SPRT-Based Efficient Best Arm Identification in Stochastic Bandits

Arpan Mukherjee🆔 and Ali Tajer🆔, *Senior Member, IEEE*

Abstract—**This paper investigates the best arm identification (BAI) problem in stochastic multi-armed bandits in the fixed confidence setting. The general class of the exponential family of bandits is considered. The existing algorithms for the exponential family of bandits face computational challenges. To mitigate these challenges, the BAI problem is viewed and analyzed as a sequential composite hypothesis testing task, and a framework is proposed that adopts the likelihood ratio-based tests known to be effective for sequential testing. Based on this test statistic, a BAI algorithm is designed that leverages the canonical sequential probability ratio tests for arm selection and is amenable to tractable analysis for the exponential family of bandits. This algorithm has two key features: (1) its sample complexity is asymptotically optimal, and (2) it is guaranteed to be $\delta-$PAC. Existing efficient approaches focus on the Gaussian setting and require Thompson sampling for the arm deemed the best and the challenger arm. Additionally, this paper analytically quantifies the computational expense of identifying the challenger in an existing approach. Finally, numerical experiments are provided to support the analysis.**

Index Terms—**Best arm identification, sequential probability ratio test.**

I. INTRODUCTION

THIS paper considers the problem of best arm identification (BAI) in stochastic multi-armed bandits (MABs). In a stochastic MAB, each arm generates rewards from distributions with *unknown* mean values. The objective of a learner in BAI is to identify the arm with the largest mean value, using the fewest number of samples.

The BAI problem is broadly studied under two key settings: the *fixed budget* setting and the *fixed confidence* setting. The objective in the fixed budget setting is to identify the arm having the largest mean within a pre-specified sampling budget while minimizing the decision error probability. On the other hand, in the fixed confidence setting, the learner identifies the best arm while having a guarantee on the error probability, and the objective is to minimize the sample complexity. The problem of identifying the best arm was first proposed

in [1], where the problem was posed in the fixed budget setting. More investigations in this setting can be found in [2] and [3]. Representative studies in the fixed confidence setting can be found in [4], [5], [6], [7], [8], [9]. Algorithms in this setting can be further classified into two categories, namely, non-Bayesian algorithms and Bayesian algorithms.

In the non-Bayesian setting, an optimal algorithm for parametric stochastic bandits was first proposed in [6] for the single parameter exponential family, which is based on a tracking procedure for arm selection. Specifically, in each round, [6] computes the optimal sampling proportion at the current mean estimates and selects an arm based on the optimal sampling proportion. While the track and stop (TaS) algorithm exhibits asymptotic optimality, it is computationally expensive due to solving an optimization problem to compute the optimal allocation in each iteration. To reduce the computational complexity, [10] showed that tracking the optimal proportion in intervals with exponentially increasing gaps is sufficient. However, this study focuses only on linear bandits with Gaussian noise. To address the computational complexity of track and stop, another approach to solving BAI is the gamification approach [11], [12]. In this approach, BAI is posed as an unknown two-player game in which the sampling rules are obtained from the iterative strategies of the two players, which converge to a saddle point. The state-of-the-art algorithm for BAI in the non-Bayesian parametric setting is the Frank-Wolfe sampling algorithm (FW) [13]. In this approach, the sampling rule is obtained from a single iteration of the Frank-Wolfe algorithm, which involves solving a two-player zero-sum game in each iteration. Despite being computationally efficient compared to TaS, the FW algorithm involves solving a linear program in each round.

Some of the non-Bayesian approaches to solving BAI for the non-parametric class of stochastic MABs (such as the class of sub-Gaussian stochastic MABs or bounded variance stochastic MABs) include confidence interval-based approaches (see [4], [7], [14]), successive elimination-based approaches (see [9], [14], [15], [16], [17], [18]) and tracking based approaches [19]. In the confidence interval-based approach, the learner computes the sample mean of each arm, and a confidence interval around these empirical estimates, within which the true mean exists with a high probability. The rationale behind this strategy is to gather more evidence until there is no overlap between the confidence intervals and the learner decides the best arm based on the empirical estimates. On the other hand, the successive elimination-based strategy involves eliminating the potentially suboptimal arms in each

round and continuing sampling from all other arms until only one arm remains to be eliminated.

While non-Bayesian approaches have been investigated extensively, the Bayesian setting is far less investigated. The first Bayesian algorithm was investigated in [20], which introduced the *top-two* design philosophy and designed three Bayesian algorithms based on that. The top-two algorithmic design mitigated the computational challenge encountered in TaS, and provided a computationally simple alternative to the arm selection strategy. Among them, a modification of the Thompson sampling algorithm [21], [22], [23], called the top-two Thompson sampling (TTTS) has received more attention due to its simplicity and optimality properties. The sample complexity of TTTS, however, was not analyzed in [20]. To address the sample complexity of the top-two algorithms in Gaussian bandits, an improved algorithm was devised in [24], where the sample complexity was shown to be asymptotically optimal up to a constant factor. The sample complexity of TTTS in the Gaussian setting was later analyzed in [25], and it was shown to be asymptotically optimal. However, the sample complexity analysis was based on a non-informative prior, reducing it to a non-Bayesian setting. Furthermore, despite its simplicity, TTTS faces a computational challenge in its sampling strategy, which becomes significant in the regime of diminishing error rates. To circumvent this, a transportation cost-based sampler was adopted in designing an efficient Thompson sampling-based BAI algorithm called the top-two transportation cost (T3C) [25]. However, the sample complexity of T3C is only analyzed in the Gaussian setting.

This paper leverages a sequential hypothesis testing framework for formalizing and solving the BAI problem in the fixed confidence setting. The arm selection and stopping rules are, in spirit, similar to the sequential probability ratio test [26]. BAI has been viewed as a hypothesis testing problem in a wide body of literature starting from the investigation in [6], and subsequently in [11], [13], [24], [25]. Despite that, the algorithms offered generally do not adopt the statistics known to be effective for sequential composite hypothesis testing. In contrast, in this investigation, the arm selection rules dynamically update generalized likelihood ratios that compare the relative likelihood of different arms for being among the best. We refer to this algorithm by the top-two sequential probability ratio test (TT-SPRT). The idea of using SPRT-based rules was first introduced in [27], the key advantage of which is being amenable to analysis in the broader class of the exponential family of bandits. Specifically, the analysis for Thompson sampling-based approaches relies on non-asymptotic concentration inequalities for the convergence of the posterior mean to the ground truth. These concentration results exist in the literature for Gaussian bandits [25]. However, for the broader class of the exponential family, the analysis of posterior sampling-based approaches is contingent on tail bounds for the posterior means, which need to be derived. Furthermore, empirically, computing a posterior distribution may involve Monte Carlo integration [20], in case a closed-form conjugate prior does not exist. Both of these issues make the log-likelihood ratio-based test statistic a more reasonable choice compared to posterior sampling. Prior to this, the sample complexity of

top-two algorithms [20], [24], [25] has only been analyzed in the special setting of Gaussian bandits. This investigation is a generalization of [24], [25] in the sense that the TT-SPRT algorithm is shown to be asymptotically optimal for the single-parameter exponential family. Thus, we have addressed the open question of developing an *efficient* BAI algorithm in the fixed-confidence setting for the single parameter exponential family of distributions.

We show that in the special case of Gaussian bandits, TT-SPRT addresses a computational weakness of the TTTS algorithm. Specifically, for dynamically identifying the top two arms, the TTTS sampling strategy generates random samples from the posterior distributions of arm rewards. Subsequently, the coordinate with the largest value in the first sample is deemed the best arm's index. For identifying the second arm (the challenger), TTTS keeps generating more samples until the coordinate with the largest value is distinct from the index already identified as the best arm. After enough explorations, the posterior distribution converges to the true model, and the largest coordinate of any random sample will be pointing to the best arm. This increases the delay in encountering a challenger. TT-SPRT does not have such a computational challenge. Finally, we note that our arm selection strategy is, in spirit, similar to the sequential probability ratio test [26] and [28], which is a powerful test in a wide range of sequential testing problems, owing to its optimality properties and computational simplicity.

Besides this paper (and its earlier version [27]) leveraging SPRT-based rules for the top-two algorithms with various choices of the top arm and the challenger arm have also been recently further investigated in [29], including the choice that has been considered in this paper. While the results in [29] confirm the findings in this paper, there are a few critical differences in the assumptions on the bandit model considered. Specifically, for the exponential family analysis, [29] assumes sub-Gaussian distributions and having distinguishable arm means. However, we make no such assumption for our analysis of the exponential family. In contrast, to account for this assumption, our algorithm involves a forced exploration stage.

II. BAI IN STOCHASTIC BANDITS

Consider a *K*-armed *stochastic* MAB setting with *K* arms generating rewards based on probability distributions that belong to an exponential family of distributions. Specifically, corresponding to a convex, twice-differentiable function $b : \mathbb{R} \mapsto \mathbb{R}$, we consider a single-parameter exponential family

$$\mathscr{P}_b \triangleq \left\{ \nu_\theta : \frac{d\nu_\theta}{d\nu} = \exp(\theta x - b(\theta)) \ , \ \theta \in \Theta \right\} \qquad (1)$$

where $\Theta \subseteq \mathbb{R}$ is a compact parameter space, ν_θ is the probability measure associated with parameter $\theta \in \Theta$, and ν is a reference measure on $\mathbb{R}$ such that $\nu_\theta \ll \nu$ for all $\theta \in \Theta$. The mean value of the distribution $\nu_\theta \in \mathscr{P}_b$ is given by $\mu(\theta) \triangleq \dot{b}(\theta)$, and the members of $\mathscr{P}_b$ can be uniquely identified by their mean values. We denote the *unknown* mean of arm i by μ_i and denote its associated probability distribution (probability density function for continuous and probability mass function

for discrete distributions) from $\mathscr{P}_b$ by $\nu(\cdot \mid \mu_i)$. Accordingly, the vector of mean values is denoted by $\boldsymbol{\mu} \triangleq [\mu_1, \ldots, \mu_K]$. The exponential family bandit model associated with $\boldsymbol{\mu}$ is denoted by $\nu_{\boldsymbol{\mu}} \triangleq [\nu(\cdot \mid \mu_1), \ldots, \nu(\cdot \mid \mu_K)]$. The *best* arm, assumed to be unique, has the largest mean value and it is denoted by

$$a^\star \triangleq \arg\max_{i \in [K]} \mu_i. \tag{2}$$

The gap between the expected values of the best arm and arm $i \in [K]$ is denoted by $\Delta_i \triangleq \mu_{a^\star} - \mu_i$, and the smallest gap among all possible arm pairs is captured by

$$\Delta_{\min} \triangleq \min_{i \neq j} |\mu_i - \mu_j|. \tag{3}$$

We allow $\Delta_{\min} = 0$ for the analysis of the single-parameter exponential family. In the special case of Gaussian bandits, we prove an implicit exploration property of our proposed algorithm with the additional assumption that $\Delta_{\min} > 0$. However, such an assumption is not required in general, facilitated by the forced exploration stage in our algorithm design presented in Section III. The learner performs a sequence of arm selections with the objective of identifying the best arm, i.e., $a^\star$, with the fewest number of samples (on average). In the sequential arm selection process, at time $n \in \mathbb{N}$ the leaner selects arm A_n and receives the reward X_n, generating the filtration $\mathcal{F}_n \triangleq \{(A_t, X_t) : t \in \{1, \ldots, n\}\}$. Arm selection decision A_{n+1} is assumed to be $\mathcal{F}_n$−measurable. Accordingly, the sequence of arm selections and the corresponding rewards obtained by the learner up to time n are denoted by

$$\mathcal{A}_n \triangleq \{A_1, \ldots, A_n\} \text{ and } \mathcal{X}_n \triangleq \{X_1, \ldots, X_n\}. \tag{4}$$

The sub-sequence of the rewards from only arm $i \in [K]$ up to time n is denoted by

$$\mathcal{X}_n^i \triangleq \{X_s : s \in [n], A_s = i\}. \tag{5}$$

We use the notation $D_{KL}(\nu(\cdot \mid \mu_i) \| \nu(\cdot \mid \mu_j))$ to denote the Kullback-Leibler (KL) divergence from $\nu(\cdot \mid \mu_j)$ to $\nu(\cdot \mid \mu_i)$. It can be readily verified [6] that this divergence measure can be specified as a function of μ_i and μ_j, for which we use the shorthand notation

$$\begin{aligned} d_{KL}(\mu_i \| \mu_j) &\triangleq D_{KL}(\nu(\cdot \mid \mu_i) \| \nu(\cdot \mid \mu_j)) \\ &= b(\dot{b}^{-1}(\mu_j)) - b(\dot{b}^{-1}(\mu_i)) \\ &\quad - \mu_i[\dot{b}^{-1}(\mu_j) - \dot{b}^{-1}(\mu_i)]. \end{aligned} \tag{6}$$

Let τ, which is $\mathcal{F}_\tau$-measurable, denote the *stochastic* stopping time of the BAI algorithm, that is the time instant at which the BAI algorithm terminates the search process and identifies an arm as the best arm. Accordingly, define $\hat{A}_\tau$ as the arm identified as the best arm at the stopping time. In this paper, we consider the *fixed confidence* setting, in which the learner's objective is to identify the best arm $a^\star$ with a pre-specified confidence level. We use δ−PAC and β-optimality as the canonical notions of optimality for assessing the efficiency of the BAI algorithms. These notions are formalized next. First, we specify the δ−PAC guarantee, which evaluates the fidelity of the terminal decision.

Definition 1 (δ−PAC): A BAI algorithm is δ−PAC if the algorithm has a stopping time τ adapted to $\{\mathcal{F}_t : t \in \mathbb{N}\}$, and at the stopping time it ensures

$$\mathbb{P}_{\boldsymbol{\mu}}\left\{\tau < +\infty, \hat{A}_\tau = a^\star\right\} > 1 - \delta, \tag{7}$$

where $\mathbb{P}_{\boldsymbol{\mu}}$ denotes the probability measure induced by the interaction of the BAI algorithm with the bandit instance $\nu_{\boldsymbol{\mu}}$.

For analyzing the sample complexity, we leverage a setting-specific notion of *problem complexity* defined next. The problem complexity characterizes the level of difficulty that an algorithm faces for identifying the best arm with sufficient fidelity. For this purpose, we define $\mathscr{W}(\beta)$ as the following set of K-dimensional probability simplexes

$$\begin{aligned} \mathscr{W}(\beta) &\triangleq \Big\{\boldsymbol{\omega} \triangleq [\omega_1, \ldots, \omega_K] \in [0,1]^K : \omega_{a^\star} = \beta, \\ &\qquad \sum_{i \in [K]} \omega_i = 1\Big\}. \end{aligned} \tag{8}$$

Definition 2 (Problem Complexity): For a given bandit model $\nu_{\boldsymbol{\mu}}$, the problem complexity is defined as

$$\begin{aligned} \Gamma_{\boldsymbol{\mu}}(\beta) &\triangleq \max_{\boldsymbol{\omega} \in \mathscr{W}(\beta)} \min_{i \neq a^\star} \Big(\beta\, d_{KL}(\mu_{a^\star} \| \mu_{a^\star,i}(\boldsymbol{\omega})) \\ &\qquad + \omega_i\, d_{KL}(\mu_i \| \mu_{a^\star,i}(\boldsymbol{\omega}))\Big) \end{aligned} \tag{9}$$

where we have defined

$$\mu_{i,j}(\boldsymbol{\omega}) \triangleq \frac{\mu_i\, \omega_i + \mu_j\, \omega_j}{\omega_i + \omega_j}. \tag{10}$$

Accordingly, we define the optimal sampling proportions as

$$\begin{aligned} \boldsymbol{\omega}^\star(\beta) &\triangleq \arg\max_{\boldsymbol{\omega} \in \mathscr{W}(\beta)} \min_{i \neq a^\star} \Big(\beta\, d_{KL}(\mu_{a^\star} \| \mu_{a^\star,i}(\boldsymbol{\omega})) \\ &\qquad + \omega_i\, d_{KL}(\mu_i \| \mu_{a^\star,i}(\boldsymbol{\omega}))\Big). \end{aligned} \tag{11}$$

By leveraging the notion of problem complexity, and defining $T_{n,i}$ as the number of times that arm $i \in [K]$ is pulled until n, we have the following known information-theoretic lower bound on the sample complexity of any δ−PAC BAI algorithm.

Theorem 1 (Sample Complexity - Lower bound [25]): For any δ−PAC algorithm that satisfies

$$\frac{T_{n,a^\star}}{n} \xrightarrow{n \to \infty} \beta \tag{12}$$

we almost surely have

$$\liminf_{\delta \to 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}[\tau]}{\log(1/\delta)} \geq \frac{1}{\Gamma_{\boldsymbol{\mu}}(\beta)}. \tag{13}$$

The universal lower bound in Theorem 1 provides the minimum number of samples that any δ−PAC BAI algorithm requires asymptotically, provided that β fraction of measurement effort is allocated to the best arm. Accordingly, we specify β-optimality as a measure of sample complexity.

Definition 3 (β-optimality): A δ−PAC BAI algorithm is called asymptotically β-optimal if it satisfies

$$\frac{T_{n,a^\star}}{n} \xrightarrow{n \to \infty} \beta \text{ a.s.} \text{ and } \limsup_{\delta \to 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}[\tau]}{\log(1/\delta)} \leq \frac{1}{\Gamma_{\boldsymbol{\mu}}} \tag{14}$$

where $\mathbb{E}_{\boldsymbol{\mu}}$ denotes expectation under the measure $\mathbb{P}_{\boldsymbol{\mu}}$.

III. TT-SPRT Algorithm for BAI

Hypothesis Testing Framework: We view the BAI problem as a collection of binary *composite* hypothesis testing problems as follows. For each pair of distinct arms $(i,j) \in [K]^2$, define the following composite hypothesis, testing which discerns which of the two arms i and j has a larger mean value:

$$\mathcal{H}_{i,j} : \mu_i > \mu_j. \tag{15}$$

Unlike in sequential hypothesis testing that aims to solve problem (15), our objective is *not* forming decisions for all these $\binom{K}{2}$ different hypotheses. Instead, we use them to form relevant measures that capture the likelihood of different hypotheses. We then use these measures to design the arm selection and stopping rules. Specifically, we form generalized likelihood ratios as the key ingredient for our decision rules. To this end, at time n and corresponding to each arm pair (i,j) we define the generalized log-likelihood ratio (GLLR)

$$\Lambda_n(i,j) \triangleq \log \frac{\sup_{\boldsymbol{\mu}} \mathbb{P}_{\boldsymbol{\mu}}(\mathcal{X}_n \mid \mathcal{H}_{i,j} \text{ is true})}{\sup_{\boldsymbol{\mu}} \mathbb{P}_{\boldsymbol{\mu}}(\mathcal{X}_n \mid \mathcal{H}_{j,i} \text{ is true})}. \tag{16}$$

It can be readily verified that for the exponential family of bandits, $\Lambda_n(i,j)$ can be simplified to

$$\Lambda_n(i,j) = \log \frac{\max_{\mu_i > \mu_j} \prod_{x_i \in \mathcal{X}_n^i} \nu(x_i \mid \mu_i) \prod_{x_j \in \mathcal{X}_n^j} \nu(x_j \mid \mu_j)}{\max_{\mu_j > \mu_i} \prod_{x_i \in \mathcal{X}_n^i} \nu(x_i \mid \mu_i) \prod_{x_j \in \mathcal{X}_n^j} \nu(x_j \mid \mu_j)}. \tag{17}$$

It can be readily verified (for instance see [6]) that the GLLRs in (17) have simple closed form, specified in the next lemma. To specify the closed form, we define the empirical mean

$$\mu_{n,i} \triangleq \frac{1}{T_{n,i}} \sum_{t \in [n]: a_t = i} X_t \tag{18}$$

as an empirical estimate of μ_i. Accordingly, define the weighted average terms

$$\mu_{n,i,j} \triangleq \frac{T_{n,i}\mu_{n,i} + T_{n,j}\mu_{n,j}}{T_{n,i} + T_{n,j}} \quad \forall i,j \in [K]. \tag{19}$$

Lemma 1 [6]: Under the exponential family of distributions, the GLLRs defined in (17) have a closed form given by

$$\Lambda_n(i,j) = \left[T_{n,i}\, d_{KL}(\mu_{n,i} \| \mu_{n,i,j}) + T_{n,j}\, d_{KL}(\mu_{n,j} \| \mu_{n,i,j}) \right] \times \mathbb{1}_{\{\mu_{n,i} > \mu_{n,j}\}}. \tag{20}$$

We note that in the Gaussian setting, the closed forms for the GLLRs in (20) simplify to

$$\Lambda_n(i,j) = \frac{1}{2\sigma^2} \cdot \frac{(\mu_{n,i} - \mu_{n,j})^2}{T_{n,i}^{-1} + T_{n,j}^{-1}} \cdot \mathbb{1}_{\{\mu_{n,i} > \mu_{n,j}\}}. \tag{21}$$

Based on these definitions, the detailed steps of the TT-SPRT algorithm are specified in Algorithm 1. The decision rules involved are discussed next.

Arm Selection Rule: At each instant n, the TT-SPRT identifies a *top* arm as the arm that has the largest sample mean $\mu_{n,i}$, denoted by

$$a_n^1 \in \arg\max_{i \in [K]} \mu_{n,i}. \tag{22}$$

Algorithm 1: Top-Two SPRT (TT-SPRT)

1 **Input:** β
2 **Initialize:** $n = 0$, $\mathcal{I}_n = [K]$, $\mu_{n,i} = 0 \ \forall\ i \in [K]$, $\Lambda(a_n^1, a_n^2) = 0$, $c_{n,\delta} = 0$
3 **while** $\Lambda(a_n^1, a_n^2) \leq c_{n,\delta}$ **do**
4 $\quad$ $n \leftarrow n+1$
5 $\quad$ Sample $D_n \sim \mathsf{Bern}(\beta)$
6 $\quad$ Play an arm a_n specified by (25) or (26) and obtain reward X_n
7 $\quad$ Update μ_{n,a_n} using (18)
8 $\quad$ $a_n^1 \leftarrow \arg\max_{i \in [K]} \mu_{n,i}$
9 $\quad$ Compute $\Lambda_n(a_n, i)$ for every $i \in [K] \setminus \{a_n^1\}$ using (20)
10 $\quad$ $a_n^2 \leftarrow \arg\min_{i \in [K] \setminus \{a_n^1\}} \Lambda_n(a_n^1, i)$
11 $\quad$ Update $c_{n,\delta}$ using (28) or (33)
12 $\quad$ Update $\mathcal{I}_n$ using (24)
13 **Output:** Top arm a_n^1

In addition to the top arm, we also define the *challenger* arm as the one which is the closest competitor of the *top* arm for being the *best* arm. The challenger arm at time n is the arm that minimizes the log-likelihood ratio computed with respect to the top arm a_n^1. We denote the *challenger* arm at time n by a_n^2 and it is specified by

$$a_n^2 \triangleq \arg\min_{j \in [K]: \mu_{n,j} < \mu_{n,a_n^1}} \Lambda_n(a_n^1, j). \tag{23}$$

The TT-SPRT sampling strategy consists of an explicit exploration phase to ensure that each arm is pulled sufficiently often. Specifically, the arm selection strategy works as follows. We say an arm is *under-explored*, if at time n it is selected fewer than $\sqrt{n/K}$ times. Accordingly, define the set of under-explored arms at time n as

$$\mathcal{I}_n \triangleq \left\{ i \in [K] : T_{n,i} \leq \left\lceil \sqrt{n/K} \right\rceil \right\}. \tag{24}$$

If $\mathcal{I}_n \neq \emptyset$, then TT-SPRT selects the arm in $\mathcal{I}_n$ that has been explored the least. Otherwise, it randomizes between the top arm a_n^1 and the challenger a_n^2 based on a Bernoulli random variable $D_n \sim \mathsf{Bern}(\beta)$, where $\beta \in (0,1)$ is a tunable parameter. This mechanism is included to ensure sufficient exploration of all arms. Hence, the arm action rule at time instant $n+1$ is given by

$$a_{n+1} \triangleq \begin{cases} \arg\min_{i \in \mathcal{I}_n} T_{n,i} & \text{if } \mathcal{I}_n \neq \emptyset \\ a_n^1 & \text{if } D_n = 1 \text{ and } \mathcal{I}_n = \emptyset \\ a_n^2 & \text{if } D_n = 0 \text{ and } \mathcal{I}_n = \emptyset \end{cases}. \tag{25}$$

We note that in the Gaussian bandit setting, the explicit form of the GLLR defined in (21) automatically ensures a sufficient exploration of the same order, i.e., at least $\sqrt{n/K}$ times for each arm. We will show this property in Section IV-B (Lemma 5). However, this is not true in general, and the exploration mechanism is necessary. For example, in the case of Bernoulli bandits, unlike for the Gaussian bandits, the GLLRs for some of the arms, especially the ones with low average reward could be infinite in the early stages. Thus, TT-SPRT without the forced exploration phase might never sample these

arms. Hence, in the special case of the Gaussian bandit setting, TT-SPRT does not require the explicit exploration phase, and is only based on randomizing between the top-two arms. This is formalized next.

Arm Selection Rule for Gaussian Bandits: Based on the choices of the top and challenger arms, at time n, the TT-SPRT sampling strategy selects one of the two arms based on a Bernoulli random variable $D_n \sim \mathsf{Bern}(\beta)$, where $\beta \in (0,1)$ is a tunable parameter. Specifically, given $\mathcal{F}_n$-measurable decisions a_n^1 and a_n^2 at time n, the arm to be sampled at time $n+1$ is specified by the randomized rule

$$a_{n+1} \triangleq \begin{cases} a_n^1, & \text{if } D_n = 1 \\ a_n^2, & \text{if } D_n = 0 \end{cases}. \qquad (26)$$

Stopping Rule: We specify the following thresholding-based stopping criterion, in which we select the threshold such that the algorithm ensures the $\delta-$PAC guarantee:

$$\tau \triangleq \inf\left\{ n \in \mathbb{N} : \Lambda_n\left(a_n^1, a_n^2\right) > c_{n,\delta} \right\}. \qquad (27)$$

This stopping rule at each time n evaluates the GLLR between the top and challenger arms, i.e., $\Lambda_n(a_n^1, a_n^2)$, and terminates the sampling process as soon as this GLLR exceeds a pre-specified threshold $c_{n,\delta}$. This threshold will be specified in Section IV. We note that the stopping criterion is different from that of the canonical SPRT's for sequential binary hypothesis testing, which involves two thresholds in order to control the two types of decision errors. In contrast to hypothesis testing, for BAI we have only one error event $\mathbb{P}_\mu(\hat{A}_\tau \neq a^\star)$ to control.

IV. Main Results

In this section, we establish the TT-SPRT algorithm's optimality for the general exponential family in Section IV-A. Subsequently, in Section IV-B, we specialize them to the Gaussian setting. The results in the Gaussian setting have two main distinctions from the results for the general exponential family. First, we show that the exploration mechanism for the Gaussian setting is not necessary. Secondly, we provide a less stringent stopping rule, according to which the sampling process requires fewer samples. Based on these specialized rules, we recover the optimality guarantees that were also established in [24], [25]. Even though the TT-SPRT achieves the optimality guarantees similar to those in [24], [25], it has a major computational advantage, which we discuss in Section IV-C.

A. Exponential Family of Bandits

Our main technical results are on the optimality of the sample complexity. Specifically, we demonstrate that the sample complexity of the TT-SPRT algorithm coincides with the known information-theoretic lower bounds on the sample complexity of the algorithms that are $\delta-$PAC. Hence, as the first step, and for completeness, we provide the results that establish that TT-SPRT algorithm is $\delta-$PAC. An algorithm being $\delta-$PAC, generally, depends only on its design of the terminal decision rule for identifying the best arm and is independent of the arm selection strategy. For this purpose, we leverage an

existing result for the exponential family [30], which demonstrates that the combination of any arm selection strategy, the stopping rule specified in (27), and a proper choices of the threshold $c_{n,\delta}$ ensures a $\delta-$PAC guarantee. For completeness, the result is provided in the following theorem.

Theorem 2 ($\delta-$PAC – Exponential): For the exponential family, the stopping rule in (23) with the choice of the threshold

$$c_{n,\delta} \triangleq 2\mathcal{C}_{\exp}\left(\frac{1}{2}\log\frac{K-1}{\delta}\right) + 6\log\left(\log\frac{n}{2}+1\right), \quad (28)$$

coupled with any arm selection strategy is $\delta-$PAC, where the function $\mathcal{C}_{\exp} : \mathbb{R}_+ \mapsto \mathbb{R}_+$ is defined similarly to [30], and satisfies $\mathcal{C}_{\exp}(x) \simeq x + 4\log(1 + x + \sqrt{2x})$.

Proof: Follows similarly to [30, Th. 7]. ∎

To show the optimality of TT-SPRT algorithm, we provide an upper bound on its sample complexity, which matches the information-theoretic lower bound on the sample complexity of any $\delta-$PAC BAI algorithm. Note that the explicit exploration phase incorporated in TT-SPRT ensures that each arm is explored sufficiently often and that the sample means converge to the true values. Before delineating the sample complexity achieved by the TT-SPRT, we provide the key properties of the TT-SPRT arm selection strategy. The first property pertains to the sufficiency of exploration.

Lemma 2 (Sufficient Exploration – Exponential): Under the TT-SPRT sampling strategy (25), for all $i \in [K]$ and for any $n \in \mathbb{N}$, we have $T_{n,i} \geq \sqrt{n/K} - 1$.

Proof: The proof follows similar arguments as in the proof of [6, Lemma 8]. ∎

This property ensures that each arm is sampled sufficiently often, such that the empirical estimates converge to the true mean value if the sampling strategy is allowed to continue drawing samples without stopping. The second property pertains to the convergence of the empirical mean estimates to the true mean values. Specifically, we show that there exists a time after which the empirical mean values are within an ϵ-band of their corresponding ground truth values.

Lemma 3 (Convergence in Mean – Exponential): For any $\epsilon > 0$, define N_ϵ^μ as

$$N_\epsilon^\mu \triangleq \inf\left\{s \in \mathbb{N} : |\mu_{n,i} - \mu_i| \leq \epsilon \ \forall n \geq s, \ \forall i \in [K]\right\}. (29)$$

Under the sampling rule in (25) for the exponential family we have $\mathbb{E}_\mu[N_\epsilon^\mu] < +\infty$.

Proof: See Appendix B. ∎

The convergence in mean values for the case of Gaussian bandits has been analyzed in [24], [25] under different top-two sampling rules. The analysis mainly relies on a concentration inequality for the convergence of the sample mean to the ground truth. Furthermore, the analysis relies on the assumption that $\Delta_{\min} > 0$, despite showing an implicit exploration of the arms due to the sampling rule. In contrast, our analysis for the exponential family makes no such assumption at the cost of an explicit exploration phase. Our analysis leverages the explicit exploration phase and Chernoff's inequality for the convergence in mean. We refer to Appendix B for more details. The third property is that the TT-SPRT ensures that the ratio of the sampling resources spent on arm i up to

time n, i.e., $T_{n,i}/n$ converges to the optimal sampling proportion $\omega_i^*(\beta)$ when n is sufficiently large. This is formalized next.

Lemma 4 (Convergence to Optimal Proportions): Under the sampling rule in (25) for the exponential family of bandits, there exists a stochastic time N_ϵ^ω, $\mathbb{E}_{\boldsymbol{\mu}}[N_\epsilon^\omega] < +\infty$, such that

$$\left| \frac{T_{n,i}}{n} - \omega_i^\star(\beta) \right| \leq \epsilon \quad \forall n > N_\epsilon^\omega \; \forall i \in [K]. \quad (30)$$

Proof: See Appendix C. ∎

Convergence in the sampling proportions to the β-optimal allocation is a key step in the sample complexity analysis of top-two algorithms, as shown in [24], [25]. The proof for the convergence can be broken down into two steps: 1) convergence of the sampling proportion of the best arm to β, and 2) convergence of the sampling proportion of the remaining arms to the β-optimal allocation. The novelty in the analysis of TT-SPRT comes in the second part. Specifically, TT-SPRT uses the GLLR statistic for arm selection, which differs from the statistic in [24], [25]. The analysis for Convergence in sampling proportion relies on showing that, eventually, the challenger is always contained in the set of the under-sampled arms. For details, we refer to Appendix C. Finally, we use the notion of β-optimality, which was first introduced in establishing optimality guarantees for the top-two algorithms for BAI in [24], and was adopted later for establishing the optimality of the TTTS algorithm [25]. Specifically, we show that the TT-SPRT algorithm is β-optimal, i.e., its sample complexity asymptotically matches the universal lower bound on the average sample complexity provided in Theorem 1, while satisfying the condition on the sampling proportion allocated to the best arm $a^\star$. By leveraging the three properties shown in Lemma 2-Lemma 4, we provide an upper bound on the average sample complexity of the TT-SPRT algorithm.

Theorem 3 (Sample Complexity - Upper bound): The TT-SPRT algorithm is $\delta-$PAC and almost surely satisfies

$$\frac{T_{n,a^\star}}{n} \xrightarrow{n \to \infty} \beta. \quad (31)$$

Furthermore, for the exponential family of distributions, we have

$$\lim_{\delta \to 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}[\tau]}{\log(1/\delta)} \leq \frac{1}{\Gamma_{\boldsymbol{\mu}}(\beta)} \quad (32)$$

Proof: See Appendix D. ∎

Theorem 3 along with the universal lower bound in Theorem 1 establishes the asymptotic β-optimality of TT-SPRT for the exponential family of bandits.

B. Gaussian Bandits

Next, we specialize the properties of the TT-SPRT algorithm to the Gaussian bandit setting. First, we leverage a tighter stopping condition from [25], which ensures that in the specific setting of Gaussian bandits, the stopping rule specified in (27) is $\delta-$PAC. Specifically, the threshold for the exponential family scales as $O(6 \log \log n)$, while the one for the Gaussian setting scales as $O(4 \log \log n)$.

Theorem 4 ($\delta-$PAC $-$ Gaussian): In the Gaussian bandit setting, the stopping rule in (27) with the choice of the threshold

$$c_{n,\delta} \triangleq 4 \log(4 + \log n) + 2g\left(\frac{\log(K-1) - \log \delta}{2} \right), \quad (33)$$

where we have defined $g(x) \triangleq x + \log x$, coupled with any arm selection strategy is $\delta-$PAC.

Proof: Follows similarly to [25, Th. 2]. ∎

Next, we show that the explicit exploration phase is redundant in the Gaussian setting, since the TT-SPRT sampling rule automatically ensures sufficient exploration in this setting.

Lemma 5 (Sufficient Exploration $-$ Gaussian): The TT-SPRT sampling strategy (26) ensures that there exists a random variable N_1 such that $\mathbb{E}_{\boldsymbol{\mu}}[N_1] < +\infty$ and for all $i \in [K]$ and for any $n > N_1$, we have $T_{n,i} \geq \sqrt{n/K}$ almost surely.

Proof: See [31, Appendix B]. ∎

Similarly to Lemma 3, we show that due to the implicit exploration of the TT-SPRT sampling rule for Gaussian bandits, we have convergence in the empirical mean values to the ground truth.

Lemma 6 (Convergence in Mean $-$ Gaussian): Under the sampling rule in (26) for the Gaussian setting, there exists a stochastic time M_ϵ^μ, $\mathbb{E}_{\boldsymbol{\mu}}[M_\epsilon^\mu] < +\infty$, such that for all $n > M_\epsilon^\mu$, we have:

$$|\mu_{n,i} - \mu_i| < \epsilon \quad \text{almost surely.} \quad (34)$$

Proof: See Appendix B. ∎

We show that the TT-SPRT algorithm in the Gaussian setting is β-optimal by providing an upper bound on the average sample complexity of the TT-SPRT algorithm in the Gaussian setting.

Theorem 5 (Sample Complexity - Upper bound): The TT-SPRT algorithm is $\delta-$PAC and almost surely satisfies

$$\frac{T_{n,a^\star}}{n} \xrightarrow{n \to \infty} \beta. \quad (35)$$

Furthermore, for the Gaussian model, we have

$$\lim_{\delta \to 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}[\tau]}{\log(1/\delta)} \leq \frac{1}{\Gamma_{\boldsymbol{\mu}}(\beta)}. \quad (36)$$

Proof: See Appendix D. ∎

C. Challenger Identification in Top-Two Thompson Sampling

The TTTS algorithm [20], is a Bayesian algorithm in which the reward mean values are assumed to have the prior distribution $\mathcal{N}(0, \kappa^2)$. Based on this prior, at each time n and based on $\mathcal{X}_n$, the learner computes a posterior distribution $\Pi_n \in \mathbb{R}^K \to \mathbb{R}$. Specifically, for the average reward realization $\bar{\boldsymbol{\mu}}$ and reward realization $\mathcal{X}_n = \mathbf{x}_n$:

$$\Pi_n(\bar{\boldsymbol{\mu}} \mid \mathbf{x}_n) \triangleq \prod_{i=1}^{K} \frac{1}{\sqrt{2\pi \eta_{n,i}^2}} \exp\left\{ -\frac{(\bar{\mu}_i - \tilde{\mu}_{n,i})^2}{2\eta_{n,i}^2} \right\}, \quad (37)$$

where we have defined

$$\tilde{\mu}_{n,i} \triangleq \frac{1}{T_{n,i} + \sigma^2/\kappa^2} \sum_{x \in \mathcal{X}_n^i} x, \text{ and } \eta_{n,i}^2 \triangleq \frac{\sigma^2}{T_{n,i} + \sigma^2/\kappa^2}. \quad (38)$$

While the TTTS is devised for the setting with a Gaussian prior distribution for the rewards, the sample complexity analysis for the algorithm holds for the asymptotic regime of $\kappa \to +\infty$. This assumption renders the prior distributions uninformative, and the setting becomes equivalent to that of the non-Bayesian counterpart, i.e., when the means corresponding to each arm is unknown, and we have no prior distribution over the arm means. Thus, the posterior mean corresponding to each arm $i \in [K]$ defined in (38) reduces to that of the sample mean, i.e., $\tilde{\mu}_{n,i} = \mu_{n,i}$, and the settings for both TTTS as well as TT-SPRT become equivalent. As a result, Π_n denotes the product of the K Gaussian posteriors, $\mathcal{N}(\mu_{n,i}, \sigma_{n,i}^2)$ for all $i \in [K]$, where we have defined $\sigma_{n,i}^2 \triangleq \sigma^2/T_{n,i}$. Let us denote the expectation operator with respect to Π_n by $\mathbb{E}_n$.

The arm selection strategy of the TTTS algorithm works as follows. At each time n, a random K-dimensional sample $\boldsymbol{\theta}^n \triangleq (\theta_1^n, \ldots, \theta_K^n)$ is drawn from the posterior distribution Π_n. The coordinate with the largest value is defined as the index of the top arm, denoted by $b_n^1 \triangleq \arg\max_{i \in [K]} \theta_i^n$. In order to find a challenger (the closest competitor to b_n^1), the algorithm continues sampling the posterior Π_n until a realization from Π_n is encountered such that the index of its largest coordinate is distinct from b_n^1. This is considered the challenger arm and its index is denoted by b_n^2. Encountering a challenger arm requires generating enough samples from Π_n. As n increases and the posterior distribution Π_n points to more confidence about the best arm, the number of samples required to encounter a challenger increases. We denote a sample s generated from Π_n by $\boldsymbol{\theta}_s^n \triangleq (\theta_{s,1}^n, \ldots, \theta_{s,K}^n)$. By design, clearly, $b_n^2 \triangleq \arg\max_{i \in [K]} \theta_{s,i}^n$, and $b_n^2 \neq b_n^1$. Once b_n^1 and b_n^2 are identified, the TTTS selects one of them based on a Bernoulli random variable parameterized by $\beta \in (0, 1)$. As mentioned earlier, as n increases and Π_n converges, the number of samples required for encountering a challenger also increases, and this imposes a computational challenge, especially for large n. In the next theorem, we show that the number of samples required for encountering a challenger scales at least exponentially in $\sqrt{n}$. For this purpose, we define

$$T_{\text{TTTS}}^n \triangleq \inf\left\{s \in \mathbb{N} : \exists i \in [K] , \ \theta_{s,i}^n > \theta_{s,b_n^1}^n\right\} \qquad (39)$$

as the number of posterior samples required for finding a challenger at time n.

Theorem 6 (Challenger's Sample Complexity): In   the TTTS algorithm [20], there exists a random variable N_0 such that for any $n > N_0$, the average number of posterior samples required in order to find a challenger is lower-bounded as

$$\mathbb{E}_n\big[T_{TTTS}^n\big] \geq \min_{i \in [K]\setminus\{a^\star\}} 2 \exp\big(\sqrt{\tfrac{n}{K}} \, C_{i,L}\big). \qquad (40)$$

Furthermore, for all $n > \max\{N_0, 32\sigma^2/9\Delta_{\min}^2\}$, we have

$$\mathbb{E}_n\big[T_{TTTS}^n\big] \leq \max_{i \in [K]\setminus\{a^\star\}} \sqrt{2\pi e} \exp\big(n \, C_{i,U}\big) \qquad (41)$$

where we have defined

$$C_{i,L} \triangleq \frac{(\Delta_i - \Delta_{\min}/2)^2}{4\sigma^2} \quad \text{and} \quad C_{i,U} \triangleq \frac{(\Delta_i + \Delta_{\min}/2)^2}{2\sigma^2}.$$

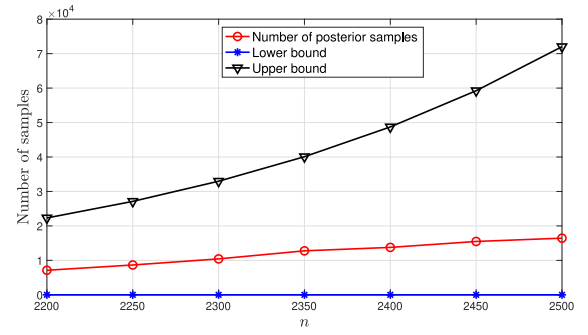*Proof:* See Appendix E.   ∎

Fig. 1.   Average number of posterior samples versus n.

We observe that the lower bound increases exponentially in $\sqrt{n}$, and thus, diverges for large values of n, i.e., when the confidence required on the decision quality is large. Furthermore, note that for a sufficiently small δ, the TTTS stopping time always exceeds N_0. Specifically, we show in Appendix E that there exists $\delta(N_0) > 0$ such that for any $\delta \in (0, \delta(N_0))$, the TTTS algorithm almost surely stops after N_0 time instants. This implies that for a large enough decision confidence, the TTTS algorithm requires a large number of posterior samples to identify a challenger, which has an exponential growth of the order of at least $\sqrt{n}$.

V. Numerical Experiments

In this section, we conduct numerical experiments to compare the performance of TT-SPRT against state-of-the-art algorithms. Specifically, we provide experiments for empirically depicting the computational difficulty of the TTTS algorithm in identifying a challenger arm. Furthermore, we compare the performance of TT-SPRT against existing BAI strategies under three different bandit models, namely, Gaussian bandits, Bernoulli bandits, and exponential bandits. While β is a tunable parameter of TT-SPRT, and its optimal choice depend on instance-specific parameters, a choice of $\beta = 0.5$ exhibits good empirical performance over several bandit instances. Specifically, the average sample complexity with $\beta = 0.5$ is at most twice the asymptotically optimal sample complexity, shown in [20, Lemma 3]. Unless otherwise stated, we have used $\beta = 0.5$ for simulations involving the top-two sampling strategies.

A. Computational Difficulty of TTTS

First, to show the computational difficulty of obtaining a challenger in TTTS, in Figure 1 we evaluate the average number of posterior samples required by TTTS to identify a challenger. As shown analytically in Theorem 6, the expected number of samples required for encountering a challenger scales at least exponentially in $\sqrt{n}$. It is observed from Figure 1 that as n increases, such average number of posterior samples increases drastically. We also plot the lower bound on the average number of posterior samples obtained in Theorem 6, which matches the scaling behavior of the actual number of samples. Furthermore, Figure 2 illustrates the scaling behavior of the number of posterior samples against various levels of the minimum gap compared to the best arm,
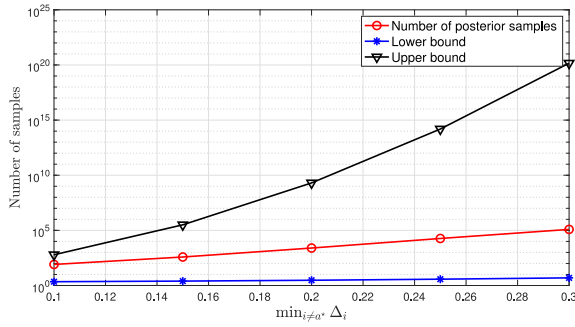
Fig. 2. Average number of posterior samples versus gaps.

Fig. 3. Sensitivity to β (Gaussian): $\boldsymbol{\mu} = [5, 4.5, 1, 1, 1]$, $\delta = 10^{-15}$.

Fig. 4. Sample complexity (Gaussian): $\boldsymbol{\mu} = [5, 4.5, 1, 1, 1]$, $\delta = 10^{-15}$.

i.e., $\min_{i \neq a^\star} \Delta_i$. For this experiment, we have set $n = 500$. It can be readily verified from Theorem 6 that as the minimum gap between the means of the best arm and any other arm increases, it becomes more challenging to obtain a challenger from the posterior distribution after it has converged sufficiently. This can be observed in Figure 2. Specifically, the number of posterior samples increases at least exponentially at a rate of $\min_{i \neq a^\star} \Delta_i^2$, and it can become extremely large when the gap is large.

B. Gaussian Bandits

Next, we compare the empirical performance of the TT-SPRT algorithm to existing strategies for BAI in the Gaussian bandit setting. Specifically, we compare against four existing BAI algorithms, which are DKM [11], LUCB [5], T3C [25], and FW [13]. Note that we have not compared our algorithm with TTTS, since TTTS requires an enormous computation time to identify a challenger in our experimental setting. T3C is a computationally efficient alternative to the TTTS algorithm, which has been proposed in [25]. Specifically, T3C is based on posterior sampling for identifying the top arm and replaces the resampling procedure for identifying a challenger in TTTS by defining the challenger based on the minimum transportation cost compared to the best arm. Despite its computational efficiency in selecting a challenger, T3C computes a posterior distribution at each time for identifying the top arm, which may involve Monte Carlo integration without conjugate priors [20]. The DKM algorithm is based on a gamification principle for algorithm design, which involves a **w** player who plays a sampling distribution and a **λ** player who plays an alternate bandit instance. We have implemented a best-response (zero-regret) **λ** player with the Adahedge **w** player, as described in [11]. The LUCB algorithm is based on identifying two arms, a top arm and a challenger, based on the current confidence intervals of the mean estimates. At each iteration, the LUCB algorithm samples both arms and updates their corresponding mean estimates. Finally, the FW algorithm is based on a single iteration of a Frank-Wolfe update step to compute a sampling proportion in each round.

We have considered two Gaussian bandit instances, one with $\Delta_{\min} = 0$, and the other with $\Delta_{\min} > 0$. These instances are characterized by the mean values $\boldsymbol{\mu}_1 = [5, 4.5, 1, 1, 1]$, and $\boldsymbol{\mu}_2 = [1, 0.85, 0.8, 0.7]$. For both instances, we set $\sigma = 1$, and $\boldsymbol{\mu}_2$ is a bandit instance from the experiments in [11]. The
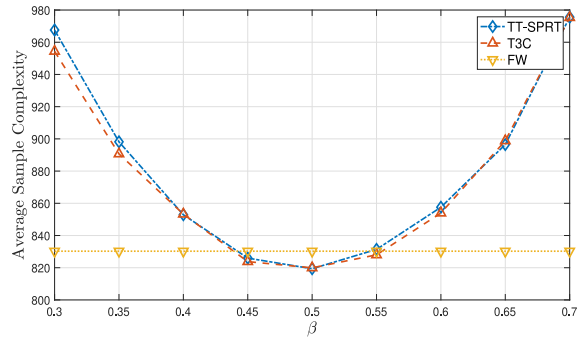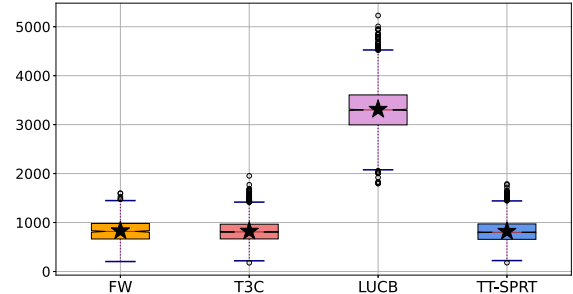

Fig. 5. Sensitivity to β (Gaussian): $\boldsymbol{\mu} = [1, 0.85, 0.8, 0.7]$, $\delta = 0.1$.
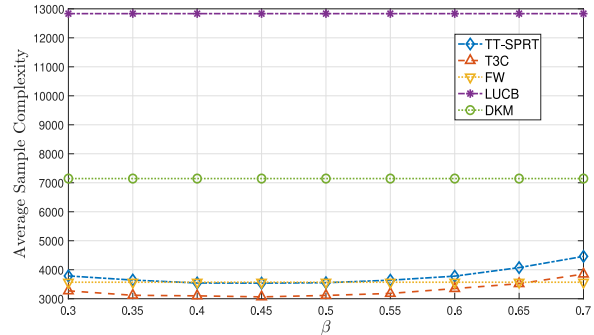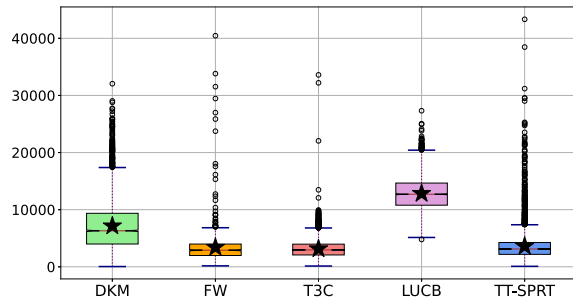

Fig. 6. Sample complexity (Gaussian): $\boldsymbol{\mu} = [1, 0.85, 0.8, 0.7]$, $\delta = 0.1$.

experiment with $\boldsymbol{\mu}_1$ is plotted in Figure 3 and Figure 4, which confirms the superior performance of TT-SPRT over the FW sampling strategy for $0.45 \leq \beta \leq 0.55$. Note that we have not plotted DKM and LUCB for this experiment due to their large empirical sample complexities, 3317.8 and 10, 335 for LUCB and DKM, respectively. The plot corresponding to bandit instance $\boldsymbol{\mu}_2$ can be found in Figure 5 and Figure 6. In this
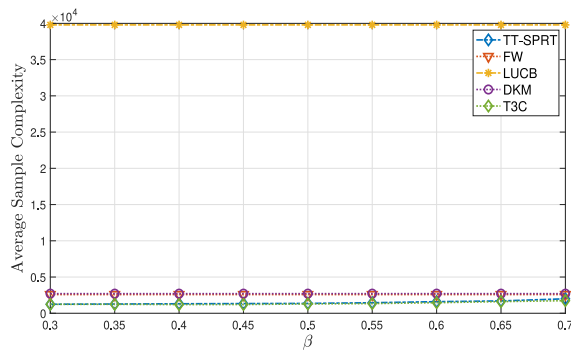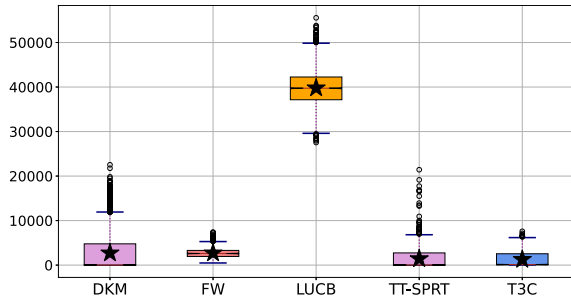
Fig. 7.   Sensitivity to β: (Bernoulli).

Fig. 9.   Sensitivity to β: (Exponential).

Fig. 8.   Sample complexity (Bernoulli).

Fig. 10.   Sample complexity (Exponential).

experiment, we observe that the empirical sample complexity of the T3C and FW algorithms is slightly better than TT-SPRT. However, this comes at an increased computational cost for the FW sampling strategy for solving a linear program in each iteration. For our implementations in MATLAB, we have used the simplex method for this purpose, which has a worst-case complexity of the order of $O(2^K)$ [32]. Furthermore, the T3C algorithm requires computing a posterior distribution in each iteration, which may involve Monte Carlo integration if a conjugate prior does not exist [20]. Note that the TT-SPRT algorithm uses explicit exploration for the instance $\boldsymbol{\mu}_1$, since it has $\Delta_{\min} = 0$. The effect of explicit exploration (as present in TT-SPRT) versus implicit exploration (as in algorithms such as T3C) is observed to vary, depending on the bandit instance. For example, in $\boldsymbol{\mu}_1$, T3C is observed to have a comparable performance as TT-SPRT. On the other hand, in $\boldsymbol{\mu}_2$, we observe that T3C performs slightly better than TT-SPRT. Overall, the necessity of explicit exploration depends on the algorithm design and the bandit instance, and whether or not it improves (or hurts) the sample complexity seems to be unclear.

C. Bernoulli Bandits

Next, we compare the sample complexity of TT-SPRT in the Bernoulli bandit setting. For comparison, we use a state-of-the-art FW sampling strategy [13], along with existing approaches such as DKM [11] and LUCB [5]. For this experiment, we have used the Bernoulli bandit instance from [6], [11], with mean values $\boldsymbol{\mu} = [0.3, 0.21, 0.2, 0.19, 0.18]$. We have set $\delta = 0.1$, and the performance comparison has been plotted in Figure 7 and Figure 8. Note that there is a difference between
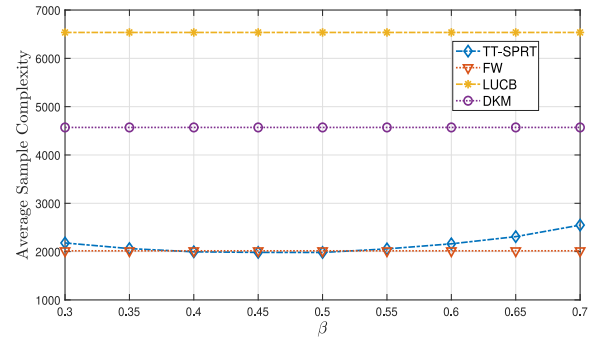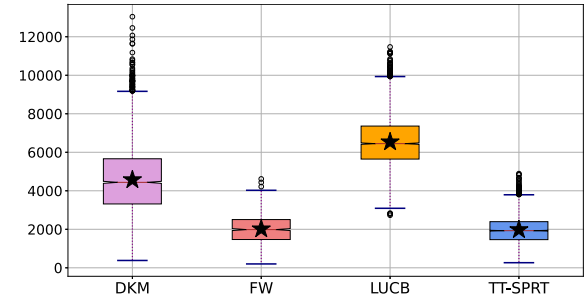
the simulations with DKM in [11], even though we use the same bandit instance and value of δ. This is because [11] uses a "stylized threshold" of $c_{n,\delta} = \log((1 + \log n)/\delta)$, which is disallowed by theory. In contrast, we use the theoretically grounded stopping thresholds [30]. Figure 8 clearly shows that TT-SPRT outperforms the state-of-the-art algorithms. Furthermore, we observe a significantly worse performance of the LUCB algorithm, which is due to the loose confidence bound on the mean estimates prescribed in [5].

D. Exponential Bandits

Finally, since TT-SPRT generalizes to any member of the single parameter exponential family, we provide an example with an exponential bandit instance with $\boldsymbol{\mu} = [0.9, 0.7, 0.5, 0.3, 0.1]$. We have set $\delta = 0.1$, and the performance of TT-SPRT is compared against that of LUCB [5], DKM [11] and FW [13] in Figure 9 and Figure 10. We can observe that TT-SPRT outperforms DKM and LUCB by a significant margin. Furthermore, its performance is comparable to the FW sampling strategy for a wide range of β, including the choice of $\beta = 0.5$, which has been used to obtain the box plot. Specifically, at $\beta = 0.5$, TT-SPRT has an empirical sample complexity of 1982.5, while that of the FW sampling strategy is 2016.5. Thus, TT-SPRT has a low empirical sample complexity in the example of exponential bandits, despite having a significantly lower computational cost compared to the state-of-the-art FW sampling rule. Furthermore, in order to highlight the computational burden incurred by the FW sampling rule, we tabulate the computation times required by TT-SPRT and other existing BAI strategies in Table I. These simulations are performed in MATLAB R2022a, on an Apple M1 pro processor equipped with 16 gigabytes of

TABLE I
AVERAGE COMPUTATION TIME (IN SECONDS)

Algorithm	FW	TT-SPRT	DKM	LUCB	T3C
Time	8.7984	0.0235	0.0388	0.0218	0.412

RAM. Clearly, TT-SPRT outperforms the FW sampling rule in terms of computational time by a wide margin. Note that we have not compared the computation times for TaS, since it is significantly worse (as much as 10 times that of the FW algorithm [33, Appendix G.1]).

VI. CONCLUSION

In this paper, we have investigated the problem of best arm identification (BAI) in stochastic multi-armed bandits (MABs). Arm selection and terminal decision rules are characterized based on generalized likelihood ratio tests with similarities to the conventional sequential probability ratio tests. The decisions rules (dynamic arm selection and stopping time) have two main properties: (1) they achieve optimality in the probably approximately correct learning framework, and, (2) they asymptotically achieve the optimal sample complexity. We have analytically characterized the optimality properties, and comparisons with the state-of-the-art are shown both analytically and numerically. We have also analyzed the computational challenges in an existing top-two sampling strategy for BAI.

APPENDIX A
RELEVANT LEMMAS

Lemma 7 [24, Lemma 5]: In the Gaussian bandit setting, there exists a random variable W_1 such that for all $i \in [K]$ and for all $n \in \mathbb{N}$, we almost surely have

$$|\mu_{n,i} - \mu_i| \le \sigma W_1 \sqrt{\frac{\log(e + T_{n,i})}{1 + T_{n,i}}}. \tag{42}$$

and $\mathbb{E}_{\boldsymbol{\mu}}[e^{sW_1}] < +\infty$ for all $s \in \mathbb{R}^+$.

Lemma 8 (Cramér-Chernoff Inequality for the Exponential Family): Consider the sequence of i.i.d. random variables $\{X_i : i \in [n]\}$ distributed according to $\mathbb{P} \in \mathscr{P}_b$ with mean μ. Then, for any $\zeta \in \mathbb{R}_+$ and $\bar{X}_n \triangleq \frac{1}{n} \sum_{i=1}^{n} X_i$ we have

$$\mathbb{P}(\bar{X}_n - \mu \ge \zeta) \le \exp(-n d_{KL}(\mu + \zeta \| \mu)), \quad (43)$$

$$\text{and} \quad \mathbb{P}(\bar{X}_n - \mu \le -\zeta) \le \exp(-n d_{KL}(\mu - \zeta \| \mu)). \quad (44)$$

Proof: Let us define $S_n \triangleq \sum_{i=1}^{n} X_i$ and denote the moment generating function associated with $\mathbb{P}$ by $\phi(\lambda) \triangleq \mathbb{E}_{\mathbb{P}}[e^{\lambda X}]$, which for the single parameter exponential $\mathscr{P}_b$ is given by

$$\phi(\lambda) = \exp\left(b\left(\dot{b}^{-1}(\mu) + \lambda\right) - b\left(\dot{b}^{-1}(\mu)\right)\right). \tag{45}$$

Hence,

$$\mathbb{P}(\bar{X}_n - \mu \ge \zeta) = \mathbb{P}\left(e^{\lambda S_n} \ge e^{n\lambda(\mu + \zeta)}\right) \tag{46}$$

$$\le \frac{\mathbb{E}_{\boldsymbol{\mu}}[e^{\lambda S_n}]}{e^{n(\lambda + \zeta)}} \tag{47}$$

$$= \prod_{i=1}^{n} \left(\frac{\mathbb{E}_{\boldsymbol{\mu}}[e^{\lambda X_i}]}{e^{\lambda(\mu + \zeta)}}\right) \tag{48}$$

$$= \left(\frac{\phi(\lambda)}{e^{\lambda(\mu + \zeta)}}\right)^n, \tag{49}$$

where (47) is holds due to Markov's inequality and (48) holds due to independence. Since (49) is valid for any $\lambda \in \mathbb{R}_+$, we have

$$\mathbb{P}(\bar{X}_n - \mu \ge \zeta) \le \left(\inf_{\lambda \in \mathbb{R}_+} \frac{\phi(\lambda)}{e^{\lambda(\mu + \zeta)}}\right)^n. \tag{50}$$

It can be readily verified that the infimum of the right-hand-side in (50) is obtained at

$$\arg\inf_{\lambda \in \mathbb{R}_+} \frac{\phi(\lambda)}{e^{\lambda(\mu + \zeta)}} = \dot{b}^{-1}(\mu + \zeta) - \dot{b}^{-1}(\mu). \tag{51}$$

Using (51), we can rewrite (50) as

$$\mathbb{P}(\bar{X}_n - \mu \ge \zeta) \le \exp\left(-n\left[b\left(\dot{b}^{-1}(\mu)\right) - b\left(\dot{b}^{-1}(\mu + \zeta)\right)\right.\right.$$
$$\left.\left. - (\mu + \zeta)\left(\dot{b}^{-1}(\mu) - \dot{b}^{-1}(\mu + \zeta)\right)\right]\right) \tag{52}$$

$$= \exp(-n D_{KL}(\mu + \zeta \| \mu)). \tag{53}$$

where (53) holds due to (6). The proof of (44) follows a similar line of arguments. ∎

Lemma 9: For any pair of distributions $\mathbb{P}, \mathbb{P}' \in \mathscr{P}_b$ with mean values μ, μ', for any $\epsilon \in \mathbb{R}_+$,

$$d_{KL}(\mu + \epsilon \| \mu') - d_{KL}(\mu \| \mu') = O(\epsilon), \tag{54}$$

$$\text{and } d_{KL}(\mu - \epsilon \| \mu') - d_{KL}(\mu \| \mu') = O(\epsilon). \tag{55}$$

Proof: Note that $b : \Theta \mapsto \mathbb{R}$ is a convex, twice-differentiable function over a compact space Θ, and hence it is Lipschitz continuous. Let us define a Lipschitz constant, based on which

$$|b(\theta) - b(\theta')| \le \mathcal{L}_1|\theta - \theta'|, \quad \forall \theta, \theta' \in \Theta. \tag{56}$$

Corresponding to any $\theta \in \Theta$, $\ddot{b}(\theta)$ is the variance of the distribution $\mathbb{P}_\theta \in \mathscr{P}_b$, and hence, $\ddot{b}(\theta) > 0$. Let us define

$$\mathcal{L}_2 \triangleq \left(\min_{\theta \in \Theta} \ddot{b}(\theta)\right)^{-1}. \tag{57}$$

Using the mean value theorem, for any pair $\theta, \theta' \in \Theta$, there exists a $\lambda \in \Theta$ for which we have:

$$\frac{\dot{b}(\theta) - \dot{b}(\theta')}{\theta - \theta'} = \ddot{b}(\lambda) \ge \frac{1}{\mathcal{L}_2}. \tag{58}$$

Using (58), for any pair $z, z' \in \mathbb{R}$, we have

$$|\dot{b}^{-1}(z) - \dot{b}^{-1}(z')| \le \mathcal{L}_2|z - z'|. \tag{59}$$

This indicates that for any $\epsilon \in \mathbb{R}_+$, we have

$$d_{KL}(\mu + \epsilon \| \mu') - d_{KL}(\mu \| \mu')$$
$$\overset{(6)}{=} b\left(\dot{b}^{-1}(\mu)\right) - b\left(\dot{b}^{-1}(\mu + \epsilon)\right)$$
$$+ \mu\left[\dot{b}^{-1}(\mu + \epsilon) - \dot{b}^{-1}(\mu)\right] + O(\epsilon) \tag{60}$$

$$\overset{(56)}{\leq} \mathcal{L}_1 \left| \dot{b}^{-1}(\mu) - \dot{b}^{-1}(\mu + \epsilon) \right|$$
$$+ \mu \left[\dot{b}^{-1}(\mu + \epsilon) - \dot{b}^{-1}(\mu) \right] + O(\epsilon) \tag{61}$$

$$\overset{(59)}{\leq} (\mathcal{L}_1 + \mu)\mathcal{L}_2 \epsilon + O(\epsilon) \tag{62}$$

$$= O(\epsilon). \tag{63}$$

Proving (55) follows similar steps.   ■

APPENDIX B
PROOF OF LEMMA 3 AND LEMMA 6

A. Exponential Family of Bandits

For any $t \in \mathbb{N}$, for a given $\epsilon \in \mathbb{R}_+$ and any bandit realization with mean $\boldsymbol{\mu}$ we have

$$\mathbb{P}_{\boldsymbol{\mu}}\big(N_\epsilon^\mu > t\big)$$
$$= \sum_{s=t+1}^{\infty} \mathbb{P}_{\boldsymbol{\mu}}\big(N_\epsilon^\mu = s\big) \tag{64}$$

$$= \sum_{s=t+1}^{\infty} \mathbb{P}_{\boldsymbol{\mu}}\Big(\exists i \in [K] : |\mu_{s-1,i} - \mu_i| > \epsilon, \text{ and,}$$

$$\forall u \geq s, \forall i \in [K], |\mu_{u,i} - \mu_i| \leq \epsilon \Big) \tag{65}$$

$$\leq \sum_{s=t}^{\infty} \mathbb{P}_{\boldsymbol{\mu}}\big(\exists i \in [K] : |\mu_{s,i} - \mu_i| > \epsilon\big) \tag{66}$$

$$\leq \sum_{i \in [K]} \sum_{s=t}^{\infty} \mathbb{P}_{\boldsymbol{\mu}}\big(|\mu_{s,i} - \mu_i| > \epsilon\big) \tag{67}$$

$$= \sum_{i \in [K]} \sum_{s=t}^{\infty} \mathbb{P}_{\boldsymbol{\mu}}\Big(|\mu_{s,i} - \mu_i| > \epsilon, \ T_{s,i} \geq \sqrt{\frac{s}{K}} - 1\Big)$$

$$+ \underbrace{\sum_{i \in [K]} \sum_{s=t}^{\infty} \mathbb{P}_{\boldsymbol{\mu}}\Big(|\mu_{s,i} - \mu_i| > \epsilon, \ T_{s,i} < \sqrt{\frac{s}{K}} - 1\Big)}_{=0}$$
$$\tag{68}$$

$$= \sum_{i \in [K]} \sum_{s=t}^{\infty} \sum_{\ell = \sqrt{s/K} - 1}^{\infty} \mathbb{P}_{\boldsymbol{\mu}}\big(|\mu_{s,i} - \mu_i| > \epsilon, \ T_{s,i} = \ell\big) \tag{69}$$

$$\leq \sum_{i \in [K]} \sum_{s=t}^{\infty} \sum_{\ell = \sqrt{s/K} - 1}^{\infty} \Big(\exp\big(-\ell d_{KL}(\mu_i + \epsilon \| \mu_i) \big)$$
$$+ \exp\big(-\ell d_{KL}(\mu_i - \epsilon \| \mu_i) \big) \Big) \tag{70}$$

$$\leq \underbrace{\sum_{i \in [K]} \sum_{s=t}^{\infty} \int_{\sqrt{s/K}-2}^{\infty} \exp\{-\ell d_{KL}(\mu_i + \epsilon \| \mu_i)\} \, d\ell}_{\triangleq A_1}$$

$$+ \underbrace{\sum_{i \in [K]} \sum_{s=t}^{\infty} \int_{\sqrt{s/K}-2}^{\infty} \exp\{-\ell d_{KL}(\mu_i - \epsilon \| \mu_i)\} \, d\ell}_{\triangleq A_2}, \tag{71}$$

where (67) is obtained by the union bound, (68) and (69) use total probability and (70) is a result of using Lemma 8.

Furthermore, for A_1 we have

$$A_1 = \sum_{i \in [K]} \sum_{s=t}^{\infty} \int_{\sqrt{s/K}-2}^{\infty} \exp\{-\ell d_{KL}(\mu_i + \epsilon \| \mu_i)\} \, d\ell$$

$$= \sum_{i \in [K]} \frac{1}{d_{KL}(\mu_i + \epsilon \| \mu_i)}$$
$$\times \sum_{s=t}^{\infty} \exp\left(-\left(\sqrt{\frac{s}{K}} - 2\right) d_{KL}(\mu_i + \epsilon \| \mu_i)\right) \tag{72}$$

$$\leq \sum_{i \in [K]} \frac{\exp(2 d_{KL}(\mu_i + \epsilon \| \mu_i))}{d_{KL}(\mu_i + \epsilon \| \mu_i)}$$
$$\times \int_t^{\infty} \exp\left(-\sqrt{\frac{s}{K}} d_{KL}(\mu_i + \epsilon \| \mu_i)\right) ds \tag{73}$$

$$= \sum_{i \in [K]} 2K \frac{\exp(2 d_{KL}(\mu_i + \epsilon \| \mu_i))}{(d_{KL}(\mu_i + \epsilon \| \mu_i))^3}$$
$$\times \left(\frac{d_{KL}(\mu_i + \epsilon \| \mu_i)}{\sqrt{K}} \sqrt{t} + 1\right)$$
$$\times \exp\left\{-\frac{d_{KL}(\mu_i + \epsilon \| \mu_i)}{\sqrt{K}} \sqrt{t}\right\}, \tag{74}$$

where (73) is obtained by upper-bounding the summation over the index s by its integration. Consequently,

$$\sum_{t \in \mathbb{N}} A_1 \leq \sum_{i \in [K]} 2K \frac{\exp(2 d_{KL}(\mu_i + \epsilon \| \mu_i))}{(d_{KL}(\mu_i + \epsilon \| \mu_i))^3}$$
$$\times \int_0^{\infty} \left(\frac{d_{KL}(\mu_i + \epsilon \| \mu_i)}{\sqrt{K}} \sqrt{t} + 1\right)$$
$$\times \exp\left\{-\frac{d_{KL}(\mu_i + \epsilon \| \mu_i)}{\sqrt{K}} \sqrt{t}\right\} dt \tag{75}$$

$$= 12K^2 \sum_{i \in [K]} \frac{\exp(2 d_{KL}(\mu_i + \epsilon \| \mu_i))}{(d_{KL}(\mu_i + \epsilon \| \mu_i))^5} < +\infty. \tag{76}$$

Similarly, we can show that

$$\sum_{t \in \mathbb{N}} A_2 \leq 12K^2 \sum_{i \in [K]} \frac{\exp(2 d_{KL}(\mu_i - \epsilon \| \mu_i))}{(d_{KL}(\mu_i - \epsilon \| \mu_i))^5} < +\infty. \tag{77}$$

Finally, using (76) and (77), we obtain

$$\mathbb{E}_{\boldsymbol{\mu}}\big[N_\epsilon^\mu\big] = \sum_{t \in \mathbb{N}} \mathbb{P}_{\boldsymbol{\mu}}\big(N_\epsilon^\mu > t\big) \leq \sum_{t \in \mathbb{N}} (A_1 + A_2) < +\infty. \tag{78}$$

B. Gaussian Bandits

For the case of Gaussian bandits, by invoking the estimator concentration specified in Lemma 7 and the sufficiency of exploration established in Lemma 5, for all $i \in [K]$ and $\forall n > N_1$, where N_1 is specified in Lemma 7, we almost surely have

$$|\mu_{n,i} - \mu_i| \overset{(42)}{\leq} \sigma W_1 \sqrt{\frac{\log(e + T_{n,i})}{1 + T_{n,i}}} \tag{79}$$

$$\leq \sigma W_1 \sqrt{\frac{\log(e + \sqrt{n/K})}{1 + \sqrt{n/K}}} \tag{80}$$

$$\leq \sigma W_1 \sqrt{\frac{2(n/K)^{1/4}}{1 + \sqrt{n/K}}}, \tag{81}$$

where (80) is a result of Lemma 5, noting that $n > N_1$, and monotonicity of $\frac{\log(e+x)}{1+x}$, and (81) holds since $\log(e + \sqrt{n/K}) \leq 2(n/K)^{1/4}$ for $n > K$. Furthermore, owing to $\frac{x^{1/4}}{(1+\sqrt{x})}$ being a decreasing function in x, corresponding to any realization of W_1 there exists L_ϵ^μ, which is a function of W_1 and an instance-dependent parameter, such that it satisfies

$$\sqrt{\frac{2(L_\epsilon^\mu/K)^{1/4}}{1 + \sqrt{L_\epsilon^\mu/K}}} < \frac{\epsilon}{\sigma W_1}. \tag{82}$$

By defining $M_\epsilon^\mu \triangleq \max\{N_1, L_\epsilon^\mu\}$, from (81) and (82), we obtain that for all $n > M_\epsilon^\mu$

$$|\mu_{n,i} - \mu_i| \leq \epsilon, \quad \forall i \in [K]. \tag{83}$$

Since both N_1 and L_ϵ^μ are a function of W_1 and instance dependent parameters, we have $\mathbb{E}_{\boldsymbol{\mu}}[N_1] < +\infty$ and $\mathbb{E}_{\boldsymbol{\mu}}[L_\epsilon^\mu] < +\infty$, since, by Lemma 7, $\mathbb{E}_{\boldsymbol{\mu}}[W_1] < +\infty$. Finally, since $M_\epsilon^\mu < N_1 + L_\epsilon^\mu$, we have $\mathbb{E}_{\boldsymbol{\mu}}[M_\epsilon^\mu] < \mathbb{E}_{\boldsymbol{\mu}}[N_1] + \mathbb{E}_{\boldsymbol{\mu}}[L_\epsilon^\mu] < +\infty$.

APPENDIX C
PROOF OF LEMMA 4

The analysis for the convergence of the sampling proportions to the optimal allocation has two main steps.

Step 1 Convergence in allocation for $a^\star$: First, we show the convergence of the selection proportions allocated to the best arm $a^\star$ to the β-optimal allocation β. For this, using Lemma 3, we have that for all $n > N_{\Delta_{\min}/2}^\mu$, $a_n^1 = a^\star$. To show this, let us assume that $a_n^1 \neq a^\star$ for some $n > N_{\Delta_{\min}/2}^\mu$. We obtain,

$$\mu_{n,a_n^1} - \mu_{n,a^\star} \leq \mu_{a_n^1} - \mu_{a^\star} + \Delta_{\min} \leq 0 \tag{84}$$

where the first inequality holds since $n > N_{\Delta_{\min}/2}^\mu$, and the second one follows by noting that $\mu_{a_n^1} - \mu_{a^\star} \leq -\Delta_{\min}$. Since (84) contradicts the definition of a_n^1, we have that $a_n^1 = a^\star$ for all $n > N_{\Delta_{\min}/2}^\mu$. Following the same line of arguments as in [24, Lemma 12] and [24, Lemma 13], the above property means that there exists a random variable N_2^ϵ, which is a function of W_1, such that for all $n > N_2^\epsilon$, we have $|T_{n,a^\star}/n - \beta| \leq \epsilon$.

Step 2 Convergence in allocation for any $i \in [K] \setminus \{a^\star\}$: The other step is to show the convergence of the sampling probabilities of the other arms $i \in [K] \setminus \{a^\star\}$ to the optimal proportions $\{\omega_i^\star(\beta) : i \in [K] \setminus \{a^\star\}\}$. We begin by defining the set of over-sampled arms as follows. For any $\epsilon \in \mathbb{R}_+$ we define

$$\mathcal{O}_n(\epsilon) \triangleq \left\{i \in [K] \setminus \{a^\star\} : T_{n,i}/n > \omega_i^\star(\beta) + \epsilon\right\}. \tag{85}$$

Furthermore, we use the notation $\bar{\mathcal{O}}_n(\epsilon) \triangleq [K] \setminus \mathcal{O}_n(\epsilon)$ to denote the set of arms that are not over-sampled. The key idea for showing the convergence of the arms $i \in [K] \setminus \{a^\star\}$ is to show that after some time, the challenger is never contained in the over-sampled set. This implies that when the means of all the arms and the proportion of the best arm have converged sufficiently, the TT-SPRT sampling strategy never samples from the over-sampled set of arms. This, in turn, leads to the convergence in allocation of the arms $i \in [K] \setminus \{a^\star\}$. This idea is formalized in the following lemma, which has been proved in [24].

Lemma 10 [24]: For any $\epsilon > 0$, for any top-two sampling strategy, assume that the following conditions are satisfied:
1) There exists a stochastic time N_2^ϵ, such that for all $n > N_2^\epsilon$, we have convergence in allocation for the best arm, i.e.,

$$\left|\frac{T_{n,a^\star}}{n} - \beta\right| \leq \epsilon \quad \forall n > N_2^\epsilon. \tag{86}$$

2) There exists N_3^ϵ such that for all $n > N_3^\epsilon$, the challenger $a_n^2 \notin \mathcal{O}_n(\epsilon/2)$.

Then, there exists N_4^ϵ such that $\mathbb{E}_{\boldsymbol{\mu}}[N_4^\epsilon] < +\infty$, and $\mathcal{O}_n(\epsilon) = \emptyset$ for all $n > N_4^\epsilon$. Furthermore, for all $n > N_4^{\epsilon/K}$, we have convergence in allocation of every arm $i \in [K] \setminus \{a^\star\}$, i.e.,

$$\left|\frac{T_{n,a^\star}}{n} - \beta\right| \leq \epsilon \quad \forall n > N_4^{\epsilon/K}, \, \forall i \in [K]. \tag{87}$$

Leveraging Lemma 10, the proof is concluded by setting $N_\epsilon^\omega \triangleq N_4^{\epsilon/K}$. We have already shown that condition 1 in Lemma 10 is satisfied by TT-SPRT for all $n > N_2^\epsilon$. Next, we will show that condition 2 is also satisfied. Specifically, we will show that there exists N_3^ϵ such that for all $n > N_3^\epsilon$, for any $i \in \mathcal{O}_n(\epsilon/2)$ and any $j \in \bar{\mathcal{O}}_n(0)$ we have

$$\Lambda_n\left(a_n^1, i\right) > \Lambda_n\left(a_n^1, j\right). \tag{88}$$

By the definition of a_n^2, this implies that $a_n^2 \notin \mathcal{O}_n(\epsilon/2)$. Next, we will show that (88) holds. For any $i \in [K] \setminus \{a^\star\}$, let us define

$$C_i\left(\beta, \omega_i^\star(\beta)\right) \triangleq \beta d_{KL}\left(\mu_{a^\star} \| \mu_{a^\star,i}\right) + \omega_i^\star(\beta) d_{KL}\left(\mu_i \| \mu_{a^\star,i}\right), \tag{89}$$

where $\mu_{a^\star,i} \triangleq (\beta\mu_{a^\star} + \omega_i^\star(\beta)\mu_i)/(\beta + \omega_i^\star(\beta))$. Next, we state a proposition characterizing the optimal allocation.

Proposition 1 [25, Proposition 1]: There exists a unique optimal allocation $\boldsymbol{\omega}^\star(\beta)$ to the optimization problem in (9), such that for any $i, j \in [K] \setminus \{a^\star\}$, $C_i(\beta, \omega_i^\star(\beta)) = C_j(\beta, \omega_j^\star(\beta))$. Furthermore,

$$C_i\left(\beta, \omega_i^\star(\beta)\right) = \min_{x \in [\mu_i, \mu_{a^\star}]} \beta d_{KL}(\mu_{a^\star} \| x) + \omega_i^\star(\beta) d_{KL}(\mu_i \| x). \tag{90}$$

Let us set $\zeta = \epsilon/2$. Leveraging Lemma 3, we obtain that for all $n > N_{\zeta^2}^\mu$, $|\mu_{n,i} - \mu_i| < \zeta^2$ for all $i \in [K]$. Next, for all $n > N_2^{\zeta^2}$, we have $|T_{n,a^\star}/n - \beta| < \zeta^2$. Let us define $N_3^\zeta \triangleq \max\{N_{\zeta^2}^\mu, N_2^{\zeta^2}, N_{\Delta_{\min}/2}^\mu\}$. Thus, for all $n > N_3^\zeta$, for any $i \in \mathcal{O}_n(\zeta)$ and for any $j \in \bar{\mathcal{O}}_n(0)$:

$$\frac{1}{n}\left(\Lambda_n\left(a_n^1, i\right) - \Lambda_n\left(a_n^1, j\right)\right)$$
$$= \frac{1}{n}\left(\Lambda_n\left(a^\star, i\right) - \Lambda_n\left(a^\star, j\right)\right) \tag{91}$$
$$= \underbrace{\frac{1}{n}\Lambda_n\left(a^\star, i\right) - C_i\left(\beta, \omega_i^\star(\beta)\right)}_{\triangleq A_3}$$
$$+ \underbrace{C_j\left(\beta, \omega_j^\star(\beta)\right) - \frac{1}{n}\Lambda_n\left(a_n^1, j\right)}_{\triangleq A_4}, \tag{92}$$

where (91) holds since $a_n^1 = a^\star$ for all $n > N_3^\zeta$, and (92) is obtained by leveraging Proposition 1. Expanding A_3, we obtain

$$A_3 = \frac{1}{n}\Lambda_n(a^\star, i) - C_i(\beta, \omega_i^\star(\beta))$$
$$\geq (\beta - \zeta^2)d_{KL}(\mu_{n,a^\star}\|\mu_{n,a^\star,i}) - \beta d_{KL}(\mu_{a^\star}\|\mu_{a^\star,i})$$
$$+ (\omega_i^\star(\beta) + \zeta)d_{KL}(\mu_{n,i}\|\mu_{n,a^\star,i}) - \omega_i^\star(\beta)d_{KL}(\mu_i\|\mu_{a^\star,i}), \quad (93)$$

where (93) holds due to the definition of N_3^ζ, along with the fact that $i \in \mathcal{O}_n(\zeta)$ and $j \in \bar{\mathcal{O}}_n(0)$. Now, (93) can be lower-bounded as

$$(\beta - \zeta^2)d_{KL}(\mu_{n,a^\star}\|\mu_{n,a^\star,i}) - \beta d_{KL}(\mu_{a^\star}\|\mu_{a^\star,i})$$
$$+ (\omega_i^\star(\beta) + \zeta)d_{KL}(\mu_{n,i}\|\mu_{n,a^\star,i}) - \omega_i^\star(\beta)d_{KL}(\mu_i\|\mu_{a^\star,i}) \quad (94)$$
$$\geq (\beta - \zeta^2)d_{KL}(\mu_{a^\star} - \zeta^2\|\mu_{n,a^\star,i}) - \beta d_{KL}(\mu_{a^\star}\|\mu_{a^\star,i})$$
$$+ (\omega_i^\star(\beta) + \zeta)d_{KL}(\mu_i + \zeta^2\|\mu_{n,a^\star,i}) - \omega_i^\star(\beta)d_{KL}(\mu_i\|\mu_{a^\star,i}) \quad (95)$$
$$\geq (\beta - \zeta^2)d_{KL}(\mu_{a^\star} - \zeta^2\|\mu_{n,a^\star,i}) - \beta d_{KL}(\mu_{a^\star}\|\mu_{n,a^\star,i})$$
$$+ (\omega_i^\star(\beta) + \zeta)d_{KL}(\mu_i + \zeta^2\|\mu_{n,a^\star,i}) - \omega_i^\star(\beta)d_{KL}(\mu_i\|\mu_{n,a^\star,i}) \quad (96)$$
$$= (\beta - \zeta^2)\Big[d_{KL}(\mu_{a^\star}\|\mu_{n,a^\star,i}) + O(\zeta^2)\Big] - \beta d_{KL}(\mu_{a^\star}\|\mu_{n,a^\star,i})$$
$$+ (\omega_i^\star(\beta) + \zeta)\Big[d_{KL}(\mu_i\|\mu_{n,a^\star,i}) + O(\zeta^2)\Big]$$
$$- \omega_i^\star(\beta)d_{KL}(\mu_i\|\mu_{n,a^\star,i}) \quad (97)$$
$$= \zeta d_{KL}(\mu_i\|\mu_{n,a^\star,i}) + O(\zeta^2) \quad (98)$$

where (95) follows from the facts that $\mu_{n,a^\star,i} \in [\mu_{n,i}, \mu_{n,a^\star}]$ and $|\mu_{n,i} - \mu_i| \leq \zeta^2$ for every $n > N_3^\zeta$, (96) follows from Proposition 1 by noting that $\beta d_{KL}(\mu_{a^\star}\|\mu_{n,a^\star,i}) + \omega_i^\star(\beta)D_{KL}(\mu_i\|\mu_{n,a^\star,i}) \geq \beta d_{KL}(\mu_{a^\star}\|\mu_{a^\star,i}) + \omega_i^\star(\beta)D_{KL}(\mu_i\|\mu_{a^\star,i})$, and (97) is a result of Lemma 9. Similarly, expanding A_4, we obtain

$$A_4 = \frac{1}{n}\Lambda_n(a_n^1, j) - C_j(\beta, \psi_j^\beta)$$
$$= \frac{T_{n,a^\star}}{n}d_{KL}(\mu_{n,a^\star}\|\mu_{n,a^\star,j})$$
$$+ \frac{T_{n,j}}{n}d_{KL}(\mu_{n,j}\|\mu_{n,a^\star,j}) - C_j(\beta, \omega_j^\star(\beta)) \quad (99)$$
$$\leq \frac{T_{n,a^\star}}{n}d_{KL}(\mu_{n,a^\star}\|\mu_{a^\star,j})$$
$$+ \frac{T_{n,j}}{n}d_{KL}(\mu_{n,j}\|\mu_{a^\star,j}) - C_j(\beta, \omega_j^\star(\beta)) \quad (100)$$
$$\leq (\beta + \zeta^2)d_{KL}(\mu_{n,a^\star}\|\mu_{a^\star,j}) - \beta d_{KL}(\mu_a^\star\|\mu_{a^\star,j})$$
$$+ \omega_j^\star(\beta)d_{KL}(\mu_{n,j}\|\mu_{a^\star,j}) - \omega_j^\star(\beta)d_{KL}(\mu_j\|\mu_{a^\star,j}) \quad (101)$$
$$\leq (\beta + \zeta^2)d_{KL}(\mu_{a^\star} + \zeta^2\|\mu_{a^\star,j}) - \beta d_{KL}(\mu_{a^\star}\|\mu_{a^\star,j})$$
$$+ \omega_j^\star(\beta)d_{KL}(\mu_j - \zeta^2\|\mu_{a^\star,j}) - \omega_j^\star(\beta)d_{KL}(\mu_j\|\mu_{a^\star,j}) \quad (102)$$
$$\leq (\beta + \zeta^2)\Big[d_{KL}(\mu_{a^\star}\|\mu_{a^\star,j}) + O(\zeta^2)\Big] - \beta d_{KL}(\mu_{a^\star}\|\mu_{a^\star,j})$$
$$+ \omega_j^\star(\beta)\Big[d_{KL}(\mu_j\|\mu_{a^\star,j}) + O(\zeta^2)\Big]$$
$$- \omega_j^\star(\beta)d_{KL}(\mu_j\|\mu_{a^\star,j}) \quad (103)$$
$$= O(\zeta^2) \quad (104)$$

where (100) holds due to the fact that [6]

$$\mu_{n,a^\star,i} \triangleq \arg\min_{x\in[\mu_{n,i},\mu_{n,a^\star}]} \frac{T_{n,a^\star}}{n}d_{KL}(\mu_{n,a^\star}\|x)$$
$$+ \frac{T_{n,j}}{n}d_{KL}(\mu_{n,j}\|x) \quad (105)$$

Equation (101)–(102) holds due to the fact that $n > N_3^\zeta$, and (103) is obtained by leveraging Lemma 9. Finally, using (98) and (104), (92) can be rewritten as

$$\frac{1}{n}\Big(\Lambda_n(a_n^1, i) - \Lambda_n(a_n^1, j)\Big)$$
$$\geq \zeta d_{KL}(\mu_i\|\mu_{n,a^\star,i}) + O(\zeta^2) > 0. \quad (106)$$

APPENDIX D
PROOF OF THEOREM 3 AND THEOREM 5

In this section, we start by providing the proof of Theorem 3 for the case of the exponential family of bandits. We begin by defining the function $\mathcal{C}_{\exp}$ used in characterizing the stopping threshold in Theorem 2. For this purpose, for any $u \geq 1$, let us define the function $h(u) \triangleq u - \log u$. Furthermore, let us define the function $\tilde{h}_z(x) : \mathbb{R}_+ \mapsto \mathbb{R}_+$ for any $z \in [1, e]$ as

$$h_z(x) \triangleq \begin{cases} e^{\frac{1}{h^{-1}(x)}}h^{-1}(x), & \text{if } x \geq h(1/\log z) \\ z(x - \log\log z), & \text{otherwise} \end{cases} \quad (107)$$

The function $\mathcal{C}_{\exp} : \mathbb{R}_+ \mapsto \mathbb{R}_+$ is defined as

$$\mathcal{C}_{\exp}(x) \triangleq 2\tilde{h}_{3/2}\left(\frac{h^{-1}(1 + x) + \log(2\zeta(2))}{2}\right) \quad (108)$$

where $\zeta(s) \triangleq \sum_{n=1}^\infty n^{-s}$. Let us define the stochastic time T_β^ϵ, which marks the convergence of the arm means and the sampling proportions,

$$T_\beta^\epsilon \triangleq \inf\Big\{n_0 \in \mathbb{N} : |\mu_{n,i} - \mu_i| \leq \epsilon,$$
$$|T_{n,i}/n - \omega_i^\star(\beta)| \leq \epsilon \ \forall i \in [K], \ \forall n > n_0\Big\}. \quad (109)$$

Leveraging Lemma 3, we have $\mathbb{E}_{\boldsymbol{\mu}}[N_\epsilon^\mu] < +\infty$, and $|\mu_{n,i} - \mu_i| \leq \epsilon$ for all $n > N_\epsilon^\mu$ and for every $i \in [K]$. Furthermore, in Lemma 4 we showed that for any $\epsilon > 0$, there exists N_ϵ^ω, $\mathbb{E}_{\boldsymbol{\mu}}[N_\epsilon^\omega] < \infty$ and for all $n > N_\epsilon^\omega$ and $i \in [K]$ we have $|T_{n,i}/n - \omega_i^\star(\beta)| < \epsilon$. Furthermore, by the definition of T_β^ϵ in (109), $T_\beta^\epsilon = \max\{N_\epsilon^\mu, N_\epsilon^\omega\}$, and thus $\mathbb{E}_{\boldsymbol{\mu}}[T_\beta^\epsilon] < +\infty$. Finally, the theorem is proved by leveraging Lemma 11 stated below.

Lemma 11: For the combination of any arm selection rule that satisfies $\mathbb{E}_{\boldsymbol{\mu}}[T_\beta^\epsilon] < +\infty$ for any $\epsilon > 0$, and the stopping rule specified in (27), we have

$$\lim_{\delta\to 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}[\tau]}{\log(1/\delta)} \leq \frac{1}{\Gamma_{\boldsymbol{\mu}}(\beta)}. \quad (110)$$

Proof: Note that by the definition of T_ϵ^β, using the same argument as in (84), we have $a_n^1 = a^\star$ for all $n \geq T_\beta^{\Delta_{\min}/2}$. Thus, for all $n \geq T_\beta^{\Delta_{\min}/2}$, we have

$$\frac{1}{n}\Lambda_n(a_n^1, a_n^2) \overset{(20)}{=} \min_{j\in[K]\setminus\{a^\star\}} \frac{T_{n,a^\star}}{n}d_{KL}(\mu_{n,a^\star}\|\mu_{n,a^\star,j})$$
$$+ \frac{T_{n,j}}{n}d_{KL}(\mu_{n,j}\|\mu_{n,a^\star,j}). \quad (111)$$

Furthermore, recall the definition of $\Gamma_{\boldsymbol{\mu}}(\beta)$,

$$\Gamma_{\boldsymbol{\mu}}(\beta) \triangleq \min_{i \in [K] \setminus a^\star} \beta d_{KL}\big(\mu_{a^\star} \| \mu_{a^\star, i}\big) \\ + \omega_i^\star(\beta) d_{KL}\big(\mu_i \| \mu_{a^\star, i}\big), \quad (112)$$

based on which, given $\epsilon > 0$, there exist $\epsilon' \in (0, \Delta_{\min}/2]$ and $T^\epsilon \triangleq T_\beta^{\epsilon'}$, such that for all we have $n > T^\epsilon$, $|\mu_{n,i} - \mu_i| \leq \epsilon'$, $|T_{n,i}/n - \omega_i^\star(\beta)| \leq \epsilon'$, and $|\Lambda(a_n^1, a_n^2)/n - \Gamma_{\boldsymbol{\mu}}(\beta)| < \epsilon$ [6], [24]. Furthermore, $\mathbb{E}_{\boldsymbol{\mu}}[T^\epsilon] = \mathbb{E}_{\boldsymbol{\mu}}[T_\beta^{\epsilon'}] < +\infty$. Furthermore, for all $t > 3 \times 10^7$, we have

$$6 \log\left(\log \frac{t}{2} + 1\right) \leq \log t. \quad (113)$$

Let us define $T_0^\epsilon \triangleq \max\{T^\epsilon, 3 \times 10^7\}$. Next, expanding the time instant right before stopping, we have

$$\tau - 1 = (\tau - 1)\mathbb{1}_{\{\tau-1 \leq T_0^\epsilon\}} + (\tau - 1)\mathbb{1}_{\{\tau-1 > T_0^\epsilon\}} \quad (114)$$

$$\leq T_0^\epsilon + (\tau - 1)\mathbb{1}_{\{\tau-1 > T_0^\epsilon\}}. \quad (115)$$

Now, due to the choice of our stopping rule in (23) along with the choice of threshold defined in Theorem 2, at $(\tau - 1)$, we have

$$\Lambda_{\tau-1}\big(a_1^{\tau-1}, a_2^{\tau-1}\big) \\ \leq \log \frac{K-1}{\delta} + 8 \log\left(1 + \frac{1}{2} \log \frac{K-1}{\delta} + \sqrt{\log \frac{K-1}{\delta}}\right) \\ + 6 \log\left(\log \frac{\tau-1}{2} + 1\right). \quad (116)$$

Furthermore, when $\tau - 1 > T_0^\epsilon$,

$$\Lambda_{\tau-1}\big(a_1^{\tau-1}, a_2^{\tau-1}\big) \geq (\tau - 1)\big(\Gamma_{\boldsymbol{\mu}}(\beta) - \epsilon\big). \quad (117)$$

Combining (116) and (117), when $\tau - 1 > T_0^\epsilon$, we have

$$(\tau - 1)\big(\Gamma_{\boldsymbol{\mu}}(\beta) - \epsilon\big) \\ \leq \log \frac{K-1}{\delta} + 8 \log\left(1 + \frac{1}{2} \log \frac{K-1}{\delta} + \sqrt{\log \frac{K-1}{\delta}}\right) \\ + 6 \log\left(\log \frac{\tau-1}{2} + 1\right) \quad (118)$$

$$\leq \log \frac{K-1}{\delta} + 8 \log\left(1 + \frac{1}{2} \log \frac{K-1}{\delta} + \sqrt{\log \frac{K-1}{\delta}}\right) \\ + \log(\tau - 1). \quad (119)$$

Furthermore, note that $f(x) \triangleq x - \frac{1}{D_1} \log D_2 x$ is a monotonically increasing function in x for $D_1, D_2 \in \mathbb{R}_+$. Thus, there exists $x_{\max}$ such that for all $x \geq x_{\max}$, $f(x) \geq 0$. Next, we will find an $\bar{x}$ such that $f(\bar{x}) \geq 0$. In our case, we have $D_1 = (\Gamma_{\boldsymbol{\mu}}(\beta) - \epsilon)$, and $D_2 = ((K-1)g(\delta))/\delta$, where we have defined

$$g(\delta) \triangleq \left(1 + \frac{1}{2} \log \frac{K-1}{\delta} + \sqrt{\log \frac{K-1}{\delta}}\right)^8. \quad (120)$$

We leverage [6, Lemma 18], which yields that

$$\tau - 1 \leq \frac{1}{\Gamma_{\boldsymbol{\mu}}(\beta) - \epsilon}\left[\log\left\{\frac{g(\delta)(K-1)e}{(\Gamma_{\boldsymbol{\mu}}(\beta) - \epsilon)\delta}\right\} \\ + \log\log\left\{\frac{g(\delta)(K-1)}{(\Gamma_{\boldsymbol{\mu}}(\beta) - \epsilon)\delta}\right\}\right]. \quad (121)$$

Thus, combining (115) and (121), we obtain

$$\tau \leq T_0^\epsilon + \frac{1}{\Gamma_{\boldsymbol{\mu}}(\beta) - \epsilon}\left[\log\left\{\frac{g(\delta)(K-1)e}{(\Gamma_{\boldsymbol{\mu}}(\beta) - \epsilon)\delta}\right\} \\ + \log\log\left\{\frac{g(\delta)(K-1)}{(\Gamma_{\boldsymbol{\mu}}(\beta) - \epsilon)\delta}\right\}\right] + 1. \quad (122)$$

Finally, taking the expectation on both sides of (122), dividing by $\log(1/\delta)$, and taking the limit of $\delta \to 0$, we obtain

$$\lim_{\delta \to 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}[\tau]}{\log(1/\delta)} \leq \frac{1}{\Gamma_{\boldsymbol{\mu}}(\beta) - \epsilon}, \quad (123)$$

where (123) is obtained by noting that $\lim_{\delta \to 0} \log g(\delta)/\log(1/\delta) = 0$, and we have used the fact that $\mathbb{E}_{\boldsymbol{\mu}}[T_0^\epsilon] < +\infty$. The proof for the exponential family of bandits is completed by taking an infimum with respect to ϵ in (123).

The proof for the Gaussian setting involves a counterpart of Lemma 11, which has been proved in [25, Lemma 1]. Specifically, [25, Lemma 1] states that for Gaussian bandits, the combination of any arm selection rule that satisfies $\mathbb{E}_{\boldsymbol{\mu}}[T_\beta^\epsilon] < +\infty$, where we have defined T_β^ϵ in (109), along with the stopping rule specified in (33), satisfies

$$\limsup_{\delta \to 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}[\tau]}{\log(1/\delta)} \leq \frac{1}{\Gamma_{\boldsymbol{\mu}}(\beta)}. \quad (124)$$

The proof is concluded by leveraging Lemma 3 and Lemma 4, ensuring $\mathbb{E}_{\boldsymbol{\mu}}[T_\beta^\epsilon] < +\infty$. ∎

APPENDIX E
PROOF OF THEOREM 6

Let us define

$$T_i^n \triangleq \inf\left\{s \in \mathbb{N}: \theta_{s,i}^n > \theta_{s,b_s^1}^n\right\} \quad (125)$$

based on which $T_{TTTS}^n = \min_{i \in [K] \setminus \{b_n^1\}} T_i^n$. Note that using [25, Lemma 12], there exists a random variable M_0, such that for all $n > M_0$, we have $b_n^1 = a^\star$. Similarly, based on [25, Sec. C.2] there exists a random variable $M_{\Delta_{\min}/4}$ such that for all $n > M_{\Delta_{\min}/4}$, we almost surely have

$$|\mu_{n,i} - \mu_i| \leq \frac{\Delta_{\min}}{4} \quad \forall i \in [K]. \quad (126)$$

Defining $N_0 \triangleq \max(M_0, M_{\Delta_{\min}/4})$, for all $n > N_0$, $s \in \mathbb{N}$, and for any $i \in [K] \setminus \{a^\star\}$ we have

$$\mathbb{P}\big(\theta_{s,i}^n > \theta_{s,a^\star}^n\big) \\ \leq \frac{1}{2} \exp\left(-\frac{(\mu_{n,a^\star} - \mu_{n,i})^2}{2\sigma^2}\left(\frac{1}{T_{n,i}} + \frac{1}{T_{n,a^\star}}\right)^{-1}\right) \quad (127)$$

$$\leq \frac{1}{2}\exp\left(-\frac{(\mu_{a^\star}-\mu_i-\Delta_{\min}/2)^2}{2\sigma^2}\left(\frac{1}{T_{n,i}}+\frac{1}{T_{n,a^\star}}\right)^{-1}\right) \tag{128}$$

$$\leq \frac{1}{2}\exp\left(-\sqrt{\frac{n}{K}}\frac{(\mu_{a^\star}-\mu_i-\Delta_{\min}/2)^2}{4\sigma^2}\right) \tag{129}$$

where (127) is obtained by leveraging [24, Lemma 1] while noting that for any two arms $i,j\in[K]$ and $n\in\mathbb{N}$, $(\theta^n_{s,i}-\theta^n_{s,j}) \sim \mathcal{N}(\mu_{n,i}-\mu_{n,j},\sigma^{i,j}_n)$, where we have defined $\sigma^{i,j}_n \triangleq \sqrt{\sigma^2(1/T_{n,i}+1/T_{n,j})}$, (128) is a result of (126), and (129) is a result of [25, Lemma 5], which states that under TTTS, there exists a random variable M_1 such that for all $n>M_1$, we have $T_{n,i}\geq\sqrt{n/K}$ for every $i\in[K]$. Furthermore, using [24, Lemma 1], we also obtain that

$$\mathbb{P}\left(\theta^n_{s,i}>\theta^n_{s,a^\star}\right)\geq\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{\left(\sigma^{a^\star,i}_n+\mu_{n,a^\star}-\mu_{n,i}\right)^2}{2\sigma^2}\right. $$
$$\left.\times\left(\frac{1}{T_{n,i}}+\frac{1}{T_{n,a^\star}}\right)^{-1}\right). \tag{130}$$

Following the same line of arguments as in (127)-(129), and noting that $T_{n,i}\leq n$ for every $i\in[K]$, from (130) we obtain

$$\mathbb{P}\left(\theta^n_{s,i}>\theta^n_{s,a^\star}\right)\geq\frac{1}{\sqrt{2\pi e}}\exp\left(-\sqrt{n}\left(\mu_{a^\star}-\mu_i+\frac{\Delta_{\min}}{2}\right)\right.$$
$$\left.\times\left(\sqrt{n}\,\frac{\mu_{a^\star}-\mu_i+\Delta_{\min}/2}{4\sigma^2}+\frac{1}{\sqrt{2}\sigma}\right)\right). \tag{131}$$

Next, note that for any $n>N_0$, we have

$$\mathbb{P}\left(T^n_i>s\right)$$
$$=1-\mathbb{P}\left(T^n_i\leq s\right) \tag{132}$$
$$=1-\sum_{\ell=1}^{s}\mathbb{P}\left(\forall m<\ell\,,\,\theta^n_{m,i}\leq\theta^n_{m,a^\star}\,,\,\theta^n_{\ell,i}>\theta^n_{\ell,a^\star}\right) \tag{133}$$
$$=1-\sum_{\ell=1}^{s}\mathbb{P}\left(\theta^n_{\ell,i}>\theta^n_{\ell,a^\star}\right)\prod_{m=1}^{\ell}\mathbb{P}\left(\theta^n_{m,i}\leq\theta^n_{m,a^\star}\right) \tag{134}$$

where (133) follows from the definition of T^n_i in (125), and (134) holds since the samples are drawn independently from the posterior Π_n. For any $m\in\mathbb{N}$, let us define $p_{n,i}\triangleq\mathbb{P}(\theta^n_{m,i}\leq\theta^n_{m,a^\star})$. Subsequently, (134) can be simplified as follows

$$\mathbb{P}\left(T^n_i>s\right)=1-\left(1-p_{n,i}\right)\sum_{\ell=1}^{s}\left(p_{n,i}\right)^{\ell-1} \tag{135}$$
$$=1-\left(1-\left(p_{n,i}\right)^s\right) \tag{136}$$
$$\geq\left(1-\varepsilon_{n,i}\right)^s \tag{137}$$

where we have defined

$$\varepsilon_{n,i}\triangleq\frac{1}{2}\exp\left(-\sqrt{\frac{n}{K}}\frac{(\mu_{a^\star}-\mu_i-\Delta_{\min}/2)^2}{4\sigma^2}\right) \tag{138}$$

and (137) is a result of (129). Finally, we have

$$\mathbb{E}_n\left[T^n_i\right]=\sum_{s=0}^{\infty}\mathbb{P}\left(T^n_i>s\right)\overset{(137)}{\geq}\sum_{s=0}^{\infty}\left(1-\varepsilon_{n,i}\right)^s=\frac{1}{\varepsilon_{n,i}}. \tag{139}$$

Similarly, an upper bound on $\mathbb{E}_{\boldsymbol{\mu}}[T^n_i]$ can be obtained by following the same line of arguments as (132)-(136), and subsequently, upper-bounding $\mathbb{P}(T^n_i>s)$ by $(1-\varepsilon'_{n,i})^s$, where

$$\varepsilon'_{n,i}\triangleq\frac{1}{\sqrt{2\pi e}}\exp\left(-\sqrt{n}\left(\mu_{a^\star}-\mu_i+\frac{\Delta_{\min}}{2}\right)\right.$$
$$\left.\times\left(\sqrt{n}\frac{\mu_{a^\star}-\mu_i+\Delta_{\min}/2}{4\sigma^2}+\frac{1}{\sqrt{2}\sigma}\right)\right), \tag{140}$$

which follows from (131). This yields

$$\mathbb{E}_n\left[T^n_i\right]\leq\sqrt{2\pi e}\exp\left(\sqrt{n}\left(\mu_{a^\star}-\mu_i+\frac{\Delta_{\min}}{2}\right)\right.$$
$$\left.\times\left(\sqrt{n}\frac{\mu_{a^\star}-\mu+\Delta_{\min}/2}{4\sigma^2}+\frac{1}{\sqrt{2}\sigma}\right)\right). \tag{141}$$

Note that for all $n>32/(9\Delta^2_{\min})$, we have

$$\frac{n(\mu_{a^\star}-\mu_i+\Delta_{\min}/2)^2}{4\sigma^2}>\frac{\sqrt{n}(\mu_{a^\star}-\mu_i+\Delta_{\min}/2)}{\sqrt{2}\sigma}. \tag{142}$$

Combining (141) and (142), we obtain the upper-bound in Theorem 6. Next, we prove that there exists $\delta(N_0)>0$ such that for any $\delta\in(0,\delta(N_0))$, the stopping time of the TTTS algorithm, denoted by τ_{TTTS}, almost surely satisfies $\tau_{TTTS}>N_0$. Note that τ_{TTTS} is defined as

$$\tau_{TTTS}\triangleq\inf\left\{n\in\mathbb{N}\colon\Lambda_n\left(b^n_1,b^n_2\right)>c_{n,\delta}\right\} \tag{143}$$

where $c_{n,\delta}$ has been defined in (33). Let us set

$$\delta(N_0)\triangleq(K-1)\exp\left(-\left(\frac{N_0(\Delta_{\max}+\Delta_{\min}/2)}{8\sigma^2}\right.\right.$$
$$\left.\left.-2\log(4+\log N_0)\right)\right). \tag{144}$$

For any $\delta\in(0,\delta(N_0))$, the GLLR almost surely satisfies

$$\forall n\in\{1,\ldots,N_0\}\quad\Lambda_n\left(b^n_1,b^n_2\right)<c_{n,\delta}. \tag{145}$$

Combining (143) and (145), for any $\delta\in(0,\delta(N_0))$, we almost surely have $\tau_{TTTS}>N_0$.

REFERENCES

[1] S. Bubeck, R. Munos, and G. Stoltz, "Pure exploration in multi-armed bandits problems," in *Proc. Int. Conf. Algorithmic Learning Theory*, Porto, Portugal, Oct. 2009, pp. 23–37.

[2] M. Hoffman, B. Shahriari, and N. Freitas, "On correlation and budget constraints in model-based bandit optimization with application to automatic machine learning," in *Proc. Int. Conf. Artif. Intell. Statist.*, Reykjavik, Iceland, Apr. 2014, pp. 365–374.

[3] J. Katz-Samuels, L. Jain, Z. Karnin, and K. G. Jamieson, "An empirical process approach to the union bound: Practical algorithms for combinatorial and linear bandits," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2020, pp. 10371–10382.

[4] V. Gabillon, M. Ghavamzadeh, and A. Lazaric, "Best arm identification: A unified approach to fixed budget and fixed confidence," in *Proc. Adv. Neural Inf. Process. Syst.*, Lake Tahoe, NV, USA, Dec. 2012, pp. 3221–3229.

[5] S. Kalyanakrishnan, A. Tewari, P. Auer, and P. Stone, "PAC subset selection in stochastic multi-armed bandits," in *Proc. Int. Conf. Mach. Learn.*, Madison, WI, USA, Jun. 2012, pp. 655–662.

[6] A. Garivier and E. Kaufmann, "Optimal best arm identification with fixed confidence," in *Proc. Conf. Learn. Theory*, New York, NY, USA, Jun. 2016, pp. 998–1027.

[7] L. Xu, J. Honda, and M. Sugiyama, "A fully adaptive algorithm for pure exploration in linear bandits," in *Proc. Int. Conf. Artif. Intell. Statist.*, Apr. 2018, pp. 843–851.

[8] K. Jamieson, M. Malloy, R. Nowak, and S. Bubeck, "lil' UCB: An optimal exploration algorithm for multi-armed bandits," in *Proc. Conf. Learn. Theory*, Barcelona, Spain, Jun. 2014, pp. 423–439.

[9] Z. Karnin, T. Koren, and O. Somekh, "Almost optimal exploration in multi-armed bandits," in *Proc. Int. Conf. Mach. Learn.*, Atlanta, GA, USA, Jun. 2013, pp. 1238–1246.

[10] Y. Jedra and A. Proutiere, "Optimal best-arm identification in linear bandits," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2020, pp. 10007–10017.

[11] R. Degenne, W. M. Koolen, and P. Ménard, "Non-asymptotic pure exploration by solving games," in *Proc. Adv. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, Dec. 2019, pp. 14465–14474.

[12] R. Degenne, P. Menard, X. Shang, and M. Valko, "Gamification of pure exploration for linear bandits," in *Proc. 37th Int. Conf. Mach. Learn.*, Vienna, Austria, Jul. 2020, pp. 2432–2442.

[13] P.-A. Wang, R.-C. Tzeng, and A. Proutière, "Fast pure exploration via frank-wolfe," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2021, pp. 5810–5821.

[14] K. Jamieson and R. Nowak, "Best-arm identification algorithms for multi-armed bandits in the fixed confidence setting," in *Proc. Annu. Conf. Inf. Sci. Syst. (CISS)*, Princeton, NJ, USA, Mar. 2014, pp. 1–6.

[15] E. Even-Dar, S. Mannor, and Y. Mansour, "Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems," *J. Mach. Learn. Res.*, vol. 7, no. 39, pp. 1079–1105, 2006.

[16] T. Fiez, L. Jain, K. Jamieson, and L. Ratliff, "Sequential experimental design for transductive linear bandits," in *Proc. Adv. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, Nov. 2019, pp. 10666–10676.

[17] M. Soare, A. Lazaric, and R. Munos, "Best-arm identification in linear bandits," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, Montreal, QC, Canada, Dec. 2014, pp. 828–836.

[18] C. Tao, S. Blanco, and Y. Zhou, "Best arm identification in linear bandits with linear dimension dependency," in *Proc. Int. Conf. Machine Learning*, Stockholm, Sweden, Jul. 2018, pp. 4877–4886.

[19] S. Agrawal, S. Juneja, and P. Glynn, "Optimal δ-correct best-arm selection for heavy-tailed distributions," in *Proc. Int. Conf. Algorithmic Learn. Theory*, San Diego, CA, USA, Feb. 2020, pp. 61–110.

[20] D. Russo, "Simple bayesian algorithms for best-arm identification," *Oper. Res.*, vol. 68, no. 6, pp. 1625–1647, Apr. 2020.

[21] W. R. Thompson, "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples," *Biometrika*, vol. 25, nos. 3–4, pp. 285–294, 1933.

[22] L. Tang, R. Rosales, A. Singh, and D. Agarwal, "Automatic ad format selection via contextual bandits," in *Proc. CIKM*, San Francisco, CA, USA, Oct. 2013, pp. 1587–1594.

[23] S. Agarwal and N. Goyal, "Analysis of thompson sampling for the multi-armed bandit problem," in *Proc. Annu. Conf. Learn. Theory*, Edinburgh, Scotland, Jun. 2012, p. 39.

[24] C. Qin, D. Klabjan, and D. Russo, "Improving the expected improvement algorithm," in *Proc. Adv. Neural Inf. Process. Syst.*, Long Beach, CA, USA, Dec. 2017, pp. 5381–5391.

[25] X. Shang, R. de Heide, P. Menard, E. Kaufmann, and M. Valko, "Fixed-confidence guarantees for Bayesian best-arm identification," in *Proc. Int. Conf. Artif. Intell. Statist.*, Sicily, Italy, Aug. 2020, pp. 1823–1832.

[26] A. Wald, "Sequential tests of statistical hypotheses," *Ann. Math. Stat.*, vol. 16, no. 2, pp. 117–186, Jun. 1945.

[27] A. Mukherjee and A. Tajer, "SPRT-based best arm identification in stochastic bandits," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Espoo, Finland, Jul. 2022, pp. 1897–1902.

[28] B. K. Ghosh, "Sequential analysis: Tests and confidence intervals," *SIAM Rev.*, vol. 29, no. 2, pp. 315–318, 1987.

[29] M. Jourdan, R. Degenne, D. Baudry, R. de Heide, and E. Kaufmann, "Top two algorithms revisited," in *Proc. Adv. Neural Inf. Process. Syst.*, New Orleans, LA, USA, Dec. 2022, pp. 26791–26803.

[30] E. Kaufmann and W. M. Koolen, "Mixture martingales revisited with applications to sequential tests and confidence intervals," *J. Mach. Learn. Res.*, vol. 22, no. 246, pp. 1–44, 2021.

[31] A. Mukherjee and A. Tajer, "SPRT-based efficient best arm identification in stochastic bandits," 2022, *arXiv:2207.11158*.

[32] A. Deza, E. Nematollahi, and T. Terlaky, "How good are interior point methods? Klee–Minty cubes tighten iteration-complexity bounds," *Math. Program.*, vol. 113, no. 1, pp. 1–14, 2008.

[33] M. Jourdan and R. Degenne, "Non-asymptotic analysis of a UCB-based top two algorithm," 2022, *arXiv 2210.05431*.