SPRT-based Best Arm Identification in Stochastic Bandits

Arpan Mukherjee and Ali Tajer
Rensselaer Polytechnic Institute

2022 IEEE International Symposium on Information Theory (ISIT) | 978-1-6654-2159-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/ISIT50566.2022.9834534

***Abstract*—This paper investigates the problem of best arm identification (BAI) in stochastic multi-armed bandits in the fixed confidence setting. A novel formulation based on sequential hypothesis testing is provided, and an algorithm for BAI is proposed that, in spirit, follows the structure of the canonical sequential probability ratio test (SPRT). The algorithm has three features: (1) its sample complexity is asymptotically optimal, (2) it is guaranteed to be δ-PAC, and (3) it addresses the computational challenge of the state-of-the-art approaches. Specifically, the existing approaches rely on Thompson sampling for dynamically identifying the best arm and a challenger. This paper shows that identifying the challenger can be computationally expensive and demonstrates that the SPRT-based approach addresses that computational weakness.**

I. Introduction

A. Relevant Literature

In this paper, we consider the problem of best arm identification (BAI) in stochastic multi-armed bandits (MABs). The BAI problem is studied, broadly, under two settings: the *fixed budget* setting and the *fixed confidence* setting. The objective in the fixed budget setting is to identify the arm with 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 ensuring a guarantee on the error probability, and the objective is to minimize the sample complexity.

The BAI problem was first investigated in [1], which focused on the fixed budget setting. More studies on this setting can be found in [2] and [3]. Representative studies in the fixed confidence setting can be found in [4]–[7]. Algorithms in this setting can be classified into two categories: non-Bayesian algorithms and Bayesian algorithms. Some of the non-Bayesian approaches to BAI in stochastic MABs include confidence interval-based approaches (see [4] and [7]) and successive elimination-based approaches (see [8]).

In the confidence interval-based approach, the learner computes the sample mean of each arm as an empirical estimate for it, and a confidence interval around these estimates. the true means lie in these intervals with a high probability. The rationale behind this strategy is to gather more evidence until there is no overlap among the confidence intervals, and then the learner decides the best arm based on the empirical estimates. The successive elimination-based strategy, on the other hand, involves eliminating the potentially suboptimal arms in each round and sampling from all other arms until only one arm remains to be eliminated. The state-of-the-art in the non-Bayesian setting is the track and stop strategy of [6], which tracks the optimal allocation of arms, and selects the next action based on the estimated optimal allocation computed at the current time instant. This algorithm is asymptotically optimal up to a constant factor and faces the computational challenge of computing the optimal allocation at each time instance.

While non-Bayesian approaches have been investigated extensively, the Bayesian setting is far less investigated. The first Bayesian algorithms were introduced in [9], based on a *top-two* design philosophy. Among them, the top-two Thompson sampling (TTTS) algorithm has received more attention due to its simplicity and optimality properties. Owing to the simplicity of the top-two design principle, an improvement of the expected improvement algorithm was proposed in [10], and it was shown to be asymptotically optimal up to a constant factor. The sample complexity of TTTS in the Gaussian setting was later analyzed in [11], where it was shown to be asymptotically optimal.

B. Contribution

Despite its simplicity, the TTTS algorithm faces a computational challenge. Specifically, for dynamically identifying the top two arms, it generates random samples from the posterior distributions of the rewards. The coordinate with the largest value in the first sample is deemed as the coordinate of the best arm. 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. The computational challenge of TTTS stems from the following behavior: after enough explorations, the posterior distribution of the average reward converges to the true model, and the largest coordinate of any random sample will be pointing to the best arm. This significantly increases the delay for encountering a challenger.

In this paper, we propose a sequential hypothesis testing framework for formulating and solving the BAI problem in the fixed confidence setting. In this framework, we design a BAI algorithm that mitigates the computational difficulty of TTTS while maintaining the optimality guarantees. The combination of arm selection and stopping rules are, in spirit, similar to the sequential probability ratio test (SPRT) [12]. The arm selection rules involve dynamically updating generalized likelihood ratios that compare the relative likelihood of different arms for being among the best arms. We refer to this algorithm by the top-two SPRT (TT-SPRT). While achieving the same optimality guarantees as those of TTTS, TT-SPRT

This research was supported in part by the U. S. National Science Foundation under the CAREER Award ECCS-1554482 and the Rensselaer-IBM Artificial Intelligence Research Collaboration (AIRC) center.

does not face the delay that TTTS faces for identifying the challengers. To establish this, we analyze the number of samples that TTTS should generate for encountering the challenger. TT-SPRT, on the other hand, does not require gathering additional samples for identifying the challenger. To establish the optimality of TT-SPRT, we provide an upper bound on its sample complexity and show that it matches the universal lower bound asymptotically.

II. BAI in Stochastic Bandits

Consider a K-armed Gaussian stochastic MAB, such that arm $i \in [K] \triangleq \{1, \cdots, K\}$ generates rewards according to the Gaussian distribution $\mathcal{N}(\mu_i, \sigma^2)$, which we denote by $f_i(\cdot \mid \mu_i)$. The vector of mean values $\boldsymbol{\mu} \triangleq [\mu_1, \ldots, \mu_K]$ is unknown, and the mean values are assumed to be distinct. The arm with the largest mean value is denoted by

$$I^\star \triangleq \underset{i \in [K]}{\arg\max}\ \mu_i \ . \tag{1}$$

Furthermore, the gap between the means of the best arm $I^\star$ and any other arm $i \in [K] \setminus \{I^\star\}$ is denoted by

$$\Delta_i \triangleq \mu_{I^\star} - \mu_i \ , \tag{2}$$

and the smallest gap among all possible pairs is represented by

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

The value of the variance σ^2 is assumed to be known. At each time instant $n \in \mathbb{N}$, the learner selects and samples an arm, denoted by I_n, and nature generates a random reward denoted by Y_{n,I_n}. The sequence of arm selections and the corresponding rewards obtained up to time n are denoted by

$$\mathcal{I}^n \triangleq \{I_1, \cdots, I_n\}, \ \text{ and } \ \mathcal{Y}^n \triangleq \{Y_{1,I_1}, \cdots, Y_{n,I_n}\} \ . \tag{4}$$

Furthermore, the sequence of rewards accumulated from arm $i \in [K]$ up to time n is denoted by

$$\mathcal{Y}_i^n \triangleq \{Y_{\ell,I_\ell} : I_\ell = i, \ell \in [n]\} \ . \tag{5}$$

The objective of the learner is to identify the arm with the largest mean using as few samples as possible. Any sequential BAI algorithm has a stochastic stopping time, at which the algorithm identifies an arm as the best arm. Let τ denote the stochastic stopping time and $\hat{I}_\tau$ denote 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 $I^\star$ with a pre-specified confidence level. We use two notions of optimality. First, we require the BAI algorithm to have a terminal decision $\hat{I}_\tau$ such that the probability of error falls below a pre-specified guarantee δ, which we call the δ-PAC guarantee. Secondly, we require that the average sample complexity required by the algorithm in order to reach a confident decision should match the universal lower bound asymptotically, which we call β-optimality. Both these notions are specified in Section IV.

III. Top-Two SPRT for BAI

We provide an SPRT-based algorithm for BAI, referred to as top-two (TT)-SPRT. We will describe the specifics of arm selection strategy and stopping rule in this section, and provide the attendant performance guarantees in Section IV.

A. Posing BAI as Hypothesis Testing

We propose to view and analyze the BAI problem as a collection of binary composite hypothesis testing problems. Specifically, at each time instance $n \in \mathbb{N}$, for any distinct pair of arms $(i, j) \in [K] \times [K]$, we specify the binary composite hypothesis test

$$\mathcal{H}_{i,j} : \mu_i > \mu_j \ , \tag{6}$$

which aims to determine whether arm i has a larger mean compared to arm j. Thus, we have $\binom{K}{2}$ different hypotheses. For each pair of arms (i, j), looking at the sequence of rewards drawn from these two arms until time instant n, i.e., $\mathcal{Y}_i^n$ and $\mathcal{Y}_j^n$, this test compares the order of the two mean values μ_i and μ_j. Such a test, even though being closely related to the sequential hypothesis testing problem, is distinct from the objective of the BAI problem. First, the objective of solving the hypothesis test in (6) is reaching a terminal decision about the true model. In the BAI problem, on the other hand, we are not interested in ordering all the arms and we need to identify only the best arm. Secondly, we need to dynamically select the arms over time, a decision that does not exist in binary hypothesis testing.

We consider a generalized likelihood ratio test for forming the arm selection and BAI decisions. Specifically, at time n and corresponding to each pair (i, j) we define the generalized log-likelihood ratio (GLLR)

$$\Lambda_n(i,j) \triangleq \log \frac{\sup_{\boldsymbol{\mu}} \mathbb{P}(\mathcal{Y}^n \mid \mathcal{H}_{i,j})}{\sup_{\boldsymbol{\mu}} \mathbb{P}(\mathcal{Y}^n \mid \mathcal{H}_{j,i})} \ . \tag{7}$$

It can be readily verified that $\Lambda_n(i, j)$ can be simplified to

$$\Lambda_n(i,j) \triangleq \log \frac{\max\limits_{\mu_i > \mu_j} f_i(\mathcal{Y}_i^n \mid \mu_i) f_j(\mathcal{Y}_j^n \mid \mu_j)}{\max\limits_{\mu_j > \mu_i} f_i(\mathcal{Y}_i^n \mid \mu_i) f_j(\mathcal{Y}_j^n \mid \mu_j)} \ . \tag{8}$$

As shown in [6], when f_i belongs to the exponential family of distributions, i.e., $f_i(y \mid \mu_i) = \exp(\mu_i y - b(\mu_i))$, where $b : \mathbb{R} \mapsto \mathbb{R}$ is a convex, twice-differentiable function of μ_i, $\Lambda_n(i, j)$ takes a closed-form specified in the following lemma. To specify the closed form, we define $T_{n,i}$ as the number of times that arm $i \in [K]$ is pulled up to time n, and define the sample mean

$$\mu_{n,i} \triangleq \frac{1}{T_{n,i}} \sum_{y \in \mathcal{Y}_i^n} y \ , \tag{9}$$

as an empirical estimate of μ_i. Accordingly, we define the weighted average of the empirical means of arms i and j as

$$\mu_{n,i,j} \triangleq \frac{T_{n,i}\mu_{n,i} + T_{n,j}\mu_{n,j}}{T_{n,i} + T_{n,j}} \ . \tag{10}$$

Lemma 1 ([6]). *Under the exponential family of distributions, the GLLR defined in (8) has a closed form given by*

$$\Lambda_n(i,j) = \Big[T_{n,i}\mathsf{D}_{\mathsf{KL}}(\mu_{n,i}\|\mu_{n,i,j}) + T_{n,j}\mathsf{D}_{\mathsf{KL}}(\mu_{n,j}\|\mu_{n,i,j})\Big] \times \mathbb{1}_{\{\mu_{n,i}>\mu_{n,j}\}} \ , \tag{11}$$

where $\mathsf{D}_{\mathsf{KL}}(\mu_i\|\mu_j)$ *denotes the Kullback-Leibler (KL) divergence between two distributions with parameters* μ_i *and* μ_j.

Based on this lemma, for Gaussian distributions, the closed form in (11) simplifies to:

$$\Lambda_n(i,j) = \frac{(\mu_{n,i}-\mu_{n,j})^2}{2\sigma^2\left(\frac{1}{T_{n,i}}+\frac{1}{T_{n,j}}\right)} \cdot \mathbb{1}_{\{\mu_{n,i}>\mu_{n,j}\}} \ . \tag{12}$$

B. Sampling Strategy

At each instant, the TT-SPRT identifies the arm that has a positive log-likelihood ratio with respect to every other arm. Note that we have exactly one such arm, which we refer to as the *top* arm. We denote the top arm by I_1^n, i.e.,

$$I_1^n = i \quad \text{such that} \quad \Lambda_n(i,j) > 0 \quad \forall\, j \in [K]\setminus\{i\} \ . \tag{13}$$

In the Gaussian setting the top arm has the largest sample mean $\mu_{n,i}$ at time n. Hence,

$$I_1^n = \underset{i\in[K]}{\arg\max}\ \mu_{n,i} \ . \tag{14}$$

Besides the top arm, we also define the *challenger* arm as the one that 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 I_1^n. We denote the *challenger* arm at time n by I_2^n and it is given by

$$I_2^n \triangleq \underset{j\in[K]\setminus\{I_1^n\}}{\arg\min}\ \Lambda_n(I_1^n, j) \ . \tag{15}$$

Based on the choices of the top and challenger arms, at time n, our 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, our action at time $n+1$ is specified by

$$I_{n+1} \triangleq \begin{cases} I_1^n \ , & \text{if } \ D_n = 1 \\ I_2^n \ , & \text{if } \ D_n = 0 \end{cases} \ . \tag{16}$$

C. Stopping Rule

We adopt a thresholding-based stopping criterion, where we design the threshold in a way that the algorithm meets the δ-PAC guarantee. At each instant of time, we evaluate the GLLR between the top-two arms I_1^n and I_2^n. The procedure stops as soon as the GLLR $\Lambda_n(I_1^n, I_2^n)$ exceeds a threshold $c_{n,\delta}$, which is specified in Section IV. This threshold will be selected to meet the guarantee on the desired confidence level $1-\delta$. Specifically, the stopping rule is

$$\tau \triangleq \inf\Big\{n \in \mathbb{N} : \Lambda_n(I_1^n, I_2^n) > c_{n,\delta}\Big\} \ . \tag{17}$$

Hence, the TT-SPRT algorithm stops sampling and identifies an arm as soon as it has gathered sufficient evidence that it can distinguish the top arm I_1^n from the challenger I_2^n. Note that the stopping criterion is different from that of the SPRT's, which uses an upper and a lower threshold that are designed based on the required Type-I and Type-II error guarantees. In the BAI problem, we require a guarantee on the overall probability of error. Hence, instead of specifying different confidence levels for every possible incorrect decision, we have a guarantee on the aggregate probability of error, i.e., $\mathbb{P}(\tau < +\infty, \hat{I}_\tau \neq I^\star)$. Thus, it suffices to have one threshold $c_{n,\delta}$, which can be controlled by δ. The choice of $c_{n,\delta}$ as a function of δ and n and the attendant performance guarantees will be discussed in the next section. Specifically, we will analyze the sample complexity and the performance in the probably approximately correct (PAC) learning framework.

IV. Main Results

In this section, we establish the optimality of the TT-SPRT algorithm. To furnish context, we first briefly review the state-of-the-art algorithm and establish a result that shows its computational weakness. This computational weakness is the key motivation for TT-SPRT.

A. Challenger Identification in Top-Two Thompson Sampling

The TTTS algorithm [9], 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 $\mathcal{Y}^n$, the learner computes a posterior distribution $\Pi_n \in \mathbb{R}^K \to \mathbb{R}$. Specifically, for the average reward realization $\tilde{\boldsymbol{\mu}}$:

$$\Pi_n(\tilde{\boldsymbol{\mu}} \mid \mathcal{Y}^n) \triangleq \mathbb{P}(\boldsymbol{\mu} = \tilde{\boldsymbol{\mu}} \mid \mathcal{Y}^n) \ . \tag{18}$$

The marginal posterior reward distribution of arm $i \in [K]$ is Gaussian with mean $\tilde{\mu}_{n,i}$ and variance $\eta_{n,i}^2$ given by

$$\tilde{\mu}_{n,i} \triangleq \frac{1}{T_{n,i}+\sigma^2/\kappa^2}\sum_{y\in\mathcal{Y}_i^n} y \ , \tag{19}$$

$$\eta_{n,i}^2 \triangleq \frac{\sigma^2}{T_{n,i}+\sigma^2/\kappa^2} \ . \tag{20}$$

While the TTTS is devised for the setting with a Gaussian prior distribution for the rewards, the 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 (19) reduces to that of the sample mean, i.e., $\tilde{\mu}_{n,i} = \mu_{n,i}$, and the setting for both TTTS as well as TT-SPRT becomes 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 $\sigma_{n,i}^2 \triangleq \sigma^2/T_{n,i}$.

The arm selection strategy of the TTTS algorithm works as follows. At each time n, a random K-dimensional sample

$\boldsymbol{\theta}^n \triangleq (\theta_1^n, \cdots \theta_K^n)$ 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 $J_1^n \triangleq \arg\max_{i\in[K]} \theta_i^n$. In order to find a challenger (the closest competitor to J_1^n), 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 J_1^n. This is considered the challenger arm and its index is denoted by J_2^n. 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 define a sample s generated from Π_n by $\boldsymbol{\theta}_s^n \triangleq (\theta_{s,1}^n, \cdots, \theta_{s,K}^n)$. By design, clearly, $J_2^n \triangleq \arg\max_{i\in[K]} \theta_{s,i}^n$, and $J_2^n \neq J_1^n$. Once J_1^n and J_2^n are identified, the TTTS selects one of them based on a Bernoulli random variable parameterized by $\beta \in (0,1)$. 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

$$T_{\mathsf{TTTS}}^n \triangleq \inf\{s \in \mathbb{N} : \exists i \in [K], \theta_{s,i}^n > \theta_{s,J_1^n}^n\} \ , \tag{21}$$

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

Theorem 1. *In the TTTS algorithm [9], there exists $N_0 \in \mathbb{N}$ such that at any time $n > N_0$, the average number of posterior samples required in order to find a challenger is lower-bounded as*

$$\mathbb{E}[T_{\mathsf{TTTS}}^n] \geq \min_{i\in[K]\setminus\{I^\star\}} 2\exp\left(\sqrt{\frac{n}{K}}\, C_{i,\mathsf{L}}\right) \ , \tag{22}$$

where we have defined

$$C_{i,\mathsf{L}} \triangleq \frac{(\Delta_i - \Delta_{\min}/2)^2}{4\sigma^2} \ . \tag{23}$$

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.

B. δ-PAC Guarantee

Next, we present the results related to the optimality of TT-SPRT. We first state the result characterizing its optimality with respect to the decision confidence, and then analyze the average sample complexity of TT-SPRT. To characterize the decision confidence, we use the notion of δ-PAC defined next.

Definition 1 (δ-PAC)**.** *We say that a BAI algorithm is δ-PAC with a confidence level $\delta \in (0,1)$, if it guarantees that*

$$\mathbb{P}\{\tau < +\infty,\ \hat{I}_\tau = I^\star\} > 1 - \delta \ , \tag{24}$$

where $\mathbb{P}$ denotes the measure induced by the interaction of the BAI algorithm with the bandit instances.

For proving the δ-PAC guarantee we leverage an existing result in [11] that proves that any arm selection strategy coupled with the stopping rule specified in (17) and a properly chosen threshold $c_{n,\delta}$ ensures δ-PAC guarantee. With appropriate adjustments, this result ensures δ-PAC guarantee for the TT-SPRT algorithm as well. For completeness, this result is presented in the following theorem.

Theorem 2. *The stopping rule in (17) with the choice of the threshold*

$$c_{n,\delta} \triangleq 4\log(4 + \log n) + 2f\left(\frac{\log((K-1)/\delta)}{2}\right) \ , \tag{25}$$

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

Proof. Follows similarly to [11, Theorem 2]. ∎

Next, we analyze the sample complexity of TT-SPRT. First, we define a few quantities that are instrumental to stating the result.

C. Sample Complexity

To analyze the sample complexity, that is the expected value of the stochastic stopping time, we start by defining the notion of *problem complexity*. Problem complexity characterizes the level of difficulty that an algorithm faces to identify the best arm with sufficient fidelity. It is an instance-dependent quantity and it is defined as

$$\Gamma_\beta^\star \triangleq \max_{\boldsymbol{\omega}^\beta : \omega_{I^\star}^\beta = \beta} \min_{i \neq I^\star} \frac{(\mu_{I^\star} - \mu_i)^2}{2\sigma^2(1/\omega_i + 1/\beta)} \ , \tag{26}$$

where $\boldsymbol{\omega}^\beta \triangleq [\omega_1^\beta, \cdots, \omega_K^\beta]$ denotes a K-dimensional probability simplex satisfying $\omega_{I^\star}^\beta = \beta$. In the Gaussian bandit setting, the optimal sampling proportions that maximize (26) are obtained by solving [10] and [11]:

$$\frac{(\mu_{I^\star} - \mu_i)^2}{1/\omega_i^\beta + 1/\beta} = \frac{(\mu_{I^\star} - \mu_j)^2}{1/\omega_j^\beta + 1/\beta} \ , \quad \forall\, i,j \in [K] \setminus \{I^\star\} \ . \tag{27}$$

To show that TT-SPRT achieves the optimal sample complexity, we provide an upper bound on its sample complexity, which matches the following known information-theoretic lower bound on the sample complexity of any BAI algorithm.

Lemma 2 (Lower bound [11])**.** *Under any δ-PAC strategy that almost surely satisfies $\frac{T_{n,I^\star}}{n} \to \beta$, we have*

$$\liminf_{\delta\to 0} \frac{\mathbb{E}[\tau]}{\log(1/\delta)} \geq \frac{1}{\Gamma_\beta^\star} \ . \tag{28}$$

The universal lower bound in Lemma 2 provides the minimum number of samples that any δ-PAC BAI algorithm requires asymptotically, provided that β fraction of measurements is allocated to the best arm. We define the notion of β-optimality, which was first introduced in establishing optimality guarantees for the top-two algorithms for BAI in [10], and was adopted later for establishing the optimality of the TTTS algorithm [11].

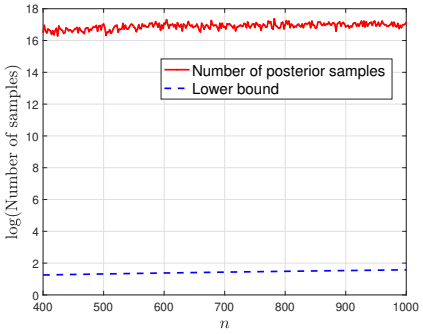

Fig. 1. Average number of posterior samples versus n

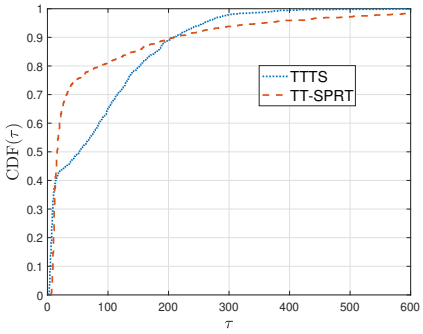

Fig. 2. Empirical CDF

Definition 2 (β-optimality)**.** *A BAI strategy is called asymptotically β-optimal, if it satisfies:*

$$\frac{T_{n,I^\star}}{n} \xrightarrow{n\to\infty} \beta \text{ a.s. }, \text{ and } \limsup_{\delta\to 0} \frac{\mathbb{E}[\tau]}{\log(1/\delta)} \leq \frac{1}{\Gamma^\star_\beta} . \quad (29)$$

Finally, we show that the TT-SPRT algorithm is β-optimal, i.e., its sample complexity matches the universal lower bound provided in Lemma 2 asymptotically, while satisfying the condition on the measurements allocated to the best arm $I^\star$.

Theorem 3. *The TT-SPRT algorithm, which consists of the sampling rule in (16) and stopping rule in (17), is asymptotically β-optimal, i.e., it satisfies:*

$$\frac{T_{n,I^\star}}{n} \xrightarrow{n\to\infty} \beta \text{ a.s. }, \text{ and } \limsup_{\delta\to 0} \frac{\mathbb{E}[\tau]}{\log(1/\delta)} \leq \frac{1}{\Gamma^\star_\beta} . \quad (30)$$

V. Numerical Experiments

In this section, we conduct numerical experiments to compare the performance of TT-SPRT against that of TTTS as the state-of-the-art algorithm. For all the experiments, we use a Gaussian bandit model in which we have set $\boldsymbol{\mu} \triangleq [5, 4.5, 1, 1, 1]$ and $\sigma^2 = 1$. This bandit model has also been used in [10] for comparing the empirical performance of BAI algorithms. We set $\beta = 0.98$ for choosing between the top two arms in TT-SPRT. Furthermore, all experiments are averaged over 1000 independent Monte Carlo trials.

In order to show the computational difficulty of obtaining a challenger in TTTS, in Figure 1 we plot the average number of posterior samples required by TTTS to identify a challenger and compare ir with the lower bound in Theorem 1. It is observed that the actual number of samples required by TTTS is considerably more than the lower bound.

Next, we compare the sample complexities of TT-SPRT and TTTS. We note that a more recent algorithm called T3C was also proposed in [11]. However, its performance was shown to be nearly comparable to, if not worse than that of TTTS. Hence, we restrict our comparison only to TTTS. For this comparison, we set $\delta = 10^{-3}$. The sample complexity for TT-SPRT in this setting is 73.741, which is about 10% lower than that of TTTS, which is 81.655. Figure 2 depicts the cumulative distribution functions (CDFs) of the stopping time τ for TT-SPRT and TTTS. The improvement in the average sample complexity of TT-SPRT is a result of the fact that a large fraction of the realizations of the stopping time take smaller values, compared to the realizations of TTTS. Specifically. we observe that 81% of the realizations for TT-SPRT are below 100, whereas only 65.9% of those corresponding to TTTS are below 100. This indicates that TT-SPRT requires fewer number of samples in order to distinguish the best arm from the second arm.

VI. Conclusions

In this paper, we have investigated the problem of best arm identification in stochastic multi-armed bandits. We have proposed a sequential hypothesis testing framework to formalize and analyze this problem. We have characterized the arm selection and terminal decision rules based on generalized likelihood ratio tests. The decisions rules (dynamic arm selection and stopping time) have three main properties: (1) they achieve optimality in the probably approximately correct learning framework, (2) they asymptotically achieve the optimal sample complexity, and (3) they address the computational shortcoming of the state-the-art-of approaches. We have analytically characterized the optimality properties, and compared with the state-of-the-art both analytically and numerically.

References

[1] S. Bubeck, R. Munos, and G. Stoltz, "Pure exploration in multi-armed bandits problems," in *Proc. International Conference on Algorithmic Learning Theory*, Porto, Portugal, October 2009.

[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. International Conference on Artificial Intelligence and Statistics*, Reykjavik, Iceland, April 2014.

[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. Advances in Neural Information Processing Systems*, Virtual, December 2020.

[4] V. Gabillon, M. Ghavamzadeh, and A. Lazaric, "Best arm identification: A unified approach to fixed budget and fixed confidence," in *Proc. Advances in Neural Information Processing Systems*, Lake Tahoe, NV, December 2012.

[5] S. Kalyanakrishnan, A. Tewari, P. Auer, and P. Stone, "PAC subset selection in stochastic multi-armed bandits," in *Proc. International Conference on Machine Learning*, Madison, WI, June 2012.

[6] A. Garivier and E. Kaufmann, "Optimal best arm identification with fixed confidence," in *Proc. Conference on Learning Theory*, New York, NY, June 2016.

[7] L. Xu, J. Honda, and M. Sugiyama, "A fully adaptive algorithm for pure exploration in linear bandits," in *Proc. International Conference on Artificial Intelligence and Statistics*, Lanzarote, Canary Islands, April 2018.

[8] E. Even-Dar, S. Mannor, and Y. Mansour, "Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems," *Journal of Machine Learning Research*, vol. 7, no. 39, pp. 1079–1105, 2006.

[9] Daniel Russo, "Simple bayesian algorithms for best-arm identification," *Operations Research*, vol. 68, no. 6, pp. 1625–1647, April 2020.

[10] C. Qin, D. Klabjan, and D. Russo, "Improving the expected improvement algorithm," in *Proc. Advances in Neural Information Processing Systems*, Long Beach, CA, December 2017.

[11] X. Shang, R. de Heide, P. Menard, E. Kaufmann, and M. Valko, "Fixed-confidence guarantees for Bayesian best-arm identification," in *Proc. International Conference on Artificial Intelligence and Statistics*, Sicily, Italy, August 2020, pp. 1823–1832.

[12] A. Wald, "Sequential tests of statistical hypotheses," *The Annals of Mathematical Statistics*, vol. 16, no. 2, pp. 117–186, June 1945.