems but become significantly amplified in large-scale ones. These observations highlight the necessity of designing and analyzing algorithms that remain effective in large-scale, non-sub-Gaussian environments.

Coinciding with the challenges posed by non-sub-Gaussianity, recent years have seen a growing need to address large-scale pure exploration problems involving a significant number of alternatives (Jamieson et al. 2013, Luo et al. 2015, Ni et al. 2017). Practical examples, such as those studied by Luo et al. (2015) and Pei et al. (2024), feature thousands to over one million alternatives. To model these settings, Jamieson et al. (2013) and later Zhong and Hong (2022) consider a new *large-scale asymptotic regime*, in which $k \to \infty$. Under this regime, understanding how an algorithm's required sampling budget scales with k becomes crucial. Sample-optimal algorithms—those requiring a budget growing at the minimal possible rate, often $\mathcal{O}(k)$—are particularly desirable and have become a key benchmark for large-scale pure exploration (Hong et al. 2022, Zhang and Peng 2024, Li et al. 2025). However, this large-scale regime challenges the assumptions that underlie small-scale pure exploration algorithms. In particular, the central limit theorem argument, which assumes a sufficiently large number of observations per alternative, may no longer hold. With only $\mathcal{O}(k)$ total sampling budget, the average number of observations per alternative remains bounded, and the cumulative effect of increased likelihood of extreme values can become severe as k grows. Yet, to the best of our knowledge, existing works in this regime continue to assume Gaussianity or sub-Gaussianity, leaving this critical issue of non-sub-Gaussianity largely unexplored.

In this work, motivated by the versatility of UCB algorithms, we explore their performance in the challenging setting of large-scale pure exploration with non-sub-Gaussian distributions. Interestingly, we find that UCB algorithms can achieve the sample optimality in this regime without requiring sub-Gaussianity. We study a broad class of UCB algorithms, develop a unified analysis of their PCS, and establish sufficient conditions under which they achieve the sample optimality. These findings may offer a promising approach to addressing the challenge of moving beyond sub-Gaussianity in large-scale pure exploration and further highlight the applicability of UCB algorithms in the pure exploration landscape.

UCB algorithms make sampling decisions in a sequential manner, typically selecting at each round the alternative with the highest current UCB value. The definition of the UCB value varies across algorithmic variants. In this work, we focus on a class of UCB algorithms in which the UCB value for each alternative is defined as the sample mean plus a non-negative bonus term that depends deterministically on its own sample size. This class represents the simplest category of UCB algorithms. To support a unified analytical framework, we consider a meta-UCB algorithm in which the bonus function is arbitrary but required to vanish asymptotically—that is, it converges to zero as the sample size approaches infinity. This meta-algorithm encompasses a number of well-known UCB variants, including UCBE (Audibert et al. 2010) and MOSS (Audibert and Bubeck 2009). Additional examples are discussed in Section 2.2. We adopt a fixed-budget formulation of pure exploration, in which the total sampling budget is predetermined. Under this formulation, a decision rule (or selection standard) is applied at the end of the sampling process to select one alternative as the best, based on the collected sample information. While the most common rule is to select the alternative with the largest sample mean, we consider an alternative rule that selects the alternative with the *largest sample size* (Jamieson et al. 2014, Andradóttir and

Prudius 2009). Interestingly, this selection rule may facilitate a simplified analysis and enable us to obtain a rich set of theoretical results.

We begin our analysis of the meta-UCB algorithm by studying its PCS. Despite the clear structure of UCB algorithms, analyzing their PCS is far from straightforward due to their sequential and adaptive nature. Existing analyses often rely heavily on sub-Gaussian assumptions and result in vacuous (e.g., negative) PCS lower bounds when applied to large-scale problems. Inspired by ideas from Li et al. (2025), we approach the analysis from a boundary-crossing perspective. Specifically, we treat the stochastic UCB process (defined in Section 3.1) of the best alternative as a boundary and view the UCB processes of the non-best alternatives as crossing below this boundary during the sampling process. This perspective enables us to derive upper bounds on the number of observations allocated to each non-best alternative in terms of their boundary-crossing times. Subsequently, the selection standard of the largest sample size enables us to establish a clean PCS lower bound for the meta-UCB algorithm. Notably, this PCS lower bound is distribution-free. Moreover, for any specific UCB algorithm within the meta class, its PCS lower bound can be obtained directly by substituting its bonus function into the lower bound. The PCS lower bound also provides intuitive structural insights into the dual role of the bonus function in UCB algorithms: enabling effective exploration while controlling the associated sampling cost. See Section 3.2 for a more detailed discussion.

Building on the PCS analysis, we examine the performance of the meta-UCB algorithm in solving large-scale, non-sub-Gaussian pure exploration problems. We begin with the indifference-zone (IZ) formulation, which assumes that the mean gap between the best and second-best alternatives remains constant as the number of alternatives increases. This formulation, also referred to as the sparse configuration (Jamieson et al. 2013), is standard in the study of sample optimality (Hong et al. 2022, Li et al. 2025). Under the IZ formulation, we first consider scenarios where the performance distributions of all alternatives follow a location-scale structure—a generalization of the Gaussian case—and assume only that their variances are bounded. This includes the important case where the alternatives are generated from a linear model (Zhou and Ryzhov 2025). Under this condition, we prove that the meta-UCB algorithm achieves the sample optimality: with a total sampling budget that grows linearly with k, the algorithm maintains a non-zero PCS asymptotically. Furthermore, we consider the broader setting where the location-scale structure does not hold. In this case, we impose a mild moment condition: each alternative has a bounded absolute moment of order $q > 3$. This condition is broad enough to encompass all light-tailed distributions and a wide class of heavy-tailed distributions. Under this assumption, we develop new technical tools to analyze the boundary-crossing times of UCB processes and once again establish the sample optimality of the meta-UCB algorithm. These sample optimality results demonstrate that the UCB approach can be efficient in solving large-scale pure exploration problems, not only for traditional sub-Gaussian settings, but also for heavy-tailed ones.

We then explore whether the meta-UCB algorithm can achieve the sample optimality beyond the IZ formulation, to further test the robustness of the UCB approach. Without the IZ assumption, the mean gap between the best and some other alternatives may become arbitrarily small as k increases. In this challenging setting, rather than relaxing the objective to identifying a "good enough" alternative (Ni et al. 2017, Eckman and Henderson 2018), we continue to focus on the PCS. To achieve the sample optimality, we identify a key condition: the UCB value of the best alternative should always exceed its true mean with non-zero probability. This requirement is natural for UCB algorithms—it aligns with the interpretation of (anytime-valid) confidence bounds—but may not be satisfied by existing IZ-requiring algorithms (Li et al. 2025). To model the non-IZ scenarios, we consider a representative problem configuration where the mean gaps shrink at a certain rate as k grows. Under the finite absolute moment condition of order $q > 3$, we show that the meta-UCB algorithm can still achieve the sample optimality, provided the mean-gap shrinkage rate satisfies a condition that links the admissible rate to the heaviness of the tails, as captured by q. As the problem becomes more heavy-tailed, achieving sample optimality becomes more difficult, and the allowable shrinkage rate becomes slower. This intuitive and positive result further highlights the robustness of UCB algorithms for large-scale pure exploration.

Lastly, we conduct a series of numerical experiments to empirically validate our theoretical findings. We evaluate several UCB algorithms and find that, under the proposed conditions, they indeed exhibit sample optimality when solving large-scale pure exploration problems with non-sub-Gaussian alternatives. In addition, we examine the classic UCB1 algorithm of Auer et al. (2002), a representative example of algorithms that fall outside our meta-UCB class. Interestingly, it performs significantly worse than those within the meta-UCB class and appears to lack sample optimality. To understand this performance gap, we analyze the budget allocation behavior of the algorithms. We observe that algorithms in the meta-UCB class do display boundary-crossing dynamics, while UCB1 does not. As a result, UCB1 tends to over-explore, allocating the sampling budget more uniformly across all alternatives, which leads to inefficiency and ultimately undermines sample optimality. This comparison highlights the importance of our selected class of UCB algorithms.

It is important to acknowledge that our perspective in this paper is primarily theoretical. We do not aim to design new UCB algorithms, or optimize the choice of bonus functions, or address practical implementation issues. That said, we believe the proposed new results and insights can serve as a foundation for future work aimed at designing practically effective UCB algorithms for large-scale, non-sub-Gaussian (especially heavy-tailed) pure exploration problems. In addition, we emphasize that our analysis focuses on algorithms with UCB values defined as the sample mean plus a bonus function. Other forms of UCB are also worth exploring, particularly those based on robust mean estimators—such as truncated or weighted sample means—that are designed to handle heavy-tailed distributions (Bubeck et al. 2013, Glynn and Juneja 2018). We discuss and conjecture in Section 5.2 that the sample optimality results may extend to these broader classes when the UCB values of different alternatives are decoupled—that is, each value depends only on the alternative's own sample size and observations. However, formally verifying these results would likely require case-by-case analysis, which we leave for future work.

1.1. Related Literature

The ranking and selection and best arm identification literature has proposed several approaches to relax the (sub-)Gaussianity assumption. The central-limit-theorem approach approximates the distribution of scaled partial sums of non-Gaussian observations with a Gaussian limit in asymptotic regimes where the sample size of each alternative tends to infinity (Toscano-Palmerin and Frazier 2015, Lee and Nelson 2016, Fan et al. 2016). The large-deviation (LD) approach, pioneered by Glynn and Juneja (2004), leverages the LD rate function of each alternative to solve the optimal budget allocation problem and is naturally suited for light-tailed observations (see, e.g., Gao et al. 2017). It has also been extended by Broadie et al. (2007) and Blanchet et al. (2008) to accommodate heavy-tailed distributions; see Dong and Zhu (2016) for an overview of these two approaches. A third class of methods employs robust mean estimators—such as the truncated sample mean—to mitigate the influence of heavy tails (Yu et al. 2018, Glynn and Juneja 2018). Most recently, Agrawal et al. (2020, 2021) introduce a change-of-measure technique for minimizing the sampling budget needed to achieve a target PCS in heavy-tailed settings. In addition, there are also algorithmic approaches, including those of Garivier and Kaufmann (2016) and Russo (2020), that naturally accommodate alternative distributions from the same one-parameter exponential family. While these approaches provide valuable tools and insights for handling non-sub-Gaussian distributions, they are primarily developed for small-scale problems. In contrast, our work focuses on the large-scale setting and demonstrates that simple UCB algorithms can remain robust and sample-efficient even in the presence of heavy tails.

Our work also responds to the growing interest in understanding the fundamental behaviors of UCB algorithms (see, e.g., Fan and Glynn 2024) and their performance in unconventional settings, such as learning with heavy-tailed observations (Bubeck et al. 2013) or with a large number of arms (Chan and Hu 2019, Bayati et al. 2020). These studies often report negative results, showing that traditional UCB algorithms may suffer from sub-optimal performance in such challenging environments. In contrast, we present sample optimality results for large-scale pure exploration under mild distributional assumptions. These results may offer new insights into their performance and further highlight their potential for pure exploration.

The remainder of this paper is organized as follows. In Section 2, we provide the problem formulation and introduce the meta-UCB algorithm. In Section 3, we analyze the PCS of the meta-UCB algorithm and examine the role of the bonus function. In Section 4, we establish the sample optimality of the meta-UCB algorithm under the IZ formulation. In Section 5, we explore extensions beyond the IZ formulation and discuss broader classes of UCB algorithms not covered by our analysis. In Section 6, we conduct numerical experiments to verify our theoretical results. Finally, we conclude the paper in Section 7 and include supplementary technical details in the E-Companion.

2. Problem Statement

In this section, we first introduce the notations and preliminaries for large-scale pure exploration. We then present a general class of Upper Confidence Bound (UCB) algorithms, referred to as the meta-UCB algorithm, which serves as our primary focus for addressing this problem under general distributional settings.

2.1. Notations and Preliminaries

Let k denote the number of alternatives in a pure exploration problem, which represents the scale of the problem. For each alternative $i = 1, \ldots, k$, let X_i be a random variable representing its performance, and let $\mu_i = \mathrm{E}[X_i]$ denote its unknown mean. The objective is to identify the best alternative—that is, the one with the largest mean performance $\max_{i=1,\cdots,k} \mu_i$. Without loss of generality, we assume that alternative 1 is the best. Since the means μ_i are unknown, one must sample the alternatives—i.e., collect independent and identically distributed (i.i.d.) observations $X_{i,j}$, $j = 1, 2, \ldots$, from each alternative i—to obtain sample information, so that an informed selection decision can be made. In this paper, we adopt the fixed-budget formulation, where the total sampling budget B is predetermined. By sequentially allocating this budget across alternatives, an algorithm selects one alternative, denoted by i^*, as the best based on the collected sample information, once the budget is exhausted. The efficiency of such an algorithm is measured by the probability of correct selection (PCS), defined as $\Pr\{i^* = 1\}$.

The performance of pure exploration algorithms depends heavily on the distributional characteristics of the alternatives. Conventionally, it is often assumed that each X_i follows a Gaussian or, more generally, a sub-Gaussian distribution (Kim and Nelson 2006, Audibert et al. 2010). While these assumptions offer technical convenience, they may limit the applicability of the resulting algorithms. In this paper, we explore settings where such assumptions do not hold—specifically, non-sub-Gaussian problems, including heavy-tailed cases. As previously discussed, the challenges posed by non-sub-Gaussianity may become particularly pronounced in large-scale problems, where the number of alternatives k is large. In large-scale pure exploration, the key performance measure is how the total sampling budget must grow with k to maintain meaningful selection performance, specifically, a non-zero PCS. This scaling behavior has been formalized in recent work (Jamieson et al. 2013, Tanczos et al. 2017, Zhong and Hong 2022), which shows that achieving a non-zero PCS as $k \to \infty$ requires a total sampling budget that grows at least at order $\mathcal{O}(k)$. An algorithm that meets this lower bound can be said to be *sample optimal*. In the fixed-budget setting, Hong et al. (2022) formalize this idea and define the sample optimality as follows.

DEFINITION 1 (SAMPLE OPTIMALITY). A pure exploration algorithm is sample optimal if there exists a pair of constants $\alpha \in (0,1)$ and $c > 0$ such that

$$\liminf_{k\to\infty} \text{PCS} > \alpha \text{ for } B = ck. \tag{1}$$

This notion of sample optimality provides a benchmark for evaluating algorithm performance in large-scale pure exploration and serves as the primary metric in this paper. While existing work on large-scale pure exploration has focused exclusively on problems where all alternatives follow sub-Gaussian distributions, we investigate the performance of UCB algorithms in the more challenging non-sub-Gaussian regime.

2.2. A Meta-UCB Algorithm

UCB algorithms allocate the sampling budget sequentially—i.e., one observation at a time—based on a UCB value assigned to each alternative. This value serves as an "optimistic" estimate of the unknown mean and is typically defined as the upper bound of a confidence interval. The gap between the UCB value and the sample mean captures the *exploration bonus* (Lattimore and Szepesvári 2020). A wide variety of UCB algorithms have been proposed, differing in how the UCB values are constructed. In this work, we focus on a class of simple UCB algorithms in which the UCB value for each alternative i is defined as the sum of the sample mean $\bar{X}_i(n_i) = \frac{1}{n_i}\sum_{i=1}^{n_i} X_{i,j}$ and an exploration bonus function $f(n_i)$ that depends only on its sample size n_i. This setting represents a major category of UCB algorithms in which the UCB values are *decoupled*—each depends only on the alternative's own sample information, i.e., its sample size and observations. To unify this class, we introduce a meta-UCB algorithm that allows for a general exploration bonus function. The algorithm is presented in Algorithm 1. For the exploration bonus function, we only require that it is non-negative and converges to zero as the sample size increases to infinity. This structure is formalized in the following assumption.

Algorithm 1: Meta-UCB

Input: k alternatives $X_1, \ldots, X_k$, the total sampling budget B

for $i = 1$ **to** k **do**
  take one observation x_i from alternative X_i, set $n_i = 1$ and initialize $U_i(n_i)$;
end
while $\sum_{i=1}^{k} n_i < B$ **do**
  select the alternative $s = \arg\max_{i \in \{1,\ldots,k\}} U_i(n_i)$;
  take one observation x_s from X_s, set $n_s = n_s + 1$, and update $U_s(n_s)$;
end

Output: alternative $b = \arg\max_{i \in \{1,\ldots,k\}} n_i$

ASSUMPTION 1. *For each alternative $i = 1, \cdots, k$, the UCB value takes the form $U_i(n_i) = \bar{X}_i(n_i) + f(n_i)$, where the exploration bonus function f is non-negative and satisfies $\lim_{n_i \to \infty} f(n_i) = 0$.*

Intuitively, this assumption requires that as the sample size $n_i \to \infty$, the exploration bonus associated with the sample estimate of the true mean μ_i vanishes, so that the UCB value $U_i(n_i)$ converges to μ_i. This behavior is essential to ensure that the best alternative can be identified whenever all the alternatives have been sufficiently evaluated. The assumption is broad enough to encompass many well-known UCB algorithms. Three illustrative examples are provided below. Additional examples of bonus functions satisfying Assumption 1 can be found in, e.g., Degenne et al. (2019), Lattimore and Szepesvári (2020), Zhong et al. (2021), and Zhang and Ying (2023).

EXAMPLE 1 (THE UCB-E ALGORITHM OF AUDIBERT ET AL. 2010).

$$U_i(n_i) = \bar{X}_i(n_i) + \sqrt{\frac{a}{n_i}},$$

where a is a constant irrelevant to n_i.

EXAMPLE 2 (THE MOSS ALGORITHM OF AUDIBERT AND BUBECK 2010).

$$U_i(n_i) = \bar{X}_i(n_i) + \sqrt{\frac{1}{n_i} \max\left\{\log\left(\frac{c}{n_i}\right), 0\right\}},$$

where $c = B/k$.

EXAMPLE 3 (THE LIL-UCB ALGORITHM OF JAMIESON ET AL. 2014).

$$U_i(n_i) = \bar{X}_i(n_i) + (1+\beta)(1+\sqrt{\varepsilon})\sqrt{\frac{2\sigma^2(1+\varepsilon)\log\left(\frac{\log((1+\varepsilon)n_i)}{\delta}\right)}{n_i}},$$

where $\beta, \varepsilon, \sigma$ and δ are constants irrelevant to n_i.

It is worth highlighting that the decoupling structure in UCB algorithms satisfying Assumption 1 has important implications for both their performance in pure exploration and the techniques used to analyze them. We elaborate on these implications in Section 3.1 and Section 6.3. Readers familiar with the machine learning literature may recognize that many UCB algorithms, such as the well-known UCB1 (Auer et al. 2002), do not satisfy this decoupling structure. Interestingly, numerical experiments in Section 6 show that UCB1 behaves quite differently and may perform significantly worse compared to algorithms conforming to Assumption 1. We will discuss algorithms such as UCB1 that do not satisfy Assumption 1 in Section 5.2.

We would also like to highlight that in Algorithm 1, the alternative with the largest sample size is selected as the best when the total sampling budget is exhausted. This selection criterion, though less common, has been adopted in prior work—for example, by Jamieson et al. (2014)—and, interestingly, has also appeared in early simulation optimization literature (see, e.g., Andradóttir and Prudius 2009). While the more conventional and intuitive selection criterion is the largest sample mean, we will show that using the largest sample size as the selection criterion offers analytical advantages that facilitate performance analysis.

3. PCS Analysis of the Meta-UCB Algorithm

In this section, we analyze the PCS of the meta-UCB algorithm when applied to a generic pure exploration problem. This analysis lays the foundation for further understanding the algorithm's performance in solving large-scale, non-sub-Gaussian problems. Prior studies on UCB algorithms have attempted similar analyses, but they often fail to yield meaningful insights for large-scale problems—resulting in uninformative bounds (e.g., lower bounds of $-\infty$ for the PCS as $k \to \infty$)—or rely heavily on sub-Gaussian assumptions (Audibert et al. 2010, Jamieson et al. 2014). We approach the analysis from a boundary-crossing perspective inspired by Li et al. (2025), where this perspective was used to examine the sample means of different alternatives to understand the performance of greedy algorithms for Gaussian pure exploration. In Section 3.1, we introduce the boundary-crossing perspective to review the sequential sampling process through UCB value processes and derive a neat and distribution-free PCS lower bound for the meta-UCB algorithm. This PCS analysis may highlight the role of the exploration bonus in pure exploration, which we further explore in Section 3.2.

3.1. A Boundary-Crossing Perspective

To characterize the PCS of the meta-UCB algorithm, we adopt a boundary-crossing perspective to analyze its sampling process. Under this perspective, we attempt to regard the best alternative as a *boundary* for all non-best alternatives and model the sampling dynamics of each non-best alternative as a *boundary-crossing process* relative to the boundary. To illustrate the key ideas, let us consider a concrete example: the UCBE algorithm with UCB value $U_i(n) = \bar{X}_i(n) + \sqrt{0.2/n}$, applied to solve an example problem with only two alternatives where $X_1 \sim \text{Normal}(0.1, 0.25)$ and $X_2 \sim \text{Normal}(0, 0.25)$. Figure 1 shows a realized sampling process of the UCB algorithm, illustrating how the sample means and UCB values for both alternatives evolve as the total sample size increases.

Figure 1 A Sampling Process of the UCB algorithm with $U_i(n) = \bar{X}_i(n) + \sqrt{0.2/n}$ for a Problem with 2 Alternatives.

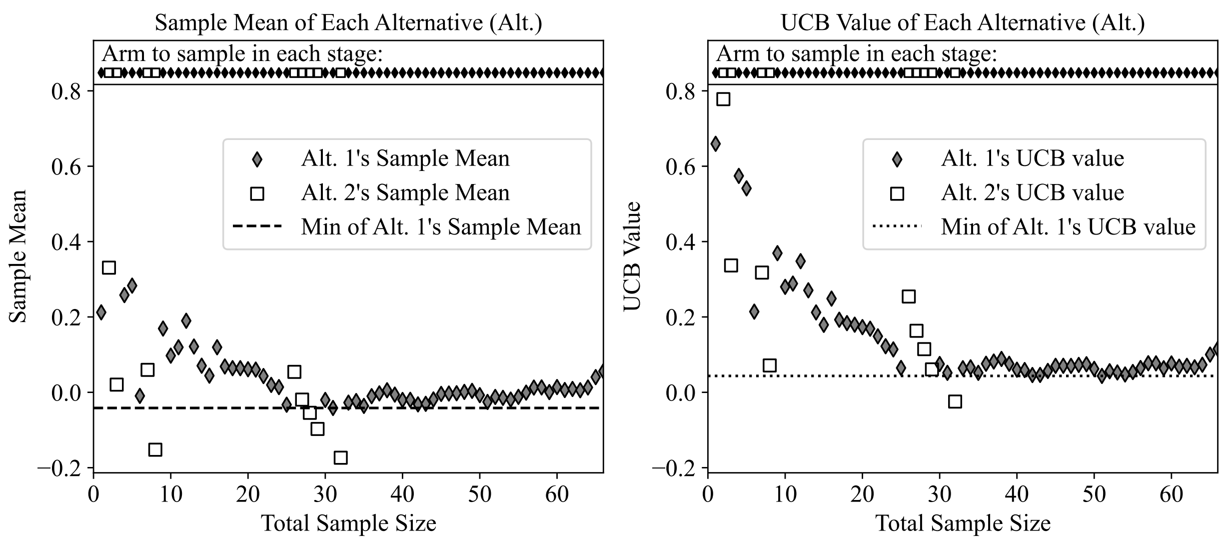


Figure 1 reveals a critical observation: *once the UCB value of alternative 2 falls below the minimum of alternative 1's UCB values, it may never be selected and sampled again*. This phenomenon arises from the decoupled structure of UCB values across different alternatives, as highlighted in Assumption 1. Since the UCB values are computed independently for each alternative, sampling alternative 1 does not affect the UCB value of alternative 2. As a result, regardless of how many additional observations alternative 1 receives, alternative 2 remains uncompetitive once its UCB value falls below the minimum of alternative 1's UCB values. In this sense, the minimum UCB value of the best alternative effectively forms a boundary for all non-best alternatives. This observation motivates a boundary-crossing perspective for analyzing the meta-UCB algorithm.

To formalize the observation, we define the stochastic process $\{U_i(n), n = 1, 2, \ldots, \infty\}$ as alternative i's *UCB value process*. These UCB value processes for all alternatives collectively determine the meta-UCB algorithm's sampling process across the alternatives. For the best alternative, we define

$$U_1^* = \inf_{n \in [1,\infty)} U_i(n)$$

as the minimum of its UCB value process. Notice that U_1^* is a random variable defined over the entire UCB value process from $n=1$ to ∞, and it is independent of the sampling behavior of the algorithm. In cases with $U_i(n) > \mu_i$ for all $n \geq 1$ (see Section 5.1 for examples), $U_1^* = \mu_1$. Inspired by Figure 1, we propose using U_1^* as the boundary for all non-best alternatives. From this boundary-crossing perspective, we arrive at the following property of the algorithm.

PROPERTY 1. *During the sampling process of the meta-UCB algorithm, for every non-best alternative $i = 2, \cdots, k$, once its UCB value $U_i(n)$ falls below U_1^*, it will be dominated by alternative 1 and never be sampled again.*

Property 1 is particularly useful, as it allows us to directly obtain an upper bound on the allocated sample size of each non-best alternative. This result is summarized in the following property.

PROPERTY 2. *For the meta-UCB algorithm, regardless of the total sampling budget B, the total number of observations allocated to each non-best alternative $i = 2, \cdots, k$, denoted by $n_i(B)$, satisfies*

$$n_i(B) \leq T_i^f(U_1^*) := \inf\{n \geq 1 : U_i(n) < U_1^*\}, \quad a.s., \tag{2}$$

where $T_i^f(U_1^*)$ denotes the boundary-crossing time of alternative i regarding the boundary U_1^*.

REMARK 1. Both U_1^* and $n_i(B)$ are random variables defined on the same sample path, which is composed of the UCB value processes $\{U_i(n), n = 1, 2, \ldots, \infty\}$ of all alternatives. Notice that $n_i(B)$ may be infinite when U_1^* is lower than the minimum UCB value of alternative i. However, this possibility does not affect the subsequent PCS analysis.

With this property, we are ready to study the PCS of the meta-UCB algorithm. Recall that the algorithm selects the alternative with the largest sample size as the best alternative when the total sampling budget is exhausted. Thus, the PCS of the meta-UCB algorithm can be expressed as

$$\text{PCS} = \Pr\left\{ n_1(B) > \max_{i=2,\cdots,k} n_i(B) \right\} = \Pr\left\{ B > \max_{i=2,\cdots,k} n_i(B) + \sum_{i=2}^{k} n_i(B) \right\}. \tag{3}$$

Combining Equations (2) and (3), we conclude that under a total sampling budget B, the PCS of the meta-UCB algorithm satisfies

$$\text{PCS} \geq \Pr\left\{ B > \max_{i=2,\cdots,k} T_i^f(U_1^*) + \sum_{i=2}^{k} T_i^f(U_1^*) \right\} \geq \Pr\left\{ B > 2\sum_{i=2}^{k} T_i^f(U_1^*) \right\}. \tag{4}$$

Equation (4) presents a lower bound on the PCS for the meta-UCB algorithm. Notably, this result is distribution-free. The bound depends on the distributional information of the problem only through the boundary-crossing times. Moreover, the PCS bound holds for any problem instance with a unique best alternative and applies to all UCB algorithms that satisfy Assumption 1. This distribution-free PCS lower bound sets the stage for our subsequent analysis of the meta-UCB algorithm—and therefore a class of UCB algorithms—in solving pure exploration problems with general performance distributions. These analyses are developed in Section 4.

To the best of our knowledge, Equation (4) provides the first PCS lower bound for UCB algorithms that is expressed in terms of boundary-crossing times. Both the simplicity of the bound and the elegance of its derivation are noteworthy. From Equations (3) and (4) we can see that a key contributor to this simplicity is our choice of selection criteria: let the algorithm select the alternative with the largest sample size at the end of the sampling process. While other selection criteria—such as the largest sample mean or UCB value—could be considered, we leave the analysis of those alternatives to future work. Empirically, numerical experiments suggest that the choice of selection criteria does not lead to a significant performance difference; see Section EC.2 for the details. Hence, we focus on the largest sample size criterion in this paper.

3.2. Effect of Exploration Bonus

The PCS analysis above provides an alternative lens through which to understand the effect of the exploration bonus function in UCB algorithms. To simplify the presentation, in the following we focus on the special case with $k=2$. The insights gained in this two-alternative scenario can be easily generalized to cases with $k>2$. Directly from Property 2, we have

$$\Pr\left\{n_1(B)>n_2(B) \text{ when } B>2T_2^f(U_1^*)\right\}=1, \tag{5}$$

which means that $2T_2^f(U_1^*)$ can serve as an upper bound for the required sampling budget to ensure a correct selection. To further expose the effect of f on $T_2^f(U_1^*)$, we introduce a constant $\gamma\in[0,\infty)$ to define a constant boundary $\mu_1-\gamma$. We then consider the associated boundary-crossing time

$$T_2^f(\mu_1-\gamma):=\inf\{n\geq 1: U_2(n)<\mu_1-\gamma\}=\inf\{n\geq 1:\bar{X}_2(n)+f(n)<\mu_1-\gamma\}.$$

Conditional on the event $\{U_1^*>\mu_1-\gamma\}$, we know that $T_2^f(\mu_1-\gamma)\geq T_2^f(U_1^*)$. Consequently,

$$\begin{aligned}
&\Pr\left\{n_1(B)>n_2(B) \text{ when } B>2T_2^f(\mu_1-\gamma)\right\}\\
&\quad\geq \Pr\left\{n_1(B)>n_2(B) \text{ when } B>2T_2^f(\mu_1-\gamma)\mid U_1^*\geq\mu_1-\gamma\right\}\Pr\left\{U_1^*\geq\mu_1-\gamma\right\}\\
&\quad\geq \Pr\left\{n_1(B)>n_2(B) \text{ when } B>2T_2^f(U_1^*)\mid U_1^*\geq\mu_1-\gamma\right\}\Pr\left\{U_1^*\geq\mu_1-\gamma\right\}\\
&\quad= \Pr\left\{U_1^*\geq\mu_1-\gamma\right\}=\Pr\left\{\bar{X}_1(n)+f(n)\geq\mu_1-\gamma,\,\forall n\geq 1\right\}=:P_f,
\end{aligned} \tag{6}$$

where the first equality arises from Equation (5). Equation (6) is noteworthy as it suggests that $2T_2^f(\mu_1 - \gamma)$ can be viewed as an upper bound for the meta-UCB algorithm's required sampling budget to guarantee a PCS of at least P_f. This result highlights the dual role played by the non-negative exploration bonus function f in UCB algorithms. On the one hand, it controls exploration. Increasing the exploration bonus amplifies the optimism in the UCB values, thereby improving the chance of correctly identifying the best alternative P_f. This aligns with the conventional understanding that the exploration bonus helps promote better selection. On the other hand, a higher exploration bonus also increases the boundary-crossing time $T_2^f(\mu_1 - \gamma)$, which implies a larger required sampling budget. In this light, boundary-crossing times offer a tangible way to quantify the cost of using the exploration bonus.

4. Sample Optimality of the Meta-UCB Algorithm

By leveraging the distribution-free PCS lower bound in Equation (4), we now investigate the properties of the meta-UCB algorithm in solving non-sub-Gaussian pure exploration problems. In Section 4.1, we formalize the sufficient conditions for the meta-UCB algorithm to achieve the sample optimality. In Section 4.2, we present two important classes of non-sub-Gaussian scenarios in pure exploration. In Sections 4.3.1 and 4.3.2, we demonstrate that the sufficient conditions can be satisfied under these two scenarios.

4.1. Sufficient Conditions for Sample Optimality

Recall from Equation (4) that the PCS of the meta-UCB algorithm satisfies

$$\mathrm{PCS} \geq \Pr\left\{ B > 2\sum_{i=2}^{k} T_i^f(U_1^*) \right\}. \tag{7}$$

To study the sample optimality, we need to drive $k \to \infty$ and analyze the behavior of the sum of $k-1$ boundary-crossing times. A key technical challenge arises from the dependence across these times due to the shared random boundary U_1^*. The conditioning arguments in Section 3.2 provide a way to get around this. Consider the event $\{U_1^* \geq \mu_1 - \gamma_0\}$ for some $\gamma_0 \in [0, \infty)$. Then Equation (7) implies

$$\begin{aligned} \mathrm{PCS} &\geq \Pr\left\{ B > 2\sum_{i=2}^{k} T_i^f(U_1^*) \middle| U_1^* \geq \mu_1 - \gamma_0 \right\} \Pr\left\{ U_1^* \geq \mu_1 - \gamma_0 \right\} \\ &\geq \Pr\left\{ B \geq 2\sum_{i=2}^{k} T_i^f(\mu_1 - \gamma_0) \right\} \Pr\left\{ U_1^* \geq \mu_1 - \gamma_0 \right\}. \end{aligned} \tag{8}$$

This decomposition effectively decouples the best alternative from the non-best ones, and the boundary-crossing times $T_i^f(\mu_1 - \gamma_0)$ are mutually independent as now the boundary is a constant. From Equation (8), considering the asymptotic regime of letting $k \to \infty$, we can see that the following conditions would be sufficient to show the sample optimality.

CONDITION 1. *For some* $\gamma_0 \in [0,\infty)$, $\Pr\{U_1^* \geq \mu_1 - \gamma_0\} = \Pr\{U_1(n) \geq \mu_1 - \gamma_0, \forall n \geq 1\} > 0$.

CONDITION 2. *For the* $\gamma_0 \in [0,\infty)$ *satisfying Condition 1, the sum* $\sum_{i=2}^{k} T_i^f(\mu_1 - \gamma_0)$ *grows at most linearly in* k*, i.e., there exists a constant* $c(\gamma_0) > 0$ *such that* $\limsup_{k\to\infty} \frac{1}{k}\sum_{i=2}^{k} T_i^f(\mu_1 - \gamma_0) \leq c(\gamma_0)$ *almost surely.*

Under these two conditions, it is evident that given a total sampling budget $B = 2c(\gamma_0)k$, the PCS of meta-UCB satisfies

$$\liminf_{k\to\infty} \mathrm{PCS} \geq \Pr\{U_1^* \geq \mu_1 - \gamma_0\} \liminf_{k\to\infty} \Pr\left\{ c(\gamma_0) \geq \frac{1}{k}\sum_{i=2}^{k} T_i^f(\mu_1 - \gamma_0) \right\} = \Pr\{U_1^* \geq \mu_1 - \gamma_0\} > 0,$$

which satisfies the notion of sample optimality as defined in Definition 1. Therefore, if these two conditions can be verified, the sample optimality of the meta-UCB algorithm would follow. Condition 1 requires that the UCB value of the best alternative (alternative 1) remains above $\mu_1 - \gamma_0$ with a non-zero probability. This aligns with the interpretation of upper confidence bounds in UCB algorithms (particularly in the case with $\gamma_0 = 0$). Notably, this condition only concerns alternative 1 and does not impose requirements on the other alternatives. This is intuitive—ensuring a uniformly non-vanishing confidence level across all k alternatives becomes increasingly infeasible as $k \to \infty$, due to multiplicity effects. Condition 2 is essentially a strong law of large numbers-type requirement for the boundary-crossing times $T_i^f(\mu_1 - \gamma_0)$. Both conditions ultimately depend on the distributional properties of the alternatives.

4.2. Non-Sub-Gaussian Pure Exploration Problems

Conditions 1 and 2 do not hold uniformly across all possible distributions of the alternatives. In the following, we introduce two meaningful and practically relevant non-sub-Gaussian settings. We begin with scenarios where the alternatives exhibit a location-scale structure, formalized in Assumption 2. This structure generalizes the conventional Gaussian setting where each alternative X_i is determined with a location parameter, i.e., its mean μ_i, and a scale parameter, i.e., its standard deviation σ_i. Under the structure, we only require that the reference random variable Y_i has a finite mean and variance. Consequently, Assumption 2 accommodates at least two important classes of non-sub-Gaussian settings. First, it includes cases where the alternatives follow a common location-scale family, such as the Student's t distribution, which is heavy-tailed and has attracted attention in the literature (e.g., Shin et al. 2018). Such location-scale families are widely used in decision analysis (Meyer 1987). Second, it encompasses settings where alternatives are generated through a linear model; see, e.g., Zhou and Ryzhov (2025) and the reference therein.

ASSUMPTION 2 **(Location-Scale Structure).** *For each alternative $i = 1, \ldots, k$,*

$$X_i = \mu_i + \sigma_i Y_i,$$

where Y_i has a continuous distribution and is i.i.d. across different alternatives. Without loss of generality, assume that $\mathrm{E}[Y_i] = 0$ *and* $\mathrm{Var}[Y_i] = 1$*. Furthermore,* $\max_{i=1,\cdots,k} \sigma_i < \infty$.

REMARK 2. The continuity condition in Assumptions 2 and 3 is imposed solely to avoid ties in the UCB values, thereby streamlining the analysis. It is not required for implementing the meta-UCB algorithm in practice, where ties can be broken arbitrarily.

In scenarios where a location-scale structure does not hold among the alternatives, we consider a general moment condition on their performance distributions, as stated in Assumption 3. Similar moment assumptions have been widely adopted in the study of non-sub-Gaussian pure exploration problems (see, e.g., Glynn and Juneja 2018, Agrawal et al. 2020).

ASSUMPTION 3 **(Moment Condition).** *For each alternative $i = 1, \cdots, k$, X_i has a continuous distribution, and there exists a pair of constants $q > 3$ and $M > 0$ such that*

$$\max_{i=1,\cdots,k} \mathrm{E}[|X_i|^q] \leq M$$

regardless of how large k is.

REMARK 3. A consequence of Assumption 3 is that the variance of each alternative is upper bounded by the same constant $4M^{\frac{2}{q}}$, as shown in Lemma EC.4.

This assumption is pragmatically mild. It encompasses most commonly encountered distributions that possess an absolute moment of order $q > 3$. This includes all light-tailed distributions—such as the exponential—and extends to many heavy-tailed ones, including the lognormal distribution, the Student's t distribution (with degrees of freedom higher than q), and the Pareto distribution (with a shape parameter higher than q), highlighting its versatility. It also allows the observations of different alternatives to follow distinct distribution families. The existence of a common M for all k alternatives rules out pathological cases where moments of some alternatives diverge as $k \to \infty$. It also limits the difficulty of the problem. It is not related to the tail behavior of the individual alternative. Instead, it is a shared regularity for all alternatives. Besides, we don't require the values of M or q to be known, as they are not inputs of the meta-UCB algorithm.

Beyond these two distributional assumptions, we introduce the following structural assumption on the problem configurations of pure exploration when considering the asymptotic regime of letting $k \to \infty$. It requires that there exists at least a positive constant mean gap between the best and all other alternatives, and the variances of all alternatives are uniformly bounded.

ASSUMPTION 4. *There exists positive constants γ, $\underline{\sigma}^2$, and $\bar{\sigma}^2$ such that $\mu_1 - \max_{i=2,\cdots,k} \mu_i \geq \gamma$ and $\underline{\sigma}^2 \leq \sigma_i^2 \leq \bar{\sigma}^2$ for all $i = 1, \cdots, k$ regardless of how large k is, where $\sigma_i^2 = \mathrm{Var}[X_i]$.*

The existence of a constant mean gap $\gamma > 0$ is a conventional, prevalent assumption in the ranking and selection literature, typically referred to as the indifference-zone (IZ) formulation (Bechhofer 1954, Kim and Nelson 2006). The formulation is also known as the sparse configuration in the best arm identification literature (Jamieson et al. 2013). Intuitively, the existence of a constant mean gap makes the problem easier to solve, as it prevents scenarios where some alternatives become nearly indistinguishable from the best alternative. Recent studies of large-scale ranking and selection all rely on such formulation to study the sample optimality for the PCS (e.g., Zhong and Hong 2022). We follow these works to adopt the formulation in this section. We will discuss scenarios where the IZ formulation fails later in Section 5.1. The existence of $\underline{\sigma}^2$ and $\bar{\sigma}^2$ prevents scenarios where the variances of some alternatives explode or diminish. Notice that we don't require the values of γ, $\underline{\sigma}^2$, and $\bar{\sigma}^2$ to be known as they are not inputs of the meta-UCB algorithm.

4.3. Sample Optimality

To demonstrate the sample optimality of the meta-UCB algorithm, we now examine Conditions 1 and 2 under the distributional assumptions introduced above. First, we observe that due to Assumption 4, the following lemma holds. This lemma shows that when an IZ mean gap $\gamma > 0$ exists and $\sigma_1^2 < \infty$, Condition 1 is satisfied for any $\gamma_0 \in (0, \gamma)$, regardless of the underlying distribution of alternative 1. Intuitively, the IZ mean gap provides a buffer zone that ensures a non-zero probability for the UCB value of the best alternative to stay above the boundary $\mu_1 - \gamma_0$ regardless of the exploration bonus function. The proof of the lemma is included in Section EC.1.1.

LEMMA 1. *Suppose that Assumption 4 holds. Then, for any bonus function f satisfying Assumption 1 and for any $\gamma_0 \in (0, \gamma)$, we have*

$$\Pr\{U_1^* \geq \mu_1 - \gamma_0\} \geq \exp\left(-\frac{\pi^2 \sigma_1^2}{6\gamma_0^2}\right) > 0.$$

In the remainder of this section, we focus on verifying Condition 2, which concerns the properties of the boundary-crossing times of the non-best alternatives $T_i^f(\mu_1 - \gamma_0)$. These times can be rewritten as

$$T_i^f(\mu_1 - \gamma_0) = \inf\{n \geq 1 : \bar{X}_i(n) + f(n) < \mu_1 - \gamma_0\} = \inf\{n \geq 1 : \bar{X}_i(n) < \mu_1 - \gamma_0 - f(n)\}, \quad (9)$$

which denotes the first boundary-crossing time of the sample mean $\bar{X}_i(n)$ for the boundary $\mu_1 - \gamma_0 - f(n)$. Analyzing this boundary-crossing time is challenging, since the boundary depends on the general (most likely non-linear) function f that satisfies Assumption 1. To address this, we introduce the following lemma.

Lemma 2. *Let $X_1, X_2, \cdots$ be a sequence of i.i.d. random variables with zero mean and $\bar{X}(n) = \frac{1}{n}\sum_{j=1}^{n} X_j$. For any bonus function f satisfying Assumption 1 and any constant $b > 0$, let $n^f(b)$ denote a positive integer such that $f(n) < b$ for all $n \geq n^f(b)$. Then, it holds that*

$$\inf\{n \geq 1 : \bar{X}(n) < b - f(n)\} \leq \inf\{n \geq n^f(b/2) : \bar{X}(n) < b/2\} := \bar{T}^f(b/2) \quad a.s.$$

Proof. To show this conclusion, it suffices to demonstrate that $\bar{X}(n) < b - f(n)$ at time $n = \bar{T}^f(b/2)$. Firstly, since $\bar{T}^f(b/2) \geq n^f(b/2)$, it follows from the definition of $n^f(b/2)$ that $f(\bar{T}^f(b/2)) < b/2$. Next, the definition of $\bar{T}^f(b/2)$ indicates that $\bar{Y}(\bar{T}^f(b/2)) < b/2$. Combining these two statements, we get that $\bar{X}(\bar{T}^f(b/2)) < b/2 = b - b/2 < b - f(\bar{T}^f(b/2))$. The proof is completed. $\square$

Lemma 2 is particularly useful. It shifts the exploration bonus function f from the boundary to the starting point of the boundary-crossing process. As we will show below, this transformation facilitates a more tractable analysis.

4.3.1. Sample Optimality under the Location-Scale Structure We now examine Condition 2 for problems satisfying the location-scale structure in Assumption 2. Define $\bar{Y}_i(n) = \frac{\bar{X}_i(n)-\mu_i}{\sigma_i}$ for each alternative i. From Equation (9), we have that under Assumptions 2 and 4, for some $\gamma_0 \in (0, \gamma)$,

$$\begin{aligned}
T_i^f(\mu_1 - \gamma_0) &= \inf\{n \geq 1 : \bar{X}_i(n) < \mu_1 - \gamma - f(n)\} \\
&= \inf\left\{n \geq 1 : \bar{Y}_i(n) < \frac{\mu_1 - \mu_i - \gamma_0}{\sigma_i} - \frac{f(n)}{\sigma_i}\right\} \quad \text{(by Assumption 2)} \\
&\leq \inf\left\{n \geq 1 : \bar{Y}_i(n) < \frac{\gamma - \gamma_0}{\bar{\sigma}} - \frac{f(n)}{\underline{\sigma}}\right\} \quad \text{(by Assumption 4)} \\
&\leq \inf\left\{n \geq n^{f/\underline{\sigma}}\left(\frac{\gamma - \gamma_0}{2\bar{\sigma}}\right) : \bar{Y}_i(n) < \frac{\gamma - \gamma_0}{2\bar{\sigma}}\right\} \quad \text{(by Lemma 2)} \\
&:= \bar{T}_i^{f/\underline{\sigma}}\left(\frac{\gamma - \gamma_0}{\bar{\sigma}}\right).
\end{aligned} \tag{10}$$

Notice that $\bar{T}_i^{f/\underline{\sigma}}((\gamma - \gamma_0)/\bar{\sigma})$ are i.i.d. across $i = 2, \cdots, k$, since Y_i are i.i.d. across all alternatives. Hence, Equation (10) provides an i.i.d. upper bound for each boundary-crossing time $T_i^f(\mu_1 - \gamma_0)$. Then, we verify Condition 2 by applying the strong law of large numbers (SLLN) to the sum of $\bar{T}_i^{f/\underline{\sigma}}((\gamma - \gamma_0)/\bar{\sigma})$. The following lemma establishes that the expectation of these boundary-crossing times is finite, thereby justifying the use of the SLLN. The proof of the lemma is provided in Section EC.1.2.

Lemma 3. *Suppose that Assumptions 1 and 2 hold. For each alternative $i = 2, \cdots, k$,*

$$\mathrm{E}\left[\bar{T}_i^{f/\underline{\sigma}}\left(\frac{\gamma - \gamma_0}{\bar{\sigma}}\right)\right] \leq C^f(\gamma - \gamma_0, \underline{\sigma}, \bar{\sigma}) < \infty,$$

where $C^f(x, y, z) = n^f\left(\frac{xy}{2z}\right)\exp\left(\frac{2z^2\pi^2}{3x^2 n^f\left(\frac{xy}{2z}\right)}\right)$ for $x, y, z \in (0, \infty)$.

This effectively confirms Condition 2 under the location-scale structure. Consequently, we can readily conclude that the meta-UCB algorithm achieves sample optimality under Assumption 2, which is formally presented in the following theorem.

THEOREM 1. *Suppose that Assumptions 1, 2, and 4 hold. If the total sampling budget $B = ck$ and $c > 2C^f(\gamma, \underline{\sigma}, \bar{\sigma})$, the PCS of the meta-UCB algorithm satisfies*

$$\liminf_{k\to\infty} \mathrm{PCS} \geq \Pr\{U_1^* \geq \mu_1 - \gamma_0\} > 0,$$

where $\gamma_0 = \sup\{\gamma_0 \in (0,\gamma) : 2C^f(\gamma - \gamma_0, \underline{\sigma}, \bar{\sigma}) \leq c\}$.

Theorem 1 confirms the sample optimality of the meta-UCB algorithm, thereby establishing the sample optimality of all UCB algorithms that satisfy Assumption 1 for large-scale pure exploration problems with a location-scale structure. This result demonstrates that UCB algorithms can be sample optimal without relying on Gaussian or sub-Gaussian assumptions, taking an important first step toward moving beyond sub-Gaussianity in large-scale pure exploration. Besides establishing sample optimality, Theorem 1 provides a lower bound for the PCS in the limit $k \to \infty$. This lower bound is defined on the distributional property of the best alternative and a constant γ_0 purely determined by the means and variances of all alternatives, sampling budget, and the exploration bonus function.

4.3.2. Sample Optimality under the Moment Condition We now examine Condition 2 under the moment condition specified in Assumption 3. As before, we use Lemma 2 to obtain upper bounds on the boundary-crossing times $T_i^f(\mu_1 - \gamma_0)$ to justify the use of an SLLN. Specifically, from Equation (9), we have that under Assumption 4, for some $\gamma_0 \in (0, \gamma)$,

$$\begin{aligned} T_i^f(\mu_1 - \gamma_0) &\leq \inf\{n \geq 1 : \bar{X}_i(n) - \mu_i < \gamma - \gamma_0 - f(n)\} \\ &\leq \inf\left\{n \geq n^f\left(\frac{\gamma - \gamma_0}{2}\right) : \bar{X}_i(n) - \mu_i < \frac{\gamma - \gamma_0}{2}\right\} := \hat{T}_i^f(\gamma - \gamma_0), \end{aligned} \tag{11}$$

where the last inequality arises from Lemma 2. While the terms $\{\hat{T}_i^f(\gamma - \gamma_0) : i = 2, \cdots, k\}$ are mutually independent, they are generally *not* identically distributed under Assumption 3. As a result, the ordinary SLLN, which requires only independence and finite mean under identical distributions, is not applicable. To overcome this, we invoke a more general form known as the Kolmogorov's SLLN (Theorem 2.3.10 of Sen and Singer 1994), which is stated in the following lemma.

LEMMA 4 **(Kolmogorov's SLLN).** *Let $T_1, T_2, \ldots$ denote a sequence of independent random variables. If $\sum_{i=1}^k \mathrm{Var}[T_i]/i^2 < \infty$, it holds that as $k \to \infty$,*

$$\frac{1}{k}\sum_{i=1}^k T_i - \frac{1}{k}\sum_{i=1}^k \mathrm{E}[T_i] \to 0 \quad \textit{a.s.}$$

Compared to the ordinary SLLN, Lemma 4 relaxes the requirement of identical distributions among the random variables, but introduces an additional condition on the weighted sum of their variances. To apply the Kolmogorov's SLLN to the sum $\sum_{i=2}^{k} \hat{T}_i^f(\gamma - \gamma_0)$, thereby controlling $\sum_{i=2}^{k} T_i^f(\mu_1 - \gamma_0)$, we need to verify specific properties of the means and variances of each $\hat{T}_i^f(\gamma - \gamma_0)$. To this end, we first establish the following lemma. The proof of this lemma is included in EC.1.3.

LEMMA 5. *Let $X_1, X_2, \cdots$ be a sequence of i.i.d. random variables with zero mean and $\bar{X}(n) = \frac{1}{n}\sum_{j=1}^{n} X_j$. Then, the first boundary-crossing time $T(b; n_0) = \inf\{n \geq n_0 : \bar{X}(n) < b\}$ w.r.t. a fixed boundary $b > 0$ satisfies that for any $n_0 \geq 1$,*

$$\mathrm{E}[T(b;n_0)] \leq \sum_{m=n_0}^{\infty} \mathrm{P}(\bar{X}(m) \geq b) + n_0 \quad \textit{and} \quad \mathrm{Var}[T(b;n_0)] \leq 2\sum_{m=1}^{\infty} m\mathrm{P}(\bar{X}(m+n_0-1) \geq b).$$

Lemma 5 connects the mean and variance of boundary-crossing times to the tail behavior of the associated random variable. Motivated by this result, we extend Nagaev's inequality (Lemma EC.2) to derive concentration bounds based on the moment order q and its upper bound M, as stated in the following lemma. The proof of this lemma is included in EC.1.4.

LEMMA 6. *Let $X_1, X_2, \cdots$ be a sequence of i.i.d. random variables with mean μ and $\bar{X}(n) = \frac{1}{n}\sum_{j=1}^{n} X_j$. If there exists a pair of positive constants $q > 2$ and $M > 0$ such that $\mathrm{E}\left[|X_i|^q\right] \leq M$, it holds that for any $x \geq 0$,*

$$P\left(\bar{X}(n) - \mu \geq x\right) \leq a_1 n^{-q+1} x^{-q} + \exp\left\{-a_2 n x^2\right\},$$

where $a_1 = (2 + 4/q)^q M$ and $a_2 = (q+2)^{-2} e^{-q} M^{-2/q}/2$.

Together, Lemmas 5 and 6 lead to Lemma 7, which provides finite bounds on both the mean and variance of the boundary-crossing times. In the lemma, c_1, c_2, and c_3 are positive constants that depend only on q and M. Their explicit expressions are provided in Section EC.1.5, where the proof of the lemma is also included.

LEMMA 7. *Suppose that Assumptions 1 and 3 hold. For any $\gamma_0 \in (0, \gamma)$, the boundary-crossing time $\hat{T}_i^f(\gamma - \gamma_0)$ of each alternative $i = 2. \cdots, k$ satisfies*

$$\mathrm{E}\left[\hat{T}_i^f(\gamma - \gamma_0)\right] \leq C\left(\frac{\gamma - \gamma_0}{2}; n^f\left(\frac{\gamma - \gamma_0}{2}\right)\right) \textit{ and } \mathrm{Var}\left[\hat{T}_i^f(\gamma - \gamma_0)\right] \leq D\left(\frac{\gamma - \gamma_0}{2}; n^f\left(\frac{\gamma - \gamma_0}{2}\right)\right).$$

where $C(b;n) = c_1 b^{-q} n^{-q+2} + \frac{\exp\left(-nc_2 b^2\right)}{1 - \exp(-c_2 b^2)} + n$ and $D(b;n) = c_3 b^{-q} n^{-q+3} + \frac{2n\exp\left(-nc_2 b^2\right)}{\left(1 - \exp(-c_2 b^2)\right)^2}$.

Lemma 7 provides sufficient conditions for applying Kolmogorov's SLLN to the sum of boundary-crossing times $\sum_{i=2}^{k} \hat{T}_i^f(\gamma - \gamma_0)$, thereby establishing Condition 2 for the meta-UCB algorithm under the moment condition. Specifically, from Lemma 4, we have

$$\limsup_{k\to\infty} \; \frac{1}{k}\sum_{i=2}^{k} T_i^f(\mu_1 - \gamma_0) - C\left(\frac{\gamma - \gamma_0}{2}; n^f\left(\frac{\gamma - \gamma_0}{2}\right)\right) \leq \limsup_{k\to\infty} \frac{1}{k-1}\sum_{i=2}^{k}\left(\hat{T}_i^f(\gamma - \gamma_0) - \mathrm{E}\left[\hat{T}_i^f(\gamma - \gamma_0)\right]\right) = 0 \tag{12}$$

This result confirms Condition 2, and in turn establishes the sample optimality of the meta-UCB algorithm under Assumption 3. In analogy to Theorem 1, we state the following theorem.

THEOREM 2. *Suppose that Assumptions 1, 3, and 4 hold. If the total sampling budget $B = ck$ with $c > 2C(\gamma/2; n^f(\gamma/2))$, the PCS of the meta-UCB algorithm satisfies*

$$\liminf_{k\to\infty} \mathrm{PCS} \geq \Pr\left\{U_1^*(n) \geq \mu_1 - \gamma_0\right\} > 0,$$

where $\gamma_0 = \sup\{\gamma_0 \in (0,\gamma) : 2C((\gamma-\gamma_0)/2; n^f((\gamma-\gamma_0)/2)) \leq c\}$.

Theorem 2 further establishes the sample optimality of the meta-UCB algorithm for large-scale pure exploration problems with generally distributed observations. Compared to Theorem 1, this result accommodates greater distributional heterogeneity by allowing the alternatives to follow different distribution families without requiring a location-scale structure. This added generality, however, comes at the cost of assuming the existence of bounded absolute moments of order of some $q > 3$ for the alternatives.

5. Extensions

Notice that Theorems 1 and 2 are established under the indifference-zone (IZ) formulation in Assumption 4, which requires a constant mean gap between the best alternative and all others, for UCB algorithms whose UCB values satisfy Assumption 1. In this section, we explore extensions beyond this scope. In Section 5.1, we examine the performance of UCB algorithms when the IZ assumption is relaxed. In Section 5.2, we discuss broader classes of UCB values that go beyond Assumption 1.

5.1. Moving Beyond the Indifference-Zone Formulation

The IZ formulation in Assumption 4 may be argued to be impractical for large-scale problems, where the mean differences between the best alternative and some non-best ones may shrink as k grows (Ni et al. 2017). Motivated by this, we examine whether the meta-UCB algorithm can achieve the sample optimality in scenarios where Assumption 4 does not hold. To model such scenarios, we consider a representative configuration in which the mean gap between the best alternative and alternative i, denoted $\Delta_i = \mu_1 - \mu_i$, shrinks at a polynomial rate, namely at the rate of $\mathcal{O}(k^{-\beta})$ for some $\beta > 0$. Inspired by Jamieson et al. (2013), we adopt the following mean configuration:

$$\mu_i = \mu_1 - (i/k)^\beta \quad \text{for } i = 2, \ldots, k. \tag{13}$$

This formulation captures two important features commonly observed in large-scale problems where Assumption 4 fails to hold: (1) the means are spread across a range and become increasingly dense as the number of alternatives grows; (2) the smallest mean gap Δ_2 diminishes as $k \to \infty$. The parameter β intuitively characterizes the hardness of the selection problem: a larger β implies that the non-best means are more tightly clustered near the best mean μ_1, making it more difficult to identify the true best alternative. When $\beta = 0$, the configuration satisfies Assumption 4, recovering the "easiest" IZ formulation.

Under this configuration, for the meta-UCB algorithm to achieve the sample optimality, the following two requirements—similar to those discussed in Section 4.1—serve as sufficient conditions. Intuitively, these two conditions can be regarded as adjustments to Conditions 1 and 2, achieved by enforcing $\gamma = 0$.

CONDITION 3. $\Pr\{U_1^* \geq \mu_1\} > 0$.

CONDITION 4. *There exists a constant* $c > 0$ *such that* $\limsup_{k\to\infty} \frac{1}{k}\sum_{i=2}^{k} T_i^f(\mu_1) \leq c$.

5.1.1. Satisfaction of Condition 3. To ensure that Condition 3 holds, the PCS lower bound of the meta-UCB algorithm must satisfy

$$\Pr\{U_1^*(n) \geq \mu_1\} = \Pr\{\forall n \geq 1 : \mu_1 \leq \bar{X}_1(n) + f(n)\} > 1 - \alpha \tag{14}$$

for some constant $\alpha \in (0,1)$. Importantly, this condition is not automatically guaranteed by Assumption 1. For example, if $f(\cdot) = 0$, the meta-UCB algorithm reduces to a purely greedy algorithm that always samples the alternative with the highest sample mean. In this case, the condition fails to hold, since the sample mean process will almost surely fall below the true mean at some point. Therefore, to satisfy Equation (14), the choice of the bonus function must be made more carefully. Interestingly, Equation (14) essentially requires that $\bar{X}_1(n) + f(n)$ serves as a time-uniform (or anytime) upper confidence sequence for the mean μ_1, as discussed in Howard et al. (2021). Such confidence sequences can be constructed for non-sub-Gaussian observations. Wang and Ramdas (2023) provide a detailed discussion on this topic. Building on the idea of allocating the total error $\alpha = \sum_{n=1}^{\infty} \alpha_n$ across all values of n and applying a union bound, we derive a class of bonus functions that satisfy Assumption 1 and ensure the confidence requirement in the following lemma. The proof is provided in Section EC.1.6.

LEMMA 8. *Suppose that* $\mathrm{E}[|X_1|^q] \leq M$ *for some* $q > 2$ *and* $M < \infty$. *For any* $1 < q' < q - 1$ *and* $\alpha \in (0,1)$, *Equation* (14) *holds with the exploration bonus function*

$$f(n) = \max\left\{ \sqrt[q]{\frac{2a_1 z(q')}{\alpha n^{q-1-q'}}}, \sqrt{\frac{\log(2z(q'+1)) + (q'+1)\log n + \log(1/\alpha)}{a_2 n}} \right\}, \tag{15}$$

where $z(x) = \sum_{n=1}^{\infty} \frac{1}{n^x}$ *is the Riemann zeta function.*

These bonus functions require known bounds on the q-th moment (i.e., a known M), which is a common assumption adopted in Glynn and Juneja (2018) and many other works. From Lemma 8, we observe that if $q' \leq q/2 - 1$ for any $q > 4$, then a bonus function of order $\mathcal{O}(\sqrt{\log n/n})$ is sufficient to maintain a valid confidence sequence. This result is near optimal, as the law of iterated logarithm shows that the sample mean of any random variable with finite variance scales at most as $\mathcal{O}(\sqrt{\log\log n/n})$ asymptotically.

5.1.2. Satisfaction of Condition 4. We now turn to Condition 4, which concerns the behavior of the boundary-crossing times and depends on the choice of the bonus function. For clarity of exposition, we adopt the bonus function in Equation (15) where $q' = 2$, though the arguments can be extended to any $q' > 2$; see Section EC.1.7. For the boundary-crossing time $T_i^f(\mu_1)$ of each non-best alternative $i = 2, \ldots, k$, we have

$$\begin{aligned} T_i^f(\mu_1) = \inf\{n \geq 1 : U_i(n) \leq \mu_1\} &= \inf\{n \geq 1 : \bar{X}_i(n) + f(n) - \mu_i \leq \mu_1 - \mu_i\} \\ &\leq \inf\left\{n \geq n^f\left(\frac{\mu_1 - \mu_i}{2}\right) : \bar{X}_i(n) - \mu_i \leq \frac{\mu_1 - \mu_i}{2}\right\} := \hat{T}_i^f(\Delta_i), \end{aligned} \tag{16}$$

where $\Delta_i := \mu_1 - \mu_i = (i/k)^\beta$ and the last inequality arises from Lemma 2. The next lemma provides properties of the upper bound $\hat{T}_i^f(\Delta_i)$, which allows us to show that Condition 4 can be satisfied. The proof is based on Lemmas 5 and 6 and deferred to Section EC.1.7.

LEMMA 9. *Suppose that Assumption 3 holds. For the exploration bonus function in Equation (15), it holds that for any $\beta < \min\left\{\frac{q-3}{q}, \frac{1}{2}\right\}$,*

$$\sum_{i=2}^{k} \mathrm{E}[\hat{T}_i^f(\Delta_i)] \leq \left[d_0 + \frac{d_1}{1 - \beta q/(q-3)} + \frac{d_2}{(1-2\beta)} + \frac{d_3\beta}{(1-2\beta)^2}\right]k \quad \text{and} \quad \mathrm{Var}[\hat{T}_i^f(\Delta_i)] \leq d_4,$$

where d_0, d_1, d_2, d_3, and d_4 are positive constants depending only on q, M, and α.

With this lemma, by arguments similar to Equation (12), we can apply the Kolmogorov's SLLN (Lemma 4) to show that Condition 4 is indeed satisfied. With Conditions 3 and 4 both satisfied, we obtain the following proposition.

PROPOSITION 1. *Suppose that Assumptions 1 and 3 hold. Then, for the bonus function in Equation (15) and the problem configuration in Equation (13), if $\beta < \min\left\{\frac{q-3}{q}, \frac{1}{2}\right\}$, we have that given a total sampling budget $B = \left[d_0 + \frac{d_1}{1-\beta q/(q-3)} + \frac{d_2}{(1-2\beta)} + \frac{d_3\beta}{(1-2\beta)^2}\right]k$, the PCS of the meta-UCB algorithm satisfies*

$$\liminf_{k\to\infty} \mathrm{PCS} \geq \Pr\{U_1^* \geq \mu_1\} > 0.$$

Proposition 1 is a possibility result—it showcases that the meta-UCB algorithm *can* achieve the sample optimality even without the IZ formulation, for large-scale pure exploration with generally distributed alternatives. Interestingly, it also highlights how the tail behavior of the performance distributions affects the ability to achieve sample optimality. Specifically, the meta-UCB algorithm is not universally sample optimal: the heavier the tails (i.e., smaller q), the smaller the allowable β. Conversely, as q increases (lighter tails), the range of admissible β values expands, allowing for faster-shrinking mean gaps while still maintaining sample optimality. This possibility result, and the accompanying insights, may provide a deeper understanding of the performance of UCB algorithms in large-scale pure exploration problems.

5.2. On the Other UCB Algorithms

So far, we have focused on UCB algorithms whose UCB values take the form of a sample mean plus a deterministic bonus function of the sample size (i.e., those satisfying Assumption 1). However, it is natural to ask whether these results may extend to an even broader class of algorithms to allow greater flexibility in how the UCB values can be constructed. For example, when dealing with heavy-tailed observations under moment conditions like Assumption 3, robust mean estimators such as the truncated sample mean or the weighted sample mean and the associated confidence bounds (or sequences) have proven particularly useful (Bubeck et al. 2013, Glynn and Juneja 2018, Wang and Ramdas 2023). In addition, auxiliary sample statistics such as the sample variance can also be incorporated into UCB value constructions (Audibert et al. 2009).

Although our current results do not cover these variants, we conjecture that sample optimality may still hold for a broader class of UCB algorithms, provided that their UCB value construction satisfies a form of *decoupling*—that is, the UCB value for each alternative depends only on its own sample size and sample observations. Under the decoupling condition, the boundary-crossing perspective developed in Section 3.1 remains applicable. Specifically, one can still define a UCB value process $\{U_i(n) : n = 1, 2, \dots\}$ for each alternative i that is independent of that of all other alternatives. This enables the definition of a distribution-free boundary U_1^* for the best alternative, along with corresponding boundary-crossing times $T_i^f(U_1^*)$ for the others. While the sample optimality analysis would then follow a similar structure as in Section 4, verifying the conditions for specific UCB algorithms and distributional assumptions will likely require case-by-case analysis, which we leave for future work.

We also note that there exists another major class of UCB algorithms that do not satisfy the decoupling condition. For example, many algorithms construct UCB values using global statistics, such as the total sample size across all alternatives. A prominent example is the classical UCB1 algorithm of Auer et al. (2002). In such cases, the boundary-crossing framework may break down. Interestingly, our experiments in Section 6 show that UCB1 behaves quite differently and may fail to achieve sample optimality. This contrast highlights a fundamental distinction between the two categories of UCB algorithms in pure exploration.

6. Numerical Experiments

In this section, we conduct numerical experiments to verify our theoretical results. Specifically, in Section 6.1, we introduce the problem configurations, experimental settings, and the UCB algorithms under evaluation. In Section 6.2, we demonstrate the sample optimality of these algorithms under the indifference-zone (IZ) formulation. Section 6.3 compares the budget allocation behaviors of different UCB algorithms. Finally, in Section 6.4, we show that sample optimality can also be achieved beyond the IZ formulation.

6.1. Problem Configurations, Experimental Settings, and Tested UCB Algorithms

6.1.1. Problem Configurations A problem configuration in pure exploration specifies the distributional characteristics of all alternatives and serves as a tool for probing the behavior of a given algorithm. To validate our theoretical results, we consider configurations that allow us to examine two key aspects: (1) how the distribution of the true mean values—e.g., with or without an IZ parameter—affects algorithm performance, and (2) how the algorithm behaves under different distributional families. To meet these goals, we adopt a mean-shifting model to generate problem configurations. We begin with a basic mean-shifting model, in which we first fix the base distribution of a reference random variable X_0, and then generate each alternative's performance distribution by shifting the mean as follows:

$$X_i \stackrel{d}{=} \begin{cases} X_0, & \text{if } i = 1, \\ \\ X_0 - \gamma - \lambda(i/k)^{\beta}, & \text{if } i = 2, \ldots, k. \end{cases} \tag{17}$$

where $\stackrel{d}{=}$ denotes equality in distribution. Here, γ is the IZ parameter, λ controls the overall spacing of the non-best means, and β adjusts how quickly the inferior means decay as k increases. Notice that $X_1, \ldots, X_k$ are assumed to be mutually independent. This flexible construction allows us to simulate various scenarios. Parameter values are scenario-specific and described in the relevant subsections. In our experiments, we primarily consider the following two widely studied mean configurations:

- the Slippage Configuration (SC): Corresponds to $\gamma > 0$ and $\lambda = 0$. This configuration represents a worst-case scenario among all mean structures with the same IZ parameter $\gamma > 0$.
- the Monotonic Means Configuration (MM): Corresponds to $\lambda > 0$ and $\beta > 0$. This setting is commonly used to test algorithms in more realistic scenarios than SC.

Under both the SC and MM configurations, we specify the distributional form of each alternative to align with the theoretical conditions in Sections 4.3.1 and 4.3.2. Specifically, under the location-scale structure assumed in Assumption 2 (Section 4.3.1), the structure of the mean-shifting model allows us to specify only the distribution of a base variable X_0, from which all alternatives are derived. We consider three representative non-sub-Gaussian distributions for X_0: (1) the Lognormal distribution, parameterized by log-mean μ and log-standard deviation σ, with all moments being finite; (2) the Student's t distribution, parameterized by degree of freedom (df) and scale, which has finite qth moment only when q is less than df; and (3) the Pareto distribution, parameterized by shape and scale, which has finite qth moment only when q is smaller than the shape parameter. This leads to six configurations: SC-Lognormal, SC-Student's t, SC-Pareto, MM-Lognormal, MM-Student's t, and MM-Pareto. The specific parameter values used in these configurations are reported in Section 6.2.

To evaluate the performance of UCB algorithms under non-location-scale settings, as modeled in Assumption 3 (Section 4.3.2), we adopt the mixed-distribution configuration approach proposed by Lee and Nelson (2016). We extend the mean-shifting model in Equation (17) to involve two distinct base random variables, X_{odd} and X_{even}, and define the alternative distributions as

$$X_i \overset{d}{=} \begin{cases} X_{\text{odd}}, & \text{if } i = 1, \\ X_{\text{odd}} - \gamma - \lambda(i/k)^{\beta}, & \text{if } i \geq 2 \text{ and } i \text{ is odd}, \\ X_{\text{even}} - \gamma - \lambda(i/k)^{\beta}, & \text{if } i \geq 2 \text{ and } i \text{ is even}. \end{cases} \tag{18}$$

By selecting different distributions for X_{odd} and X_{even}, we construct problem instances that deliberately violate the location-scale structure. In particular, we consider the following two mixed-distribution configurations under both SC and MM configurations: (1) Odd(t)-Even(Pareto), where X_{odd} is Student's t and X_{even} is Pareto; (2) Odd(Pareto)-Even(t), where X_{odd} is Pareto and X_{even} is Student's t. The parameters of these distributions are selected to satisfy the moment conditions in Assumption 3. The specific parameter values used in the experiments are listed in Section 6.2.

6.1.2. Tested UCB Algorithms and Experimental Settings We have analyzed and proved the sample optimality for all UCB algorithms that satisfy Assumption 1. In our experiments, we evaluate the performance of several representative algorithms from this class: (1) UCBE (Audibert et al. 2010), where the UCB value is defined in Example 1; (2) MOSS (Audibert and Bubeck 2010), where the UCB value is defined in Example 2; (3) Greedy, which uses no bonus function (i.e., $f(n_i) = 0$) and always samples the alternative with the highest sample mean. Beyond this class, we also test the famous UCB1 algorithm of Auer et al. (2002), which does not satisfy Assumption 1. Its UCB value is defined as

$$U_i(n_i) = \bar{X}_i(n_i) + \sqrt{\frac{2\log n}{n_i}}, \tag{19}$$

where $n = \sum_{i=1}^{k} n_i$ is the overall sample size of all alternatives. Although UCB1 is structurally similar to UCBE and MOSS, we will see that it may behave very differently due to its coupling across alternative through the global sample size n. Since our focus is not on algorithm tuning, we use fixed algorithm parameters (e.g., $a = 1$ for UCBE) across all experiments.

Unless otherwise specified, we evaluate algorithm performance across a range of problem sizes by varying the number of alternatives $k \in \{2^5, 2^6, \ldots, 2^{15}\}$. The total sampling budget is set as $B = ck$ for each k, where the constant c may vary by configuration. For each algorithm, we estimate the PCS in solving any problem instance based on 500 independent macro replications.

6.2. Sample Optimality under the Indifference-Zone Formulation

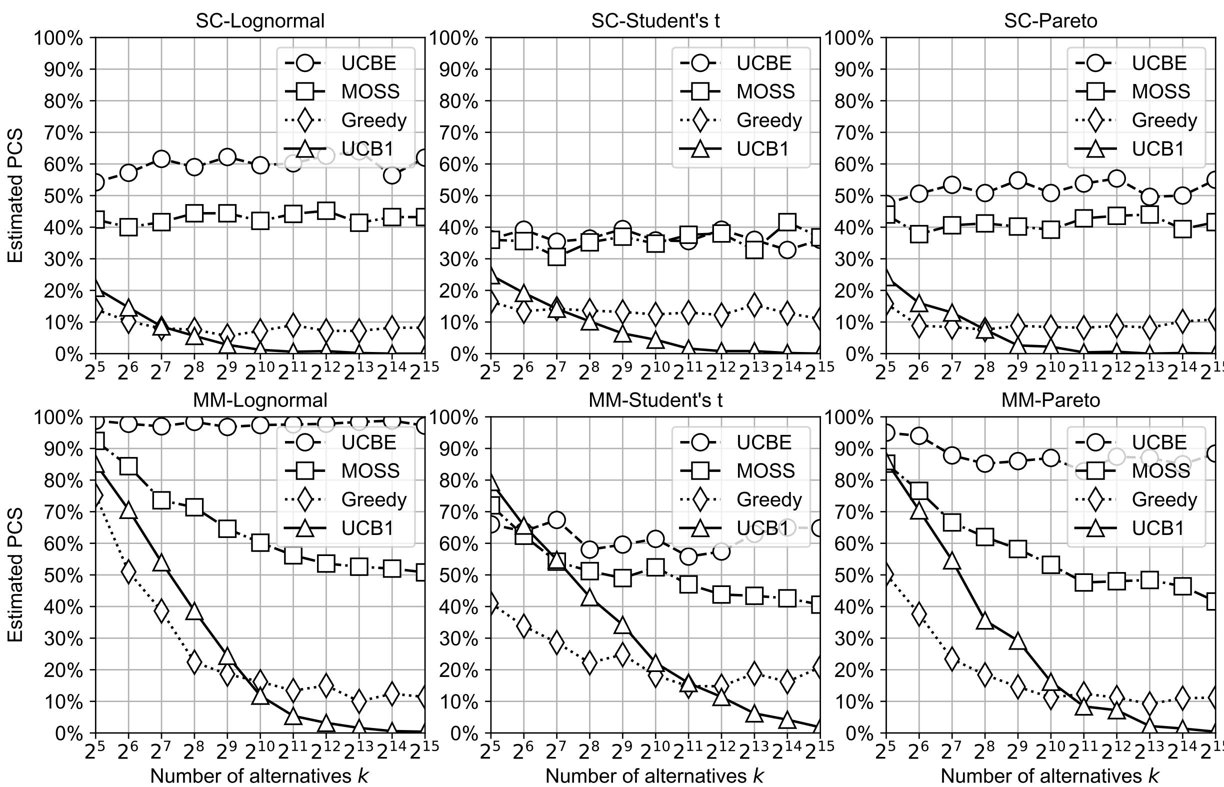
In this subsection, we validate the sample optimality of UCB algorithms under the indifference-zone (IZ) formulation. For all SC and MM configurations, we set the IZ parameter $\gamma = 0.1$. In SC configurations, we let $\lambda = 0$, and the parameter β is irrelevant. In MM configurations, we set $\lambda = 1$ and $\beta = 1$. To reflect different levels of problem difficulty, we use a total sampling budget of $B = 100k$ for each k under SC configurations and $B = 30k$ for MM configurations. For the distributional parameters of the reference random variables used in the mean-shifting model (i.e., X_0, X_{odd}, and X_{even}), we select values such that the variances are approximately 1, and the conditions in Assumption 2 (bounded variance) and Assumption 3 (bounded absolute moment of order $q > 3$) are satisfied. The distributional information of the reference random variables are summarized in Table 1.

Table 1 Distributional Parameters and Moment Conditions of the Reference Random Variables

Structure	Random Variable	Distribution Type	Parameters	q-th Moments Exist for
Location-scale (Assumption 2)	X_0	Lognormal	$\mu = -2, \sigma = 1.45$	$q < \infty$
	X_0	Student's t	df $= 3$, scale $= 0.6$	$q < 3$
	X_0	Pareto	shape $= 3$, scale $= 1.2$	$q < 3$
Mix-Distribution (Assumption 3)	X_{odd} or X_{even}	Student's t	df $= 4$, scale $= 0.7$	$q < 4$
	X_{odd} or X_{even}	Pareto	shape $= 4$, scale $= 2.1$	$q < 4$

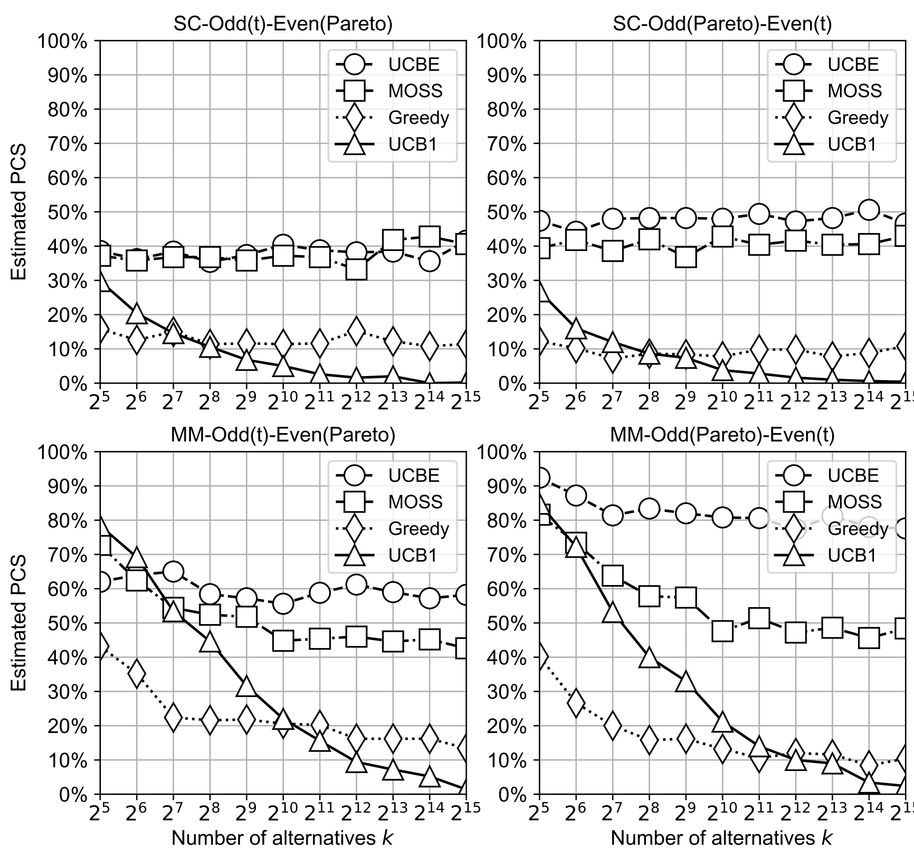
6.2.1. Location-Scale Structure We first examine the performance of UCB algorithms when the alternatives' distributions follow the location-scale structure. Figure 2 plots the PCS of each algorithm against k under the SC and MM configurations. Several key observations can be made. First, under all configurations, UCBE, MOSS, and Greedy demonstrate sample optimality. For instance, under the SC-Lognormal configuration, the PCS of UCBE remains above 50% as k increases. Although the PCS of Greedy is low when k is large, it does not approach zero, indicating that it also achieves the sample optimality. Second, both UCBE and MOSS significantly outperform Greedy, which has no bonus function. This illustrates the exploration effect of incorporating a bonus function satisfying Assumption 1 in UCB algorithms. Interestingly, UCBE generally performs better than MOSS. This is possibly due to MOSS's truncation structure, which caps the bonus and limits exploration for every alternative when its sample size exceeds c. It is also interesting to observe that the PCS levels may vary significantly across distribution families. In general, UCBE and MOSS perform better under Lognormal and Pareto configurations than under Student's t configurations. We explore this phenomenon further in the following experiment. Lastly, and notably, the UCB1 algorithm is not sample optimal under any of the tested configurations. As k increases, its PCS steadily declines toward zero. We analyze this failure and its connection to UCB1's structure in Section 6.3.

Figure 2 PCS of Different UCB Algorithms under Lognormal, Student's t, and Pareto Configurations.



6.2.2. Moment-based Condition We then examine the performance of UCB algorithms when the location-scale structure does not hold, but each alternative satisfies the moment condition of having a bounded absolute moment of order $q > 3$. Specifically, we consider the mixed-distribution configurations. Figure 3 plots the PCS of each algorithm against k under the SC and MM configurations with mixed distributions. Findings are consistent with those observed in Figure 2. Most notably, UCBE, MOSS, and Greedy all continue to demonstrate sample optimality, whereas UCB1 again fails to do so, with its PCS decaying to zero as k increases. Interestingly, we also observe that UCBE and MOSS perform better under the Odd(P)-Even(t) configurations than under the Odd(t)-Even(Pareto) configurations. This performance difference may be attributed to the distribution of the best alternative. Under the mean-shifting model in Equation (18), the key distinction between the two scenarios lies in whether the best alternative follows a Student's t or a Pareto distribution. The Student's t distribution is symmetric with support over the entire real line, while the Pareto distribution is one-sided with support on (scale, ∞). As a result, observations from the Student's t-distribution may have a higher probability of taking very low values. This can make it harder for the best alternative to be re-selected after unfavorable initial observations, even with an exploration bonus. This insight may also help explain the trend observed in Figure 2, where algorithms tend to perform better under Lognormal (also one-sided) and Pareto configurations.

Figure 3 PCS of Different UCB Algorithms under the Mixed-Distribution Configurations.



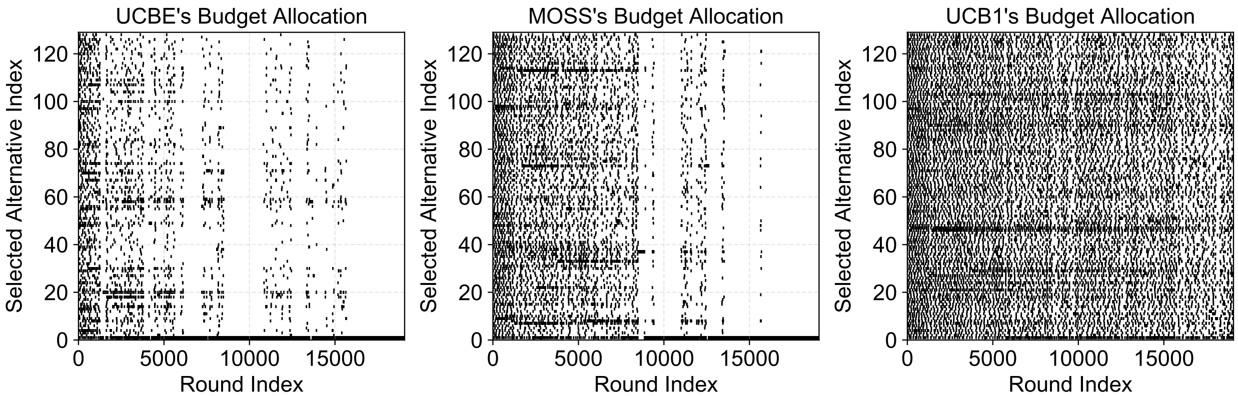
6.3. Budget Allocation Behaviors

In this subsection, we compare the budget allocation behaviors of UCBE, MOSS, and UCB1, with the goal of gaining insights into why UCBE and MOSS achieve the sample optimality, while UCB1 does not. To illustrate these allocation dynamics, we consider a relatively small-scale problem with $k = 128$ alternatives under the SC-Lognormal configuration. We visualize the sample paths of all three algorithms under a common random seed and a total sampling budget of $B = 150k$, where all algorithms correctly identify the best alternative within that sample path. These sample paths are shown in Figure 4, where every sampling round is marked by a dark dot indicating the alternative selected in that round. To further illustrate the allocation behavior, we also plot the histogram of the allocated sample sizes of all *inferior* alternatives ($i = 2, \cdots, 128$) at the end of the sampling process in Figure 5.

Figure 4 and Figure 5 show that the sample-optimal UCBE and MOSS algorithms exhibit similar sample allocation behaviors, whereas the non-sample-optimal UCB1 algorithm behaves quite differently. For UCBE and MOSS, we observe the boundary-crossing dynamics described in Section 3.1. After certain rounds, once the boundary-crossing processes of the inferior alternatives are completed, the algorithms focus almost exclusively on the best alternative. This transition may repeat several times. During the process that the UCB value of the best alternative drops to its minimum, some non-best alternatives may temporarily surpass it and get resampled to complete their boundary-crossing. Eventually, the boundary-crossing event described in Section 3.1 occurs, after which only the best alternative continues to be sampled. Furthermore, many inferior alternatives are sampled only a few times, as illustrated in the histogram in Figure 5. We believe this selective sampling behavior is key to achieving sample optimality under suitable distributional conditions. In contrast, UCB1 continues to allocate samples to most (if not all) inferior alternatives until the total sampling budget is exhausted, as shown in Figure 4. This results in a more uniform allocation across alternatives, as illustrated in the histogram in Figure 5. This behavior is somewhat surprising, as it reveals that UCB1—designed to strike an optimal balance between exploration *and* exploitation in online learning—may actually engage in *more* exploration than pure-exploration algorithms like UCBE in large-scale problems.

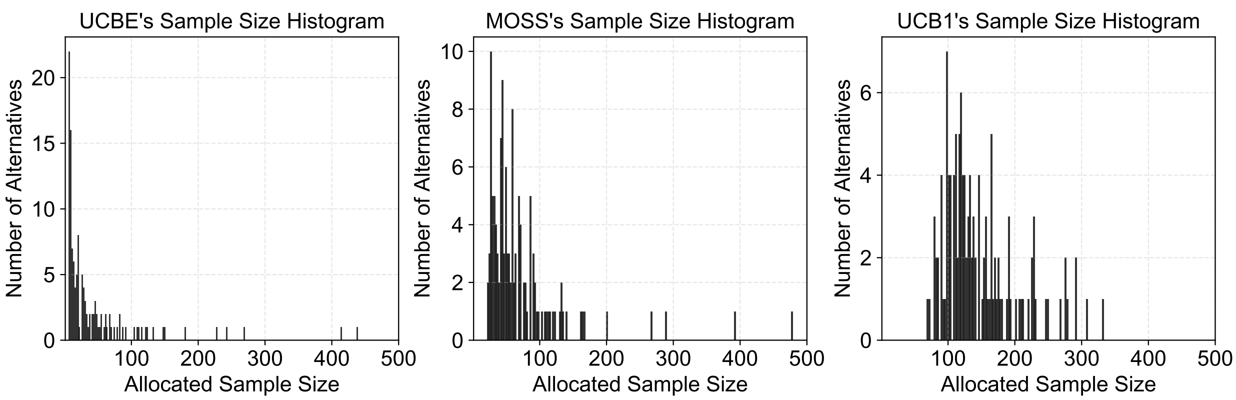
This interesting behavior of UCB1 appears to stem from the global sample size term $n = \sum_{i=1}^{k} n_i$ in the UCB value definition (Equation (19)), which introduces coupling across alternatives. Because this term grows regardless of which alternative is selected, it causes the UCB value of all alternatives—including clearly inferior ones—to increase over time, potentially leading to unnecessary resampling. Such behavior may become particularly wasteful when the number of alternatives is large, ultimately causing UCB1 to lose the sample optimality. This contrast highlights the importance of the decoupling structure inherent in bonus functions that satisfy Assumption 1 for solving large-scale pure exploration problems.

Figure 4 Budget Allocation in a Sample Path of Different UCB Algorithms for a Problem with $k = 128$.



Note. Each round is marked by a dark dot indicating the alternative selected in that round.

Figure 5 Allocated Sample Sizes of Inferior Alternatives for a Problem with $k = 128$.



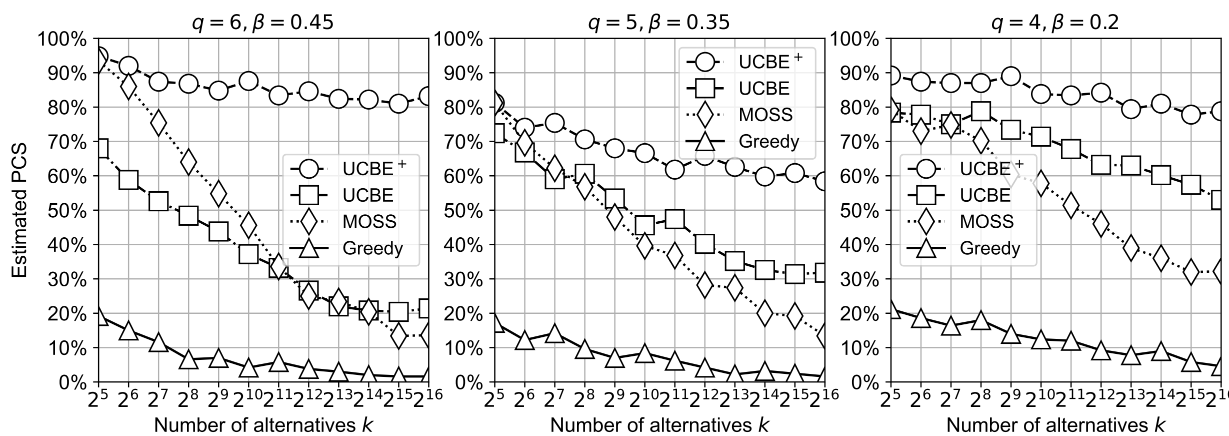
6.4. Beyond the Indifference-Zone Formulation

Lastly, we investigate the sample optimality of UCB algorithms beyond the commonly studied IZ formulation. We focus on the MM-Pareto configuration, which enables us to finely control the shape parameter of the Pareto distribution to construct scenarios with varying moment conditions. Specifically, to ensure that the Pareto distribution has a finite qth absolute moment, we set its shape parameter to $q + \varepsilon$, where $\varepsilon = 0.1$. The scale parameter is then chosen so that the resulting distribution has approximately unit variance for comparability. Under the configuration, we set $\gamma = 0$ to represent a non-IZ setting and let $\lambda = 1/4$. The parameter β is varied to induce different rates of mean gap shrinkage as follows. We consider three moment conditions by setting $q = 6$, 5, and 4, leading to increasingly heavy-tailed configurations. For each q, Proposition 1 suggests that to ensure sample optimality, the decay rate β must satisfy $\beta < \min\{(q-3)/q, 1/2\}$. Accordingly, we set $\beta = \min\{(q-3)/q, 1/2\} - 0.05$ in each case. This yields three distinct configurations, which we refer to by their corresponding (q, β) values. Then, according to Proposition 1, we set the total sampling budget $B = 10\left[\frac{1}{1-\beta q/(q-3)} + \frac{1}{1-2\beta} + \frac{\beta}{(1-2\beta)^2}\right]k$ for each (q,β) configuration. Furthermore, as shown in Lemma 8, sample optimality in non-IZ settings may require stronger conditions on the exploration bonus function, depending on the value of q. Motivated by Lemma 8 (with $q' = 2$), we construct a modified version of the UCBE algorithm, referred to as UCBE$^+$, that incorporates an adjusted bonus function instance

$$f(n_i) = \begin{cases} n_i^{\frac{3-q}{q}}, & \text{if } q < 6, \\ \sqrt{\frac{\log(n_i+2)}{n_i}}, & \text{if } q \geq 6. \end{cases}$$

We avoid using the bonus function form in Equation (15) due to its looseness for ensuring a target PCS level. Then, we compare UCBE$^+$, UCBE, MOSS, and Greedy under all three (q,β) configurations. The PCS of each algorithm across different k under the three configurations is plotted in Figure 6.

Figure 6 PCS of Different UCB Algorithms Under Non-IZ Configurations.



From Figure 6, several key observations can be made. First, UCBE$^+$ appears capable of maintaining sample optimality under varying moment conditions. When $q = 6$ or $q = 4$, its PCS stabilizes around 80%. For $q = 5$, the PCS stabilizes at a lower level, around 60%. This dip may be due to that the effect of using a smaller total budget (corresponding to a smaller β) outweighs the benefit of a wider mean gap. Second, the Greedy algorithm is no longer sample optimal. As k increases, its PCS consistently declines and may eventually approach zero. This result confirms the necessity of Condition 3 discussed following Equation 14—namely, that Greedy fails to satisfy the confidence sequence condition and therefore cannot guarantee sample optimality in non-IZ settings. This result highlights the increased challenge of achieving sample optimality when the IZ assumption is removed. Lastly, we observe that both UCBE and MOSS achieve significantly lower PCS compared to UCBE$^+$. From their PCS curves, it is difficult to conclude whether they are sample optimal or not. This suggests that their bonus functions may provide insufficient exploration under heavy-tailed, non-IZ configurations, compared to the moment-adjusted bonus used in UCBE$^+$. To summarize, these findings show that UCB algorithms equipped with adjusted bonus functions that satisfy Assumption 1 and Condition 3 can still achieve the sample optimality under the more challenging heavy-tailed, non-IZ settings for large-scale pure exploration.

7. Concluding Remarks

In this paper, we explore the performance of UCB algorithms in large-scale pure exploration problems under non-sub-Gaussian distributional assumptions. We introduce a meta-UCB algorithm with a general UCB value structure and develop a unified analysis framework based on a boundary-crossing perspective. This leads to a distribution-free lower bound on the PCS, which provides structural insights into the role of the exploration bonus function in balancing exploration and selection accuracy. Building on this PCS bound, we show that the meta-UCB algorithm achieves the sample optimality under the indifference-zone (IZ) formulation, either when the performance distributions follow a location-scale structure and have bounded variance, or when only a bounded moment of order $q > 3$ is assumed. We further extend the analysis to non-IZ scenarios and conjecture that the sample optimality can also hold for a broader class of UCB algorithms that satisfy a decoupling condition. Numerical experiments support these theoretical results and further demonstrate the robustness of the UCB approach.

This paper opens several directions for future research. On the theoretical side, preliminary numerical experiments suggest that UCB algorithms may remain sample optimal even when only low-order moments (e.g., $q \leq 2$) are finite, raising the question of whether the moment condition $q > 3$ can be further relaxed. It is also of interest to extend our results to a broader class of UCB algorithms with decoupled UCB values, as discussed in Section 5.2, which are not covered by our current framework. Additionally, we have not discussed the issue of consistency, which requires that the PCS converges to 1 as the total sampling budget grows faster than the number of alternatives. This may be achieved by incorporating an explore-first framework (Li et al. 2025) into UCB algorithms. Furthermore, from a practical perspective, the algorithm selection problem—i.e., identifying which UCB algorithms perform best under specific problem settings—remains an open question. Finally, as UCB algorithms are inherently sequential, developing parallel implementations for improved computational efficiency presents an important venue for future work.

References

Abry P, Borgnat P, Ricciato F, Scherrer A, Veitch D (2010) Revisiting an old friend: on the observability of the relation between long range dependence and heavy tail. *Telecommunication Systems* 43:147–165.

Agrawal S (2019) Recent advances in multiarmed bandits for sequential decision making. *Operations Research & Management Science in the Age of Analytics* 167–188.

Agrawal S, Juneja S, Glynn P (2020) Optimal δ-correct best-arm selection for heavy-tailed distributions. *Algorithmic Learning Theory*, 61–110 (PMLR).

Agrawal S, Koolen WM, Juneja S (2021) Optimal best-arm identification methods for tail-risk measures. *Advances in Neural Information Processing Systems* 34:25578–25590.

Andradóttir S, Prudius AA (2009) Balanced explorative and exploitative search with estimation for simulation optimization. *INFORMS Journal on Computing* 21(2):193–208.

Audibert JY, Bubeck S (2009) Minimax policies for adversarial and stochastic bandits. *COLT*, 217–226.

Audibert JY, Bubeck S (2010) Regret bounds and minimax policies under partial monitoring. *The Journal of Machine Learning Research* 11:2785–2836.

Audibert JY, Bubeck S, Munos R (2010) Best arm identification in multi-armed bandits. *COLT*, 41–53.

Audibert JY, Munos R, Szepesvári C (2009) Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science* 410(19):1876–1902.

Auer P, Cesa-Bianchi N, Fischer P (2002) Finite-time analysis of the multiarmed bandit problem. *Machine learning* 47:235–256.

Bayati M, Hamidi N, Johari R, Khosravi K (2020) Unreasonable effectiveness of greedy algorithms in multi-armed bandit with many arms. *Advances in Neural Information Processing Systems* 33:1713–1723.

Bechhofer RE (1954) A single-sample multiple decision procedure for ranking means of normal populations with known variances. *The Annals of Mathematical Statistics* 16–39.

Blanchet J, Liu J, Zwart B (2008) Large deviations perspective on ordinal optimization of heavy-tailed systems. *2008 Winter Simulation Conference*, 489–494 (IEEE).

Broadie M, Han M, Zeevi A (2007) Implications of heavy tails on simulation-based ordinal optimization. *2007 Winter Simulation Conference*, 439–447 (IEEE).

Bubeck S, Cesa-Bianchi N, Lugosi G (2013) Bandits with heavy tail. *IEEE Transactions on Information Theory* 59(11):7711–7717.

Chan HP, Hu S (2019) Optimal UCB adjustments for large arm sizes. *arXiv preprint arXiv:1909.02229* .

Chen CH, Lee LH (2011) *Stochastic simulation optimization: an optimal computing budget allocation*, volume 1 (World scientific).

Chen EJ (2011) An enhanced lognormal selection procedure. *Discrete Event Dynamic Systems* 21:205–218.

Chick SE, Inoue K (2001) New two-stage and sequential procedures for selecting the best simulated system. *Operations Research* 49(5):732–743.

Degenne R, Nedelec T, Calauzènes C, Perchet V (2019) Bridging the gap between regret minimization and best arm identification, with application to A/B tests. *The 22nd International Conference on Artificial Intelligence and Statistics*, 1988–1996 (PMLR).

Dong J, Zhu Y (2016) Three asymptotic regimes for ranking and selection with general sample distributions. *2016 Winter Simulation Conference (WSC)*, 277–288 (IEEE).

Eckman DJ, Henderson SG (2018) Guarantees on the probability of good selection. *2018 Winter Simulation Conference (WSC)*, 351–365 (IEEE).

Fan L, Glynn PW (2024) The fragility of optimized bandit algorithms. *Operations Research* .

Fan W, Hong LJ, Nelson BL (2016) Indifference-zone-free selection of the best. *Operations Research* 64(6):1499–1514.

Frazier PI, Powell WB, Dayanik S (2008) A knowledge-gradient policy for sequential information collection. *SIAM Journal on Control and Optimization* 47(5):2410–2439.

Fu MC (2016) AlphaGo and Monte Carlo tree search: the simulation optimization perspective. *2016 Winter Simulation Conference (WSC)*, 659–670 (IEEE).

Gao F, Gao S (2016) Optimal computing budget allocation with exponential underlying distribution. *2016 Winter Simulation Conference (WSC)*, 682–689 (IEEE).

Gao S, Chen W, Shi L (2017) A new budget allocation framework for the expected opportunity cost. *Operations Research* 65(3):787–803.

Garivier A, Kaufmann E (2016) Optimal best arm identification with fixed confidence. *Conference on Learning Theory*, 998–1027 (PMLR).

Glynn P, Juneja S (2004) A large deviations perspective on ordinal optimization. *Proceedings of the 36th Winter Simulation Conference (WSC)*, volume 1, 577–585 (IEEE).

Glynn P, Juneja S (2018) Selecting the best system and multi-armed bandits. *arXiv preprint arXiv:1507.04564* .

Hong LJ, Fan W, Luo J (2021) Review on ranking and selection: A new perspective. *Frontiers of Engineering Management* 8(3):321–343.

Hong LJ, Jiang G, Zhong Y (2022) Solving large-scale fixed-budget ranking and selection problems. *INFORMS Journal on Computing* 34(6):2930–2949.

Howard SR, Ramdas A, McAuliffe J, Sekhon J (2021) Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics* 49(2):1055–1080.

Jamieson K, Malloy M, Nowak R, Bubeck S (2013) On finding the largest mean among many. *arXiv preprint arXiv:1306.3917* .

Jamieson K, Malloy M, Nowak R, Bubeck S (2014) lil'UCB: An optimal exploration algorithm for multi-armed bandits. *Conference on Learning Theory*, 423–439 (PMLR).

Jamieson K, Nowak R (2014) Best-arm identification algorithms for multi-armed bandits in the fixed confidence setting. *2014 48th Annual Conference on Information Sciences and Systems (CISS)*, 1–6 (IEEE).

John TT, Chen P (2006) Lognormal selection with applications to lifetime data. *IEEE Transactions on reliability* 55(1):135–148.

Karnin Z, Koren T, Somekh O (2013) Almost optimal exploration in multi-armed bandits. *International Conference on Machine Learning*, 1238–1246 (PMLR).

Kim SH, Nelson BL (2006) Selecting the best system. *Handbooks in Operations Research and Management Science*, volume 13, 501–534 (Elsevier).

Lai TL (1987) Adaptive treatment allocation and the multi-armed bandit problem. *The Annals of Statistics* 15(3):1091–1114.

Lattimore T, Szepesvári C (2020) *Bandit algorithms* (Cambridge University Press).

Lee S, Nelson BL (2016) General-purpose ranking and selection for computer simulation. *IIE Transactions* 48(6):555–564.

Li Z, Fan W, Hong LJ (2025) The (surprising) sample optimality of greedy procedures for large-scale ranking and selection. *Management Science* 71(2):1238–1259.

Luo J, Hong LJ, Nelson BL, Wu Y (2015) Fully sequential procedures for large-scale ranking-and-selection problems in parallel computing environments. *Operations Research* 63(5):1177–1194.

Meyer J (1987) Two-moment decision models and expected utility maximization. *The American Economic Review* 421–430.

Nelson BL, Swann J, Goldsman D, Song W (2001) Simple procedures for selecting the best simulated system when the number of alternatives is large. *Operations Research* 49(6):950–963.

Ni EC, Ciocan DF, Henderson SG, Hunter SR (2017) Efficient ranking and selection in parallel computing environments. *Operations Research* 65(3):821–836.

Panahi H (2016) Model selection test for the heavy-tailed distributions under censored samples with application in financial data. *International Journal of Financial Studies* 4(4):24.

Pei L, Nelson BL, Hunter SR (2024) Parallel adaptive survivor selection. *Operations Research* 72(1):336–354.

Pryzant R, Iter D, Li J, Lee YT, Zhu C, Zeng M (2023) Automatic prompt optimization with "gradient descent" and beam search. *arXiv preprint arXiv:2305.03495* .

Qin C, You W (2025) Dual-directed algorithm design for efficient pure exploration. *Operations Research* .

Russo D (2020) Simple Bayesian algorithms for best-arm identification. *Operations research* 68(6):1625–1647.

Sen PK, Singer JM (1994) *Large sample methods in statistics: An introduction with applications* (Chapman & Hall).

Shin D, Broadie M, Zeevi A (2018) Tractable sampling strategies for ordinal optimization. *Operations Research* 66(6):1693–1712.

Strum DP, May JH, Vargas LG (2000) Modeling the uncertainty of surgical procedure times: comparison of log-normal and normal models. *The Journal of the American Society of Anesthesiologists* 92(4):1160–1167.

Tanczos E, Nowak R, Mankoff B (2017) A KL-LUCB algorithm for large-scale crowdsourcing. *Advances in Neural Information Processing Systems* 30:5896–5905.

Terayama K, Iwata H, Araki M, Okuno Y, Tsuda K (2018) Machine learning accelerates MD-based binding pose prediction between ligands and proteins. *Bioinformatics* 34(5):770–778.

Toscano-Palmerin S, Frazier PI (2015) Asymptotic validity of the bayes-inspired indifference zone procedure: The non-normal known variance case. *2015 Winter Simulation Conference (WSC)*, 3868–3879 (IEEE).

Wang H, Ramdas A (2023) Catoni-style confidence sequences for heavy-tailed mean estimation. *Stochastic Processes and Their Applications* 163:168–202.

Yu X, Shao H, Lyu MR, King I (2018) Pure exploration of multi-armed bandits with heavy-tailed payoffs. *UAI*, 937–946.

Zhang G, Li H, Peng Y (2020) Sequential sampling for a ranking and selection problem with exponential sampling distributions. *2020 Winter Simulation Conference (WSC)*, 2984–2995 (IEEE).

Zhang Q, Ying L (2023) Fast and regret optimal best arm identification: Fundamental limits and low-complexity algorithms. *Advances in Neural Information Processing Systems* 36:16729–16769.

Zhang Z, Peng Y (2024) Sample-efficient "clustering and conquer" procedures for parallel large-scale ranking and selection. *arXiv preprint arXiv:2402.02196* .

Zhong Y, Hong LJ (2022) Knockout-tournament procedures for large-scale ranking and selection in parallel computing environments. *Operations Research* 70(1):432–453.

Zhong Z, Cheung WC, Tan VY (2021) Achieving the pareto frontier of regret minimization and best arm identification in multi-armed bandits. *arXiv preprint arXiv:2110.08627* .

Zhou J, Ryzhov IO (2025) A new rate-optimal sampling allocation for linear belief models. *Operations Research* 73(1):239–250.

E-companion to

UCB for Large-Scale Pure Exploration: Beyond Sub-Gaussianity

EC.1. Proofs

To complete the desired proofs, we first prepare the following lemma, which follows directly from Corollaries 8.39 and 8.44 of Siegmund (1985).

LEMMA EC.1. *Let $X_1, X_2, \cdots$ denote a sequence of i.i.d. random variables. Define $\bar{X}(n) = n^{-1}\sum_{i=1}^{n} X_i$ and $\tau = \inf\{n \geq 1 : \bar{X}(n) < b\}$ for some constant b. Then, it holds that*

$$\Pr\{\tau = \infty\} = \exp\left(-\sum_{n=1}^{\infty} n^{-1}\Pr\{\bar{X}(n) < b\}\right) \quad \text{and} \quad \mathrm{E}[\tau] = \exp\left(\sum_{n=1}^{\infty} n^{-1}\Pr\{\bar{X}(n) \geq b\}\right).$$

EC.1.1. Proof of Lemma 1

Proof. According to the definition of U_1^*, we have

$$\begin{aligned}\Pr\{U_1^* \geq \mu_1 - \gamma_0\} &= \Pr\{\forall n \geq 1, \bar{X}_1(n) \geq \mu_1 - \gamma_0 - f(n)\} \\ &\geq \Pr\{\forall n \geq 1, \bar{X}_1(n) \geq \mu_1 - \gamma_0\}, \end{aligned} \tag{EC.1}$$

where the inequality arises because of Assumption 1. Let $\tau = \inf\{n \geq 1 : \bar{X}_1(n) - \mu_1 < -\gamma_0\}$. Then an alternative expression of Equation (EC.1) can be derived:

$$\Pr\{\forall n \geq 1, \bar{X}_1(n) \geq \mu_1 - \gamma_0\} = \Pr\{\tau = \infty\}.$$

Further, by Lemma EC.1, we may have that as long as $\gamma_0 > 0$,

$$\begin{aligned}\Pr\{\forall n \geq 1, \bar{X}_1(n) \geq \mu_1 - \gamma_0\} &= \Pr\{\tau = \infty\} \\ &= \exp\left(-\sum_{n=1}^{\infty}\frac{1}{n}\Pr\{\bar{X}_1(n) - \mu_1 < -\gamma_0\}\right) \\ &\geq \exp\left(-\sum_{n=1}^{\infty}\frac{1}{n}\Pr\{|\bar{X}_1(n) - \mu_1| \geq \gamma_0\}\right) \\ &\geq \exp\left(-\sum_{n=1}^{\infty}\frac{\sigma_1^2}{n^2\gamma_0^2}\right) \qquad \text{(by Chebyshev's inequality)} \\ &= \exp\left(-\frac{\pi^2\sigma_1^2}{6\gamma_0^2}\right), \end{aligned} \tag{EC.2}$$

which is strictly greater than zero. Combining Equations (EC.1) and (EC.2) leads to the result. $\square$

EC.1.2. Proof of Lemma 3

Proof. To simplify the notation, let $b = (\gamma - \gamma_0)/\bar{\sigma}$ and $n_b = n^{\bar{f}}(b/2)$ with $\bar{f}(\cdot) = f(\cdot)/\underline{\sigma}$. Recall that for each alternative $i = 2, \cdots, k$,

$$\bar{T}_i^{\bar{f}}(b) = \inf\left\{n \geq n_b : \bar{Y}_i(n) < b/2\right\}. \tag{EC.3}$$

Furthermore, notice that

$$\begin{aligned}\inf\left\{n \geq n_b : \bar{Y}_i(n) \leq b/2\right\} &\leq \inf\left\{n = n_b m, m \geq 1 : \bar{Y}_i(n) < b/2\right\} \\ &= n_b \inf\left\{m \geq 1 : \bar{Y}_i(n_b m) < b/2\right\}.\end{aligned} \tag{EC.4}$$

Then, combining Lemma EC.1 and Equations (EC.3) and (EC.4), we may have

$$\begin{aligned}\mathrm{E}\left[\bar{T}^{\bar{f}}(b)\right] &\leq n_b \mathrm{E}\left[\inf\left\{m \geq 1 : \bar{Y}_i(n_b m) < b/2\right\}\right] \\ &= n_b \exp\left(\sum_{m=1}^{\infty} \frac{1}{m} \Pr\left\{\bar{Y}_i(n_b m) \geq b/2\right\}\right) \\ &\leq n_b \exp\left(\sum_{m=1}^{\infty} \frac{1}{m} \Pr\left\{|\bar{Y}_i(n_b m)| \geq b/2\right\}\right).\end{aligned} \tag{EC.5}$$

Further, recall that $\bar{Y}_i(n_b m)$ denotes the sample average of $n_b m$ i.i.d. random variables with mean zero and variance one under Assumption 2. Thus,

$$\mathrm{E}[\bar{Y}_i(n_b m)] = 0 \quad \text{and} \quad \mathrm{Var}[\bar{Y}_i(n_b m)] = \frac{1}{n_b m}.$$

Consequently, from Equation (EC.5) and the Chebyshev's inequality, we have

$$\begin{aligned}\mathrm{E}\left[\bar{T}^{\bar{f}}(b)\right] &\leq n_b \exp\left(\sum_{m=1}^{\infty} \frac{1}{m} \Pr\left\{|\bar{Y}_i(n_b m)| \geq b/2\right\}\right) \\ &\leq n_b \exp\left(\sum_{m=1}^{\infty} \frac{4}{m^2 n_b b^2}\right) = n_b \exp\left(\frac{2\pi^2}{3 n_b b^2}\right).\end{aligned} \tag{EC.6}$$

Lastly, according to the definition of $n^f(\cdot)$, we have that $n_b = n^{\bar{f}}(b/2) = n^f(b\underline{\sigma}/2)$ because $\bar{f}(\cdot) = f(\cdot)/\underline{\sigma}$. Substituting this for $b = (\gamma - \gamma_0)/\bar{\sigma}$ in Equation (EC.6) leads to the result. The proof is completed. □

EC.1.3. Proof of Lemma 5

Proof. For any fixed $n_0 \geq 1$, the expectation $\mathrm{E}\left[T(b; n_0)\right]$ satisfies

$$\mathrm{E}\left[T(b; n_0)\right] - n_0 = \mathrm{E}\left[\inf\left\{m \geq 0 : \bar{X}(m + n_0) < b\right\}\right] := \mathrm{E}[M(n_0)],$$

where, for notational simplicity, we define $M(n_0) = \inf\{m \geq 0 : \bar{X}(m+n_0) < b\}$. The expectation of $M(n_0)$ can be expressed as

$$\mathrm{E}[M(n_0)] = \sum_{m=0}^{\infty} m \Pr\{M(n_0) = m\} = \sum_{m=1}^{\infty} \Pr\{M(n_0) \geq m\}.$$

By the definition of $M(n_0)$, the following fact holds: when $M(n_0) \geq m \geq 1$ is true, we must have that all the events $\bar{X}(n_0 + m - 1) \geq b$ hold. This indicates that

$$\mathrm{E}[M(n_0)] = \sum_{m=1}^{\infty} \Pr\{M(n_0) \geq m\} \leq \sum_{m=1}^{\infty} \Pr\{\bar{X}(n_0+m-1) \geq b\} = \sum_{m=n_0}^{\infty} \Pr\{\bar{X}(m) \geq b\}.$$

Therefore,

$$\mathrm{E}[T(b;n_0)] = \mathrm{E}[M(n_0)] + n_0 \leq \sum_{m=n_0}^{\infty} \Pr\{\bar{X}(m) \geq b\} + n_0.$$

This completes the bound for $\mathrm{E}[T(b;n_0)]$.

Next, we consider the variance $\mathrm{Var}[T(b;n_0)]$. We have

$$\mathrm{Var}[T(b;n_0)] = \mathrm{Var}[T(b;n_0) - n_0] = \mathrm{Var}[M(n_0)] = \mathrm{E}[M^2(n_0)] - (\mathrm{E}[M(n_0)])^2 \leq \mathrm{E}[M^2(n_0)]. \tag{EC.7}$$

Notice that $\mathrm{E}[M^2(n_0)]$ can be rewritten as

$$\begin{aligned}
\mathrm{E}[M^2(n_0)] &= \sum_{m=0}^{\infty} m^2 \Pr\{M(n_0) = m\} \\
&= \sum_{m=1}^{\infty} m^2 \left(\Pr\{M(n_0) \geq m\} - \Pr\{M(n_0) \geq m+1\}\right) \\
&= \sum_{m=1}^{\infty} m^2 \Pr\{M(n_0) \geq m\} - \sum_{m=2}^{\infty} (m-1)^2 \Pr\{M(n_0) \geq m\} \\
&\leq 2\sum_{m=1}^{\infty} m \Pr\{M(n_0) \geq m\}.
\end{aligned} \tag{EC.8}$$

Combining Equations (EC.7) and (EC.8) leads to

$$\mathrm{Var}[T(b;n_0)] \leq \mathrm{E}[M^2(n_0)] \leq 2\sum_{m=1}^{\infty} m \Pr\{M(n_0) \geq m\} \leq 2\sum_{m=1}^{\infty} m \Pr\{\bar{X}(n_0+m-1) \geq b\}.$$

This completes the proof of the desired results.

□

EC.1.4. Proof of Lemma 6

To prove the lemma, we first prepare the following lemmas.

LEMMA EC.2 **(Corollary 1.8 of Nagaev 1979).** *Let $X_1, X_2, \cdots$ be a sequence of independent random variables, $\sigma_i^2 = \mathrm{Var}[X_i]$ and $S_n = \sum_i^n X_i$ for $n \geq 1$. If $E[X_i] = 0$ and $\sum_1^n E|X_i|^q < \infty$ for some $q > 2$, then*

$$P(S_n \geq x) \leq c_q^{(1)} x^{-q} \sum_{i=1}^{n} E|X_i|^q + \exp\left\{-\frac{c_q^{(2)} x^2}{\sum_1^n \sigma_i^2}\right\},$$

where $c_q^{(1)} = (1+2/q)^q$ and $c_q^{(2)} = 2(q+2)^{-2}e^{-q}$.

LEMMA EC.3. *For a random variable X with a finite mean μ, we have that for any $p \geq 1$,*

$$\mathrm{E}[|X-\mu|^p] \leq 2^p \mathrm{E}[|X|^p]. \tag{EC.9}$$

Proof. Recall that the L^p norm of a random variable is defined as

$$\|X\|_p = \{\mathrm{E}[|X|^p]\}^{1/p}.$$

Let $p \geq 1$. By the Jensen's inequality, we know that $\|\mu\|_p = |\mu| = |\mathrm{E}[X]| \leq E[|X|] = \|X\|_1$. Then, we have

$$\|X-\mu\|_p \leq \|X\|_p + \|\mu\|_p \leq \|X\|_p + \|X\|_1 \leq \|X\|_p + \|X\|_p = 2\|X\|_p, \tag{EC.10}$$

where the first inequality comes from the Minkowski's inequality and the last inequality holds because the L^p norm of a random variable is increasing in p for $p \geq 1$. Combining Equations (EC.9) and (EC.10) leads to the result of interest. □

LEMMA EC.4. *For a random variable X with mean μ and variance σ^2, if $\mathrm{E}[|X|^q] \leq M$ for some $q > 2$, it holds that*

$$\mu \leq M^{1/q} \quad \text{and} \quad \sigma^2 \leq 4M^{2/q}.$$

Proof. Using the Jensen's inequality, we may obtain

$$\mu = \mathrm{E}[X] \leq \mathrm{E}[|X|] = \left\{\{\mathrm{E}[|X|]\}^q\right\}^{1/q} \leq \{\mathrm{E}[|X|^q]\}^{1/q} = M^{1/q}.$$

For the variance σ^2, we have

$$\sigma^2 = \mathrm{E}[(X-\mu)^2] = \left\{\{\mathrm{E}[|X-\mu|^2]\}^{q/2}\right\}^{2/q} \leq \{\mathrm{E}[|X-\mu|^q]\}^{2/q} \leq \{2^q \mathrm{E}[|X|^q]\}^{2/q} \leq 4M^{2/q},$$

where the first inequality arises from the Jensen's inequality and the second inequality arises from Lemma EC.3. This completes the proof of the desired results. □

Proof of Lemma 6. Let $\sigma^2 = \mathrm{Var}[X_i]$. Due to Lemma EC.2, we can get

$$\begin{aligned}
P\left(\bar{X}(n) - \mu \geq x\right) = P\left(\sum_{i=1}^{n}(X_i - \mu) \geq nx\right) &\leq c_q^{(1)}(nx)^{-q} n \mathrm{E}\left|X_i - \mu\right|^q + \exp\left\{-\frac{c_q^{(2)}(nx)^2}{n\sigma^2}\right\} \\
&= c_q^{(1)} n^{-q+1} x^{-q} \mathrm{E}\left|X_i - \mu\right|^q + \exp\left\{-\frac{c_q^{(2)} nx^2}{\sigma^2}\right\} \\
&\leq c_q^{(1)} n^{-q+1} x^{-q} 2^q M + \exp\left\{-\frac{c_q^{(2)} nx^2}{4M^{2/q}}\right\} \\
&= a_1 n^{-q+1} x^{-q} + \exp\left\{-a_2 nx^2\right\}
\end{aligned}$$

where the second inequality follows from Lemma EC.3 and Lemma EC.4 and

$$a_1 = (2 + 4/q)^q M \quad \text{and} \quad a_2 = (q+2)^{-2} e^{-q} M^{-2/q}/2. \tag{EC.11}$$

The proof is completed. □

EC.1.5. Proof of Lemma 7

Proof. Let $c_1 = a_1 \cdot \frac{q-1}{q-2}$, $c_2 = a_2$, and $c_3 = 2a_1 \cdot \frac{q-2}{q-3}$ with a_1 and a_2 defined in Equation (EC.11). To prove the lemma, it suffices to derive expressions for general $\mathrm{E}\left[T_i(b;n_0)\right]$ and $\mathrm{Var}\left[T_i(b;n_0)\right]$ where $T_i(b;n_0) := \inf\{n \ge n_0 : \bar{X}_i(n) - \mu_i < b\}$, since

$$\mathrm{E}\left[T_i^f(\gamma-\gamma_0)\right] = \mathrm{E}\left[T_i\left(\tfrac{\gamma-\gamma_0}{2}; n^f\left(\tfrac{\gamma-\gamma_0}{2}\right)\right)\right] \quad \text{and} \quad \mathrm{Var}\left[T_i^f(\gamma-\gamma_0)\right] = \mathrm{Var}\left[T_i\left(\tfrac{\gamma-\gamma_0}{2}; n^f\left(\tfrac{\gamma-\gamma_0}{2}\right)\right)\right].$$

For $\mathrm{E}\left[T_i(b;n_0)\right]$, under Assumption 3, we have

$$\begin{aligned}
\mathrm{E}[T_i(b;n_0) - n_0] &\le \sum_{m=n_0}^{\infty} \Pr\left\{\bar{X}_i(m) - \mu_i \ge b\right\} \text{ (by Lemma 5)} \\
&\le a_1 b^{-q} \sum_{m=n_0}^{\infty} n^{-q+1} + \sum_{m=n_0}^{\infty} \exp\left\{-a_2 m b^2\right\} \text{ (by Lemma 6).} \qquad \text{(EC.12)}
\end{aligned}$$

For the first term in Equation (EC.12), given $q > 2$ and $n_0 \ge 1$, we have

$$\sum_{m=n_0}^{\infty} m^{-q+1} \le n_0^{-q+1} + \int_{n_0}^{\infty} u^{-q+1} du = n_0^{-q+1} + (q-2)^{-1} n_0^{-q+2} \le (q-1)/(q-2) n_0^{-q+2}. \quad \text{(EC.13)}$$

For the second term in Equation (EC.12), by algebraic calculations, we may have

$$\sum_{m=n_0}^{\infty} \exp\left\{-a_2 m b^2\right\} = \frac{\exp\left(-n_0 a_2 b^2\right)}{1-\exp(-a_2 b^2)} = \frac{\exp\left(-n_0 c_2 b^2\right)}{1-\exp(-c_2 b^2)}. \quad \text{(EC.14)}$$

Combining Equations (EC.12), (EC.13), and (EC.14) leads to

$$\mathrm{E}[T_i(b;n_0)] = \mathrm{E}[T_i(b;n_0) - n_0] + n_0 \le c_1 b^{-q} n_0^{-q+2} + \frac{\exp\left(-n_0 c_2 b^2\right)}{1-\exp(-c_2 b^2)} + n_0. \quad \text{(EC.15)}$$

For $\mathrm{Var}[T_i(b;n_0)]$, similarly, under Assumption 3, we may derive

$$\begin{aligned}
\mathrm{Var}[T_i(b;n_0)] &\le 2\sum_{m=1}^{\infty} m \Pr\left\{\bar{X}_i(m+n_0-1) - \mu_i \ge b\right\} \quad \text{(by Lemma 5)} \\
&\le 2a_1 b^{-q} \sum_{m=1}^{\infty} m(m+n_0-1)^{-q+1} + 2\sum_{m=1}^{\infty} m \exp\left\{-a_2(m+n_0-1)b^2\right\} \quad \text{(by Lemma 6)} \\
&\le 2a_1 b^{-q} \sum_{m=1}^{\infty} (m+n_0-1)(m+n_0-1)^{-q+1} + 2\sum_{m=1}^{\infty} (m+n_0-1) \exp\left\{-a_2(m+n_0-1)b^2\right\} \\
&= 2a_1 b^{-q} \sum_{m=n_0}^{\infty} m^{-q+2} + 2\sum_{m=n_0}^{\infty} m \exp\left\{-a_2 m b^2\right\}. \qquad \text{(EC.16)}
\end{aligned}$$

For the first term in Equation (EC.16), given $q > 3$ and $n_0 \ge 1$, we have that similar to Equation (EC.13),

$$\sum_{m=n_0}^{\infty} m^{-q+2} \le (q-2)/(q-3) n_0^{-q+3}. \quad \text{(EC.17)}$$

For the second term in Equation (EC.16), we define an auxiliary function $f(x) := \sum_{m=n_0}^{\infty} \exp\{-xm\}$. Similar to Equation (EC.14), $f(x) = \exp(-xn_0)/(1-\exp(-x))$. By taking the first derivative of $f(x)$, we obtain

$$f'(x) = \sum_{m=n_0}^{\infty} (-m) \exp\{-xm\} = \frac{-n_0 \exp(-n_0 x) + (n_0 - 1)\exp(-(n_0+1)x)}{(1-\exp(-x))^2},$$

which indicates that

$$\sum_{m=n_0}^{\infty} m \exp\{-xm\} = \frac{n_0 \exp(-n_0 x) - (n_0 - 1)\exp(-(n_0+1)x)}{(1-\exp(-x))^2} \le \frac{n_0 \exp(-n_0 x)}{(1-\exp(-x))^2}. \quad \text{(EC.18)}$$

Plugging Equations (EC.17) and (EC.18) into Equation (EC.16) leads to

$$\mathrm{Var}[T_i(b; n_0)] \le c_3 b^{-q} n_0^{-q+3} + \frac{2n_0 \exp(-n_0 c_2 b^2)}{(1-\exp(-c_2 b^2))^2}. \quad \text{(EC.19)}$$

Substituting $b = (\gamma - \gamma_0)/2$ and $n_0 = n^f((\gamma-\gamma_0)/2)$ in Equations (EC.15) and (EC.19) leads to the results of interest. The proof is completed. □

EC.1.6. Proof of Lemma 8

Proof. For any $\alpha \in (0,1)$, define the function $f(n) = \max\{f_1(n), f_2(n)\}$, where

$$f_1(n) = \sqrt[q]{\frac{2a_1 z(q')}{\alpha n^{q-1-q'}}} \text{ and } f_2(n) = \sqrt{\frac{\log(2z(q'+1)) + (q'+1)\log n + \log(1/\alpha)}{a_2 n}}. \quad \text{(EC.20)}$$

Here, $z(x) = \sum_{n=1}^{\infty} \frac{1}{n^x}$ denotes the Riemann zeta function, which is finite for all $x > 1$, and the constants a_1 and a_2 are defined in Equation (EC.11). It is clear that for any $1 < q' < q-1$, the function $f(n)$ satisfies Assumption 1. Under this choice of bonus function, we now show that

$$\Pr\{\exists n \ge 1 : \bar{X}_1(n) + f(n) < \mu_1\} \le \alpha.$$

Given $\mathrm{E}[|X_1|^q] \le M$ for some $q > 2$ and $M < \infty$, we have that for any $n \ge 1$,

$$\begin{aligned} \Pr\{\mu_1 > \bar{X}_1(n) + f(n)\} &\le \Pr\{\mu_1 - \bar{X}_1(n) \ge f(n)\} \\ &\le a_1 n^{-q+1} f(n)^{-q} + \exp\{-a_2 n f(n)^2\} \\ &\le a_1 n^{-q+1} f_1(n)^{-q} + \exp\{-a_2 n f_2(n)^2\}, \end{aligned} \quad \text{(EC.21)}$$

where the second inequality arises from Lemma 6. For the first term in Equation (EC.21), we may have

$$a_1 n^{-q+1}\{f_1(n)\}^{-q} = a_1 n^{-q+1} \left(\frac{2a_1 z(q')}{\alpha n^{q-1-q'}}\right)^{-1} = \frac{\alpha n^{q-1-q'}}{2z(q')} \cdot n^{-q+1} = \frac{\alpha}{2z(q')n^{q'}}. \quad \text{(EC.22)}$$

Next, for the second term in Equation (EC.21), we may have

$$\exp\left\{-a_2 n\left\{f_2(n)\right\}^2\right\} = \exp\left\{-\log(2z(q'+1)) - (q'+1)\log n - \log(1/\alpha)\right\} = \frac{\alpha}{2z(q'+1)n^{q'+1}} \tag{EC.23}$$

Plugging Equations (EC.22) and (EC.23) into Equation (EC.21) and taking the union bound lead to

$$\begin{aligned}
\Pr\left\{\exists n \geq 1 \text{ s.t. } \mu_1 > \bar{X}_1(n) + f(n)\right\} &= \Pr\left\{\bigcup_{n\geq 1}\left\{\mu_1 > \bar{X}_1(n) + f(n)\right\}\right\} \\
&\leq \sum_{n=1}^{\infty} \Pr\left\{\mu_1 \geq \bar{X}_1(n) + f(n)\right\} \\
&= \sum_{n=1}^{\infty} \frac{\alpha}{2z(q')n^{q'}} + \sum_{n=1}^{\infty} \frac{\alpha}{2z(q'+1)n^{q'+1}} = \alpha.
\end{aligned}$$

The proof is completed. □

REMARK EC.1. We use $q'+1$ instead of q' in the definition of $f_2(n)$ in Equation (EC.20) to match the form that appears in Equation (EC.23), a result we will use later in Equation (EC.28).

EC.1.7. Proof of Lemma 9

To prove Lemma 9, we first prepare the following convenient lemmas.

LEMMA EC.5. *Let $a, b, c > 0$, and suppose $x > 0$. Then, the inequality $\sqrt{\frac{a + c\log n}{bn}} \leq x$ holds whenever*

$$n \geq \max\left\{\frac{2a}{bx^2}, \frac{4c}{bx^2}\log\left(\frac{2c}{bx^2}\right)\right\}.$$

Proof. To ensure this holds, it suffices to enforce

$$\frac{a}{bn} \leq \frac{x^2}{2} \quad \text{and} \quad \frac{c\log n}{bn} \leq \frac{x^2}{2}.$$

The first condition implies $n \geq \frac{2a}{bx^2}$. For the second condition, note that it is equivalent to $\frac{\log n}{n} \leq \frac{bx^2}{2c}$. Let $f(n) = \frac{\log n}{n}$. This function achieves its maximum at $n = e$, with maximum value $f(e) = 1/e$, and is decreasing for $n \geq e$. Hence, for any $y \leq 1/e$, one can show that

$$\frac{\log n}{n} \leq y \quad \text{is ensured by} \quad n \geq \frac{2}{y}\log\left(\frac{1}{y}\right).$$

Applying this bound with $y = \frac{bx^2}{2c}$ yields $n \geq \frac{4c}{bx^2}\log\left(\frac{2c}{bx^2}\right)$. Therefore, both parts of the original inequality hold whenever

$$n \geq \max\left\{\frac{2a}{bx^2}, \frac{4c}{bx^2}\log\left(\frac{2c}{bx^2}\right)\right\}.$$

The proof is completed. □

LEMMA EC.6. *For any $x > 0$,*

$$\frac{1}{1-\exp(-x)} \leq 1 + \frac{1}{x}.$$

LEMMA EC.7. *For any $0 < a < 1$ and $k \geq 2$, it holds that*

$$\sum_{i=2}^{k} \left(\frac{k}{i}\right)^a \leq \frac{k}{1-a}$$

and

$$\sum_{i=2}^{k} \left(\frac{k}{i}\right)^a \log\left(\frac{k}{i}\right)^a \leq \frac{ak}{(1-a)^2}.$$

Proof. For the first part, we have that when $a > 0$,

$$\sum_{i=2}^{k} \left(\frac{k}{i}\right)^a \leq \int_{x=1}^{k} \left(\frac{k}{x}\right)^a dx = k^a \int_1^k x^{-a} dx = \frac{k^a}{1-a}\left(k^{1-a}-1\right) \leq \frac{k}{1-a}.$$

For the second part, similarly, we may have

$$\sum_{i=2}^{k} \left(\frac{k}{i}\right)^a \log\left(\frac{k}{i}\right)^a \leq a\int_{x=1}^{k} \left(\frac{k}{x}\right)^a \log\left(\frac{k}{x}\right) dx. \tag{EC.24}$$

Define $t = \log\left(\frac{k}{x}\right)$. Then $x = ke^{-t}$, $dx = -ke^{-t}dt$, and

$$\int_{x=1}^{k} \left(\frac{k}{x}\right)^a \log\left(\frac{k}{x}\right) dx = \int_{t=\log k}^{0} e^{at}\cdot t \cdot (-ke^{-t})dt = k\int_{t=0}^{\log k} te^{-(1-a)t}dt \leq k\int_{t=0}^{\infty} te^{-(1-a)t}dt. \tag{EC.25}$$

For the exponential integral, we know that

$$\int_0^{\infty} te^{-bt}dt = \frac{1}{b^2}, \quad \text{for } b > 0. \tag{EC.26}$$

Thus, when $a < 1$, combining Equations (EC.24), (EC.25), and (EC.26) lead to the result

$$\sum_{i=2}^{k} \left(\frac{k}{i}\right)^a \log\left(\frac{k}{i}\right)^a \leq \frac{ak}{(1-a)^2}.$$

This completes the proof. □

Proof of Lemma 9. In the proof, we consider the general case with $q' \geq 2$, although in the lemma we only present the result for $q' = 2$. Recall from Equation (16) that for each alternative $i = 2, \cdots, k$,

$$\hat{T}_i^f(\Delta_i) = \inf\left\{n \geq n^f\left(\frac{\mu_1 - \mu_i}{2}\right) : \bar{X}_i(n) - \mu_i \leq \frac{\mu_1 - \mu_i}{2}\right\}.$$

To prove the lemma, we first analyze the properties of

$$T_i(b; n^f(b)) = \inf\left\{n \geq n^f(b) : \bar{X}_i(n) - \mu_i \leq b\right\}$$

for $b>0$ and the bonus function $f(n)=\max\{f_1(n),f_2(n)\}$, where $f_1(n)$ and $f_2(n)$ are defined in Equation (EC.20). For the expectation of $T_i(b;n^f(b))$, from Equation (EC.15), we have

$$\begin{aligned}
\mathrm{E}\left[T_i(b;n^f(b))\right] &\le \frac{q-1}{q-2}a_1 b^{-q} n^f(b)^{-q+2} + \frac{\exp(-a_2 n^f(b) b^2)}{1-\exp(-a_2 b^2)} + n^f(b)\\
&\le \frac{q-1}{q-2}a_1 f_1(n^f(b))^{-q} n^f(b)^{-q+2} + \frac{\exp(-a_2 n^f(b) f_2(n^f(b))^2)}{1-\exp(-a_2 b^2)} + n^f(b)\\
&= \frac{q-1}{q-2}\cdot\frac{\alpha}{2z(q')} n^f(b)^{1-q'} + \frac{\frac{\alpha}{2z(q'+1)n^f(b)^{q'+1}}}{1-\exp(-a_2 b^2)} + n^f(b)\\
&\le \alpha + \frac{\alpha}{2}\left(\frac{1}{1-\exp(-a_2 b^2)}\right) + n^f(b)\\
&\le 2\alpha + \frac{\alpha}{2a_2 b^2} + n^f(b),
\end{aligned} \tag{EC.27}$$

where the second inequality holds because $f(n)=\max\{f_1(n),f_2(n)\}$ and, by the definition of $n^f(b)$, $f(n^f(b))\le b$, the equality holds from Equations (EC.22) and (EC.23), the third inequality holds because $q>3$ and $q'\ge 2$, and the last inequality holds because of Lemma EC.6.

Similarly, for the variance of $T_i(b;n^f(b))$, from Equation (EC.19), we have

$$\begin{aligned}
\mathrm{Var}\left[T_i(b;n^f(b))\right] &\le \frac{2(q-2)}{q-3}a_1 b^{-q} n^f(b)^{-q+3} + \frac{2n^f(b)\exp(-a_2 n^f(b) b^2)}{(1-\exp(-a_2 b^2))^2}\\
&\le \frac{2(q-2)}{q-3}a_1 f_1(n^f(b))^{-q} n^f(b)^{-q+3} + \frac{2n^f(b)\exp(-a_2 n^f(b) f_2(n^f(b))^2)}{(1-\exp(-a_2 b^2))^2}\\
&= \frac{2(q-2)}{q-3}\cdot\frac{\alpha}{2z(q')} n^f(b)^{2-q'} + \frac{\frac{\alpha}{z(q'+1)n^f(b)^{q'}}}{(1-\exp(-a_2 b^2))^2}\\
&\le \frac{\alpha(q-2)}{(q-3)} + \frac{\alpha}{n^f(b)^{q'}}\cdot\frac{1}{(1-\exp(-a_2 b^2))^2}\\
&\le \frac{\alpha(q-2)}{(q-3)} + \frac{\alpha}{n^f(b)^{q'}}\left(1+\frac{2}{a_2 b^2}+\frac{1}{a_2^2 b^4}\right),
\end{aligned} \tag{EC.28}$$

where the equality holds from Equations (EC.22) and (EC.23), the third inequality holds because $q'\ge 2$, and the last inequality holds because of Lemma EC.6.

The upper bounds in Equations (EC.27) and (EC.28) depend on the value of $n^f(b)$. We now provide an explicit form of $n^f(b)$. Recall from Lemma 2 that for a given $b>0$, $n^f(b)$ is defined as a positive integer such that $\forall n\ge n^f(b)$,

$$f(n^f(b))\le b,$$

which is ensured if both

$$f_1(n^f(b))\le b \quad \text{and} \quad f_2(n^f(b))\le b.$$

From the definition of $f_1(n)$ in Equation (EC.20), we have

$$f_1(n^f(b)) = \left(\frac{2a_1 z(q')}{\alpha n^f(b)^{q-1-q'}}\right)^{1/q} \le b \quad \Rightarrow \quad n \ge \left(\frac{2a_1 z(q')}{\alpha b^q}\right)^{\frac{1}{q-1-q'}}.$$

From the definition of $f_2(n)$ in Equation (EC.20), we apply the previously established Lemma EC.5:

$$f_2(n) = \sqrt{\frac{\log(2z(q'+1)) + (q'+1)\log n + \log(1/\alpha)}{a_2 n}} \le b$$

is ensured whenever

$$n \ge \max\left\{\frac{2\left[\log(2z(q'+1)) + \log(1/\alpha)\right]}{a_2 b^2}, \frac{4(q'+1)}{a_2 b^2}\log\left(\frac{2(q'+1)}{a_2 b^2}\right)\right\}.$$

Combining the two conditions, we choose $n^f(b)$ as

$$n^f(b) = \max\left\{\underbrace{c_1 \cdot b^{-\frac{q}{q-1-q'}}}_{\text{Term A}}, \underbrace{c_2 \cdot b^{-2}}_{\text{Term B}}, \underbrace{2c_3 \cdot b^{-2}\log(c_3 b^{-2})}_{\text{Term C}}\right\}, \tag{EC.29}$$

where the constants are defined as

$$c_1 = \left(\frac{2a_1 z(q'+1)}{\alpha}\right)^{\frac{1}{q-1-q'}}, c_2 = \frac{2\left[\log(2z(q'+1)) + \log(1/\alpha)\right]}{a_2}, \text{ and } c_3 = \frac{2(q'+1)}{a_2}. \tag{EC.30}$$

Now, we are ready to prove the lemma.

Part I. Recall that for each alternative $i = 2, \dots, k$, $\Delta_i = (i/k)^\beta$. From Equation (EC.27), we have that for $\sum_{i=2}^k \mathrm{E}[\hat{T}_i^f(\Delta_i)]$,

$$\begin{aligned}\sum_{i=2}^k \mathrm{E}[\hat{T}_i^f(\Delta_i)] &= \sum_{i=2}^k \mathrm{E}\left[T_i\left(\frac{\Delta_i}{2}; n^f\left(\frac{\Delta_i}{2}\right)\right)\right] \\ &\le 2\alpha k + \sum_{i=2}^k \frac{2\alpha}{a_2}\left(\frac{k}{i}\right)^{2\beta} + \sum_{i=2}^k n^f\left(\left(\frac{i}{k}\right)^\beta/2\right).\end{aligned} \tag{EC.31}$$

For the second term in Equation (EC.31), when $2\beta < 1$, Lemma EC.7 gives

$$\sum_{i=2}^k \frac{2\alpha}{a_2}\left(\frac{k}{i}\right)^{2\beta} \le \frac{2\alpha}{a_2}\frac{k}{1-2\beta}. \tag{EC.32}$$

For the third term in Equation (EC.31), from the formula of $n^f(\cdot)$ in Equation (EC.29), observe that

$$\sum_{i=2}^k n^f\left(\frac{1}{2}\left(\frac{i}{k}\right)^\beta\right) = \sum_{i=2}^k \max\{A_i, B_i, C_i\} \le \sum_{i=2}^k A_i + \sum_{i=2}^k B_i + \sum_{i=2}^k C_i \tag{EC.33}$$

with

$$A_i = c_1 2^{\frac{q}{q-1-q'}}\left(\frac{k}{i}\right)^{\frac{\beta q}{q-1-q'}}, B_i = 4c_2\left(\frac{k}{i}\right)^{2\beta}, \text{ and } C_i = 8c_3\left(\frac{k}{i}\right)^{2\beta}\log\left(4c_3\left(\frac{k}{i}\right)^{2\beta}\right).$$

Then, Lemma EC.7 gives that when $\frac{\beta q}{q-1-q'} < 1$,

$$\sum_{i=2}^k A_i \le \frac{c_1 2^{\frac{q}{q-1-q'}} k}{1 - \beta q/(q-1-q')}, \quad \sum_{i=2}^k B_i \le \frac{4c_2 k}{1-2\beta}, \text{ and } \sum_{i=2}^k C_i \le \frac{8c_3\log(4c_3)k}{1-2\beta} + \frac{16c_3\beta k}{(1-2\beta)^2}. \tag{EC.34}$$

Combining Equations (EC.31) , (EC.32), (EC.33) and (EC.34), we may obtain when $\beta < \min\left\{\frac{q-1-q'}{q}, \frac{1}{2}\right\}$,

$$\begin{aligned}\sum_{i=2}^{k} \mathrm{E}[\hat{T}_i^f(\Delta_i)] &\le 2\alpha k + \frac{2\alpha}{a_2}\frac{k}{1-2\beta} + \frac{c_1 2^{\frac{q}{q-1-q'}} k}{1-\beta q/(q-1-q')} + \frac{4c_2 k}{1-2\beta} + \frac{8c_3\log(4c_3)k}{1-2\beta} + \frac{16c_3\beta k}{(1-2\beta)^2} \\ &= d_0 k + \frac{d_1 k}{1-\beta q/(q-1-q')} + \frac{d_2 k}{1-2\beta} + \frac{d_3\beta k}{(1-2\beta)^2},\end{aligned}$$

where $d_0 = 2\alpha, d_1 = c_1 2^{\frac{q}{q-1-q'}}, d_2 = \frac{2\alpha}{a_2} + 4c_2 + 8c_3\log(4c_3)$, and $d_3 = 16c_3$ with c_1, c_2, and c_3 in Equation (EC.30). This completes the proof for the first part.

Part II. For $\mathrm{Var}\left[\hat{T}_i^f(\Delta_i)\right]$, from Equation (EC.28), we have

$$\mathrm{Var}\left[\hat{T}_i^f(\Delta_i)\right] \le \frac{\alpha(q-2)}{(q-3)} + \frac{\alpha}{n^f(b)^{q'}}\left(1 + \frac{2}{a_2 b^2} + \frac{1}{a_2^2 b^4}\right), \tag{EC.35}$$

where $b = \Delta_i/2 = \frac{1}{2}\left(\frac{i}{k}\right)^\beta$. From the definition of $n^f(b)$ in Equation (EC.29), when $q' \ge 2$, we have the bound

$$\frac{1}{n^f(b)^{q'}} \le c_2^{-q'} b^{2q'}. \tag{EC.36}$$

Furthermore, since $b = \frac{1}{2}\left(\frac{i}{k}\right)^\beta < 1$ for all $i \ge 2$, and $q' \ge 2$, it follows that for all $i \ge 2$,

$$b^{2q'} \le 1, \quad b^{2q'-2} \le 1, \text{ and } b^{2q'-4} \le 1. \tag{EC.37}$$

Substituting Equations (EC.36) and (EC.37) into Equation (EC.35) leads to

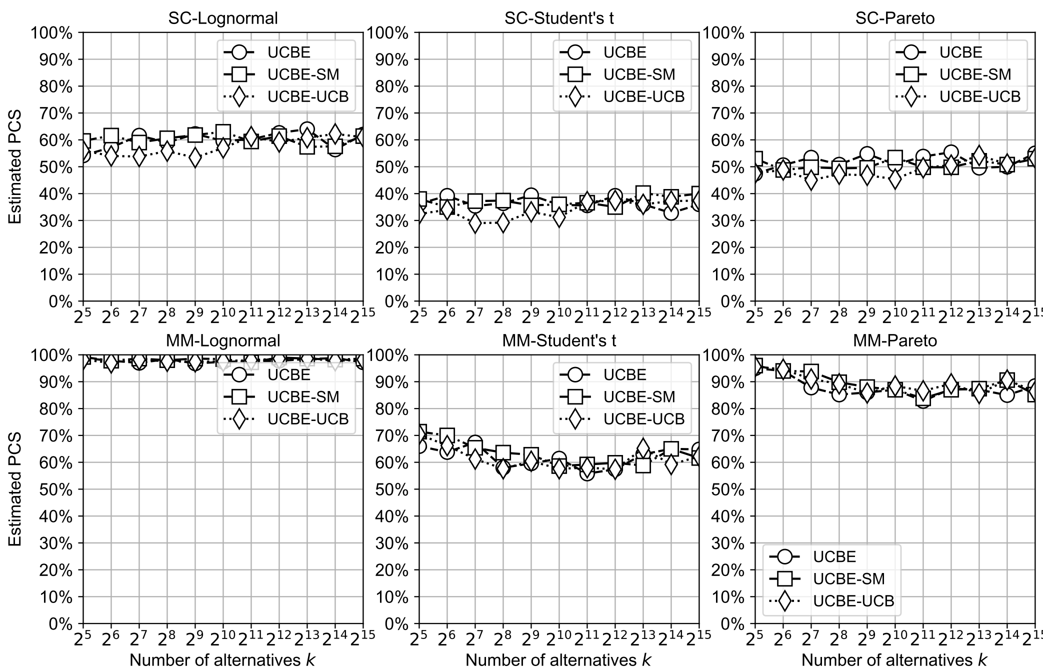
$$\begin{aligned}\mathrm{Var}\left[\hat{T}_i^f(\Delta_i)\right] &\le \frac{\alpha(q-2)}{(q-3)} + \alpha c_2^{-q'}\left(b^{2q'} + \frac{2b^{2q'-2}}{a_2} + \frac{b^{2q'-4}}{a_2^2}\right) \\ &\le \frac{\alpha(q-2)}{(q-3)} + \alpha c_2^{-q'}\left(1 + \frac{2}{a_2} + \frac{1}{a_2^2}\right) := d_4 < \infty.\end{aligned}$$

The proof is completed. □

EC.2. Impact of Selection Standards

In this section, we numerically evaluate the performance of UCB algorithms satisfying Assumption 1 under different selection standards. Specifically, we consider two base algorithms, UCBE and MOSS, both of which select the alternative with the largest sample size once the total sampling budget is exhausted. We compare each base algorithm with two variants: one that makes the final decision based on the largest sample mean (SM), and another based on the largest UCB value (UCB). This yields six algorithms in total: UCBE, UCBE-SM, UCBE-UCB, MOSS, MOSS-SM, and MOSS-UCB. For each base algorithm, we adopt the same problem configurations and experimental settings as in Section 6.2.1. The PCS results for the three UCBE variants are shown in Figure EC.1, while those for the three MOSS variants are shown in Figure EC.2. We can see that across all tested configurations, the three selection standards yield comparable PCS. However, we emphasize that this empirical similarity does not imply that our theoretical analysis can be directly extended to these alternative selection rules. Our analysis focuses on the largest sample size rule due to the technical convenience it offers in establishing sample optimality.

Figure EC.1 PCS of UCBE with Different Selection Standards.



References

Siegmund D (1985) *Sequential analysis: tests and confidence intervals* (Springer Science & Business Media).

Nagaev SV (1979) Large deviations of sums of independent random variables. *The Annals of Probability* 7(5) 745–789.

Figure EC.2 PCS of MOSS with Different Selection Standards.