Fixed-Confidence Guarantees for Bayesian Best-Arm Identification

Xuedong Shang[1,2] **Rianne de Heide**[3,4] **Emilie Kaufmann**[1,2,5] **Pierre Ménard**[1] **Michal Valko**[6,1]
[1]Inria Lille Nord Europe [2]Université de Lille [3]Leiden University [4]CWI [5]CNRS [6]DeepMind Paris

Abstract

We investigate and provide new insights on the sampling rule called Top-Two Thompson Sampling (`TTTS`). In particular, we justify its use for *fixed-confidence best-arm identification*. We further propose a variant of `TTTS` called Top-Two Transportation Cost (`T3C`), which disposes of the computational burden of `TTTS`. As our main contribution, we provide the first sample complexity analysis of `TTTS` and `T3C` when coupled with a very natural Bayesian stopping rule, for bandits with Gaussian rewards, solving one of the open questions raised by Russo (2016). We also provide new posterior convergence results for `TTTS` under two models that are commonly used in practice: bandits with Gaussian and Bernoulli rewards and conjugate priors.

1 Introduction

In multi-armed bandits, a learner repeatedly chooses an *arm* to play, and receives a reward from the associated unknown probability distribution. When the task is *best-arm* identification (BAI), the learner is not only asked to sample an arm at each stage, but is also asked to output a recommendation (i.e., a guess for the arm with the largest mean reward) after a certain period. Unlike in another well-studied bandit setting, the learner is not interested in maximizing the sum of rewards gathered during the exploration (or minimizing *regret*), but only cares about the quality of her recommendation. As such, BAI is a particular *pure exploration* setting (Bubeck et al., 2009).

Formally, we consider a finite-arm bandit model, which is a collection of K probability distributions, called arms $\mathcal{A} \triangleq \{1, \ldots, K\}$, parametrized by their means $\mu_1, \ldots, \mu_K$. We assume the (unknown) best arm is unique and we denote it by $I^\star \triangleq \arg\max_i \mu_i$. A best-arm identification strategy (I_n, J_n, τ) consists of three components. The first is a *sampling rule*, which selects an arm I_n at round n. At each round n, a vector of rewards $\mathbf{Y}_n = (Y_{n,1}, \cdots, Y_{n,K})$ is generated for all arms independently from past observations, but only Y_{n,I_n} is revealed to the learner. Let $\mathcal{F}_n$ be the σ-algebra generated by $(U_0, I_1, Y_{1,I_1}, U_1, \cdots, I_n, Y_{n,I_n}, U_n)$, then I_n is $\mathcal{F}_{n-1}$-measurable, i.e., it can only depend on the past $n-1$ observations, and some exogenous randomness, materialized into $U_{n-1} \sim \mathcal{U}([0,1])$. The second component is a $\mathcal{F}_n$-measurable *recommendation rule* J_n, which returns a guess for the best arm, and thirdly, the *stopping rule* τ, a stopping time with respect to $(\mathcal{F}_n)_{n\in\mathbb{N}}$, decides when the exploration is over.

BAI is studied within several theoretical frameworks. In this paper we consider the fixed-confidence setting, introduced by Even-dar et al. (2003). Given a risk parameter $\delta \in [0,1]$, the goal is to ensure that the probability to stop and recommend a wrong arm, $\mathbb{P}\left[J_\tau \neq I^\star \wedge \tau < \infty\right]$, is smaller than δ, while minimizing the expected total number of samples to make this accurate recommendation, $\mathbb{E}\left[\tau\right]$. The most studied alternative setting is the fixed-budget setting for which the stopping rule τ is fixed to some (known) maximal budget n, and the goal is to minimize the error probability $\mathbb{P}\left[J_n \neq I^\star\right]$ (Audibert and Bubeck, 2010). Note that these two frameworks are very different in general and do not share transferable regret bounds (see Carpentier and Locatelli 2016 for an additional discussion).

Most existing sampling rules for the fixed-confidence setting depend on the risk parameter δ. Some of them rely on confidence intervals such as `LUCB` (Kalyanakrishnan et al., 2012), `UGapE` (Gabillon et al., 2012), or `lil'UCB` (Jamieson et al., 2014); others are based on eliminations such as `SuccessiveElimination` (Even-dar et al., 2003) and `ExponentialGapElimination` (Karnin et al., 2013). The first known sampling rule for BAI that does not depend on δ is the *tracking* rule proposed by Garivier and Kaufmann (2016), which is proved to achieve

Proceedings of the 23rdInternational Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108. Copyright 2020 by the author(s).

the minimal sample complexity when combined with the Chernoff stopping rule when δ goes to zero. Such an *anytime* sampling rule (neither depending on a risk δ or a budget n) is very appealing for applications, as advocated by Jun and Nowak (2016) who introduce the anytime best-arm identification framework. In this paper, we investigate another anytime sampling rule for BAI: Top-Two Thompson Sampling (TTTS), and propose a second anytime sampling rule: Top-Two Transportation Cost (T3C).

Thompson Sampling (Thompson, 1933) is a Bayesian algorithm well known for regret minimization, for which it is now seen as a major competitor to UCB-typed approaches (Burnetas and Katehakis, 1996; Auer et al., 2002; Cappé et al., 2013). However, it is also well known that regret minimizing algorithms cannot yield optimal performance for BAI (Bubeck et al., 2011; Kaufmann and Garivier, 2017) and as we opt Thompson Sampling for BAI, then its adaptation is necessary. Such an adaptation, TTTS, was given by Russo (2016) along with two other top-two sampling rules TTPS and TTVS. By choosing between two different candidate arms in each round, these sampling rules enforce the exploration of sub-optimal arms, which would be under-sampled by vanilla Thompson sampling due to its objective of maximizing rewards.

While TTTS appears to be a good anytime sampling rule for fixed-confidence BAI when coupled with an appropriate stopping rule, so far there is no theoretical support for this employment. Indeed, the (Bayesian-flavored) asymptotic analysis of Russo (2016) shows that under TTTS, the posterior probability that $I^\star$ is the best arm converges almost surely to 1 at the best possible rate. However, this property does not by itself translate into sample complexity guarantees. Since the result of Russo (2016), Qin et al. (2017) proposed and analyzed TTEI, another Bayesian sampling rule, both in the fixed-confidence setting and in terms of posterior convergence rate. Nonetheless, similar guarantees for TTTS have been left as an open question by Russo (2016). In the present paper, we answer the question whether we can obtain fixed-confidence guarantees and optimal posterior convergence rates for TTTS. In addition, we propose T3C, a computationally more favorable variant of TTTS and extend the fixed-confidence guarantees to T3C as well.

Contributions (1) We propose a new Bayesian sampling rule, T3C, which is inspired by TTTS but easier to implement and computationally advantageous (2) We investigate two Bayesian stopping and recommendation rules and establish their δ-correctness for a bandit model with Gaussian rewards.[1] (3) We provide the first sample complexity analysis of TTTS and T3C for a Gaussian model and our proposed stopping rule. (4) Russo's posterior convergence results for TTTS were obtained under restrictive assumptions on the models and priors, which exclude the two mostly used models in practice: Gaussian bandits with Gaussian priors and bandits with Bernoulli rewards[2] with Beta priors. We prove that optimal posterior convergence rates can be obtained for those two as well.

Outline In Section 2, we restate TTTS and introduce T3C along with our proposed recommendation and stopping rules. Then, in Section 3, we describe in detail two important notions of optimality that are invoked in this paper. The main fixed-confidence analysis follows in Section 4, and further Bayesian optimality results are given in Section 5. Numerical illustrations are given in Section 6.

2 Bayesian BAI Strategies

In this section, we give an overview of the sampling rule TTTS and introduce T3C. We provide details for Bayesian updating for Gaussian and Bernoulli models respectively, and introduce associated Bayesian stopping and recommendation rules.

2.1 Sampling rules

Both TTTS and T3C employ a Bayesian machinery and make use of a prior distribution Π_1 over a set of parameters Θ, which is assumed to contain the unknown true parameter vector $\boldsymbol{\mu}$. Upon acquiring observations $(Y_{1,I_1}, \cdots, Y_{n-1,I_{n-1}})$, we update our beliefs according to Bayes' rule and obtain a posterior distribution Π_n which we assume to have density π_n w.r.t. the Lebesgue measure. Russo's analysis is requires strong regularity properties on the models and priors, which exclude two important useful cases we consider in this paper: (1) the observations of each arm i follow a Gaussian distribution $\mathcal{N}(\mu_i, \sigma^2)$ with common known variance σ^2, with imposed Gaussian prior $\mathcal{N}(\mu_{1,i}, \sigma^2_{1,i})$, (2) all arms receive Bernoulli rewards with unknown means, with a uniform ($\mathcal{B}eta(1,1)$) prior on each arm.

Gaussian model For Gaussian bandits with a $\mathcal{N}(0, \kappa^2)$ prior on each mean, the posterior distribution of μ_i at round n is Gaussian with mean and variance that are respectively given by

$$\frac{\sum_{\ell=1}^{n-1} \mathbb{1}\{I_\ell = i\} Y_{\ell, I_\ell}}{T_{n,i} + \sigma^2/\kappa^2} \quad \text{and} \quad \frac{\sigma^2}{T_{n,i} + \sigma^2/\kappa^2},$$

where $T_{n,i} \triangleq \sum_{\ell=1}^{n-1} \mathbb{1}\{I_\ell = i\}$ is the number of selections of arm i before round n. For the sake of simplic-

[1] hereafter 'Gaussian bandits' or 'Gaussian model'

[2] hereafter 'Bernoulli bandits'

ity, we consider improper Gaussian priors with $\mu_{1,i} = 0$ and $\sigma_{1,i} = +\infty$ for all $i \in \mathcal{A}$, for which

$$\mu_{n,i} = \frac{1}{T_{n,i}} \sum_{\ell=1}^{n-1} \mathbb{1}\{I_\ell = i\} Y_{\ell, I_\ell} \quad \text{and} \quad \sigma_{n,i}^2 = \frac{\sigma^2}{T_{n,i}}.$$

Observe that in this case the posterior mean $\mu_{n,i}$ coincides with the empirical mean.

Beta-Bernoulli model For Bernoulli bandits with a uniform ($\mathcal{B}eta(1,1)$) prior on each mean, the posterior distribution of μ_i at round n is a Beta distribution with shape parameters $\alpha_{n,i} = \sum_{\ell=1}^{n-1} \mathbb{1}\{I_\ell = i\} Y_{\ell, I_\ell} + 1$ and $\beta_{n,i} = T_{n,i} - \sum_{\ell=1}^{n-1} \mathbb{1}\{I_\ell = i\} Y_{\ell, I_\ell} + 1$.

Now we briefly recall TTTS and introduce T3C. The pseudo-code of TTTS and T3C are shown in Algorithm 1.

Description of TTTS At each time step n, TTTS has two potential actions: (1) with probability β, a parameter vector $\boldsymbol{\theta}$ is sampled from Π_n, and TTTS chooses to play $I_n^{(1)} \triangleq \arg\max_{i\in\mathcal{A}} \theta_i$, (2) and with probability $1-\beta$, the algorithm continues sampling new $\boldsymbol{\theta}'$ until we obtain a *challenger* $I_n^{(2)} \triangleq \arg\max_{i\in\mathcal{A}} \theta'_i$ that is different from $I_n^{(1)}$, and TTTS chooses to play $I_n^{(2)}$.

Description of T3C One drawback of TTTS is that, in practice, when the posteriors become concentrated, it takes many Thompson samples before the challenger $I_n^{(2)}$ is obtained. We thus propose a variant of TTTS, called T3C, which alleviates this computational burden. Instead of re-sampling from the posterior until a different candidate appears, we define the challenger as the arm that has the lowest *transportation cost* $W_n(I_n^{(1)}, i)$ with respect to the first candidate (with ties broken uniformly at random).

Let $\mu_{n,i}$ be the empirical mean of arm i and $\mu_{n,i,j} \triangleq (T_{n,i}\mu_{n,i} + T_{n,j}\mu_{n,j})/(T_{n,i} + T_{n,j})$, then we define

$$W_n(i,j) \triangleq \begin{cases} 0 & \text{if } \mu_{n,j} \geq \mu_{n,i}, \\ W_{n,i,j} + W_{n,j,i} & \text{otherwise}, \end{cases} \tag{1}$$

where $W_{n,i,j} \triangleq T_{n,i} d\left(\mu_{n,i}, \mu_{n,i,j}\right)$ for any i, j and $d(\mu; \mu')$ denotes the Kullback-Leibler between the distribution with mean μ and that of mean μ'. In the Gaussian case, $d(\mu;\mu') = (\mu - \mu')^2/(2\sigma^2)$ while in the Bernoulli case $d(\mu;\mu') = \mu \ln(\mu/\mu') + (1-\mu)\ln(1-\mu)/(1-\mu')$. In particular, for Gaussian bandits

$$W_n(i,j) = \frac{(\mu_{n,i} - \mu_{n,j})^2}{2\sigma^2(1/T_{n,i} + 1/T_{n,j})} \mathbb{1}\{\mu_{n,j} < \mu_{n,i}\}.$$

Note that under the Gaussian model with improper priors, one should pull each arm once at the beginning for the sake of obtaining proper posteriors.

Algorithm 1 Sampling rule (TTTS/T3C)

```
Input: $\beta$
for $n \leftarrow 1, 2, \cdots$ do
    sample $\boldsymbol{\theta} \sim \Pi_n$
    $I^{(1)} \leftarrow \arg\max_{i\in\mathcal{A}} \theta_i$
    sample $b \sim \mathcal{B}ern(\beta)$
    if $b = 1$ then
        evaluate arm $I^{(1)}$
    else
        repeat sample $\boldsymbol{\theta}' \sim \Pi_n$                    }
        $I^{(2)} \leftarrow \arg\max_{i\in\mathcal{A}} \theta'_i$            } TTTS
        until $I^{(2)} \neq I^{(1)}$                                         }
        $I^{(2)} \leftarrow \arg\min_{i\neq I^{(1)}} W_n(I^{(1)}, i)$, cf. (1)    T3C
        evaluate arm $I^{(2)}$
    end if
    update mean and variance
    $t = t + 1$
end for
```

2.2 Rationale for T3C

In order to explain how T3C can be seen as an approximation of the re-sampling performed by TTTS, we first need to define the *optimal action probabilities.*

Optimal action probability The optimal action probability $a_{n,i}$ is defined as the posterior probability that arm i is optimal. Formally, letting Θ_i be the subset of Θ such that arm i is the optimal arm,

$$\Theta_i \triangleq \left\{ \boldsymbol{\theta} \in \Theta \,\middle|\, \theta_i > \max_{j\neq i} \theta_j \right\},$$

then we define

$$a_{n,i} \triangleq \Pi_n(\Theta_i) = \int_{\Theta_i} \pi_n(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}. \tag{2}$$

With this notation, one can show that under TTTS,

$$\Pi_n\left(I_n^{(2)} = j | I_n^{(1)} = i\right) = \frac{a_{n,j}}{\sum_{k\neq i} a_{n,k}}. \tag{3}$$

Furthermore, when i coincides with the empirical best mean (and this will often be the case for $I_n^{(1)}$ when n is large due to posterior convergence) one can write

$$a_{n,j} \simeq \Pi_n\left(\theta_j \geq \theta_i\right) \simeq \exp\left(-W_n(i,j)\right),$$

where the last step is justified in Lemma 2 in the Gaussian case (and Lemma 26 in Appendix I.3 in the Bernoulli case). Hence, T3C replaces sampling from the distribution (3) by an approximation of its mode which is *easy to compute*. Note that directly computing the mode would require to compute $a_{n,j}$, which is much more costly than the computation of $W_n(i,j)$[3].

[3] TTPS (Russo, 2016) also requires the computation of $a_{n,i}$, thus we do not report simulations for it in Sec. 6.

2.3 Stopping and recommendation rules

In order to use TTTS or T3C as the sampling rule for fixed-confidence BAI, we need to additionally define stopping and recommendation rules. While Qin et al. (2017) suggest to couple TTEI with the "frequentist" Chernoff stopping rule (Garivier and Kaufmann, 2016), we propose in this section natural Bayesian stopping and recommendation rules. They both rely on the optimal action probabilities defined in (2).

Bayesian recommendation rule At time step n, a natural candidate for the best arm is the arm with largest optimal action probability, hence we define

$$J_n \triangleq \underset{i\in\mathcal{A}}{\arg\max}\, a_{n,i} \,.$$

Bayesian stopping rule In view of the recommendation rule, it is natural to stop when the posterior probability that the recommended action is optimal is large, and exceeds some threshold $c_{n,\delta}$ which gets close to 1. Hence our Bayesian stopping rule is

$$\tau_\delta \triangleq \inf\left\{ n\in\mathbb{N} : \max_{i\in\mathcal{A}} a_{n,i} \geq c_{n,\delta} \right\} . \qquad (4)$$

Links with frequentist counterparts Using the transportation cost $W_n(i,j)$ defined in (1), the Chernoff stopping rule of Garivier and Kaufmann (2016) can actually be rewritten as

$$\tau_\delta^{\text{Ch.}} \triangleq \inf\left\{ n\in\mathbb{N} : \max_{i\in\mathcal{A}} \min_{j\in\mathcal{A}\setminus\{i\}} W_n(i,j) > d_{n,\delta} \right\} . \quad (5)$$

This stopping rule is coupled with the recommendation rule $J_n = \arg\max_i \mu_{n,i}$.

As explained in that paper, $W_n(i,j)$ can be interpreted as a (log) Generalized Likelihood Ratio statistic for rejecting the hypothesis $\mathcal{H}_0 : (\mu_i < \mu_j)$. Through our Bayesian lens, we rather have in mind the approximation $\Pi_n(\theta_j > \theta_i) \simeq \exp\{-W_n(i,j)\}$, valid when $\mu_{n,i} > \mu_{n,j}$, which permits to analyze the two stopping rules using similar tools, as will be seen in the proof of Theorem 2.

As shown later in Sec. 4, τ_δ and $\tau_\delta^{\text{Ch.}}$ prove to be fairly similar for some corresponding choices of the thresholds $c_{n,\delta}$ and $d_{n,\delta}$. This similarity endorses the use of the Chernoff stopping rule in practice, which does not require the (heavy) computation of optimal action probabilities. Still, our sample complexity analysis applies to the two stopping rules, and we believe that a frequentist sample complexity analysis of a fully Bayesian-flavored BAI strategy is a nice theoretical contribution.

Useful notation We follow the notation of Russo (2016) and define the following measures of effort allocated to arm i up to time n,

$$\psi_{n,i} \triangleq \mathbb{P}\left[I_n = i|\mathcal{F}_{n-1}\right] \quad \text{and} \quad \Psi_{n,i} \triangleq \sum_{l=1}^{n} \psi_{l,i}.$$

In particular, for TTTS we have

$$\psi_{n,i} = \beta a_{n,i} + (1-\beta) a_{n,i} \sum_{j\neq i} \frac{a_{n,j}}{1-a_{n,j}},$$

while for T3C

$$\psi_{n,i} = \beta a_{n,i} + (1-\beta) \sum_{j\neq i} a_{n,j} \frac{\mathbb{1}\{W_n(j,i) = \min_{k\neq j} W_n(j,k)\}}{\#\left|\arg\min_{k\neq j} W_n(j,k)\right|}.$$

3 Two Related Optimality Notions

In the fixed-confidence setting, we aim for building δ-correct strategies, i.e. strategies that identify the best arm with high confidence on any problem instance.

Definition 1. *A strategy (I_n, J_n, τ) is δ-correct if for all bandit models $\boldsymbol{\mu}$ with a unique optimal arm, it holds that $\mathbb{P}_{\boldsymbol{\mu}}\left[J_\tau \neq I^\star \wedge \tau < \infty\right] \leq \delta$.*

Among δ-correct strategies, seek the one with the smallest sample complexity $\mathbb{E}\left[\tau_\delta\right]$. So far, TTTS has not been analyzed in terms of sample complexity; Russo (2016) focuses on posterior consistency and optimal convergence rates. Interestingly, both the smallest possible sample complexity and the fastest rate of posterior convergence can be expressed in terms of the following quantities.

Definition 2. *Let $\Sigma_K = \{\boldsymbol{\omega} : \sum_{k=1}^K \omega_k = 1, \omega_k \geq 0\}$ and define for all $i \neq I^\star$*

$$C_i(\omega, \omega') \triangleq \min_{x\in\mathcal{I}} \omega d(\mu_{I^\star}; x) + \omega' d(\mu_i; x),$$

where $d(\mu, \mu')$ is the KL-divergence defined above and $\mathcal{I} = \mathbb{R}$ in the Gaussian case and $\mathcal{I} = [0,1]$ in the Bernoulli case. We define

$$\begin{aligned} \Gamma^\star &\triangleq \max_{\boldsymbol{\omega}\in\Sigma_K} \min_{i\neq I^\star} C_i(\omega_{I^\star}, \omega_i), \\ \Gamma^\star_\beta &\triangleq \max_{\substack{\boldsymbol{\omega}\in\Sigma_K \\ \omega_{I^\star}=\beta}} \min_{i\neq I^\star} C_i(\omega_{I^\star}, \omega_i). \end{aligned} \qquad (6)$$

The quantity $C_i(\omega_{I^\star}, \omega_i)$ can be interpreted as a "transportation cost"[4] from the original bandit instance $\boldsymbol{\mu}$ to an alternative instance in which the mean of arm i is larger than that of $I^\star$, when the proportion of samples allocated to each arm is given by the vector $\boldsymbol{\omega} \in \Sigma_K$. As shown by Russo (2016), the $\boldsymbol{\omega}$ that maximizes (6) is unique, which allows us to define the *β-optimal allocation* $\boldsymbol{\omega}^\beta$ in the following proposition.

[4] for which $W_n(I^\star, i)$ is an empirical counterpart

Proposition 1. *There is a unique solution $\boldsymbol{\omega}^\beta$ to the optimization problem (6) satisfying $\omega_{I^\star}^\beta = \beta$, and for all $i, j \neq I^\star$, $C_i(\beta, \omega_i^\beta) = C_j(\beta, \omega_j^\beta)$.*

For models with more than two arms, there is no closed form expression for $\Gamma_\beta^\star$ or $\Gamma^\star$, even for Gaussian bandits with variance σ^2 for which we have

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

Bayesian β-optimality Russo (2016) proves that any sampling rule allocating a fraction β to the optimal arm ($\Psi_{n,I^\star}/n \to \beta$) satisfies $1 - a_{n,I^\star} \geq e^{-n(\Gamma_\beta^\star + o(1))}$ (a.s.).We define a *Bayesian β-optimal* sampling rule as a sampling rule matching this lower bound, i.e. satisfying $\Psi_{n,I^\star}/n \to \beta$ and $1 - a_{n,I^\star} \leq e^{-n(\Gamma_\beta^\star + o(1))}$.

Russo (2016) proves that TTTS with parameter β is Bayesian β-optimal. However, the result is valid only under strong regularity assumptions, excluding the two practically important cases of Gaussian and Bernoulli bandits. In this paper, we complete the picture by establishing Bayesian β-optimality for those models in Sec. 5. For the Gaussian bandit, Bayesian β-optimality was established for TTEI by Qin et al. (2017) with Gaussian priors, but this remained an open problem for TTTS.

A fundamental ingredient of these proofs is to establish the convergence of the allocation of measurement effort to the β-optimal allocation: $\Psi_{n,i}/n \to \omega_i^\beta$ for all i, which is equivalent to $T_{n,i}/n \to \omega_i^\beta$ (cf. Lemma 4).

β-optimality in the fixed-confidence setting In the fixed confidence setting, the performance of an algorithm is evaluated in terms of sample complexity. A lower bound given by Garivier and Kaufmann (2016) states that any δ-correct strategy satisfies $\mathbb{E}[\tau_\delta] \geq (\Gamma^\star)^{-1} \ln(1/(3\delta))$.

Observe that $\Gamma^\star = \max_{\beta\in[0,1]} \Gamma_\beta^\star$. Using the same lower bound techniques, one can also prove that under any δ-correct strategy satisfying $T_{n,I^\star}/n \to \beta$,

$$\liminf_{\delta\to 0} \frac{\mathbb{E}[\tau_\delta]}{\ln(1/\delta)} \geq \frac{1}{\Gamma_\beta^\star}.$$

This motivates the relaxed optimality notion that we introduce in this paper: A BAI strategy is called *asymptotically β-optimal* if it satisfies

$$\frac{T_{n,I^\star}}{n} \to \beta \quad \text{and} \quad \limsup_{\delta\to 0} \frac{\mathbb{E}[\tau_\delta]}{\ln(1/\delta)} \leq \frac{1}{\Gamma_\beta^\star}.$$

In the paper, we provide the first sample complexity analysis of a BAI algorithm based on TTTS (with the stopping and recommendation rules described in Sec. 2), establishing its asymptotic β-optimality.

As already observed by Qin et al. (2017), any sampling rule converging to the β-optimal allocation (i.e. satisfying $T_{n,i}/n \to w_i^\beta$ for all i) can be shown to satisfy $\limsup_{\delta\to 0} \tau_\delta/\ln(1/\delta) \leq (\Gamma_\beta^\star)^{-1}$ almost surely, when coupled with the Chernoff stopping rule. The fixed confidence optimality that we define above is stronger as it provides guarantees on $\mathbb{E}[\tau_\delta]$.

4 Fixed-Confidence Analysis

In this section, we consider Gaussian bandits and the Bayesian rules using an improper prior on the means. We state our main result below, showing that TTTS and T3C are asymptotically β-optimal in the fixed confidence setting, when coupled with appropriate stopping and recommendation rules.

Theorem 1. *With $\mathcal{C}^{gG}$ the function defined in Corollary 10 of Kaufmann and Koolen (2018), which satisfies $\mathcal{C}^{gG}(x) \simeq x + \ln(x)$, we introduce the threshold*

$$d_{n,\delta} = 4\ln(4 + \ln(n)) + 2\mathcal{C}^{gG}\left(\frac{\ln((K-1)/\delta)}{2}\right). \quad (7)$$

The TTTS and T3C sampling rules coupled with either

- *the Bayesian stopping rule (4) with threshold*
$$c_{n,\delta} = 1 - \frac{1}{\sqrt{2\pi}} e^{-\left(\sqrt{d_{n,\delta}} + \frac{1}{\sqrt{2}}\right)^2}$$
and recommendation rule $J_t = \arg\max_i a_{n,i}$, or
- *the Chernoff stopping rule (5) with threshold $d_{n,\delta}$ and recommendation rule $J_t = \arg\max_i \mu_{n,i}$,*

form a δ-correct BAI strategy. Moreover, if all the arms means are distinct, it satisfies

$$\limsup_{\delta\to 0} \frac{\mathbb{E}[\tau_\delta]}{\log(1/\delta)} \leq \frac{1}{\Gamma_\beta^\star}.$$

We now give the proof of Theorem 1, which is divided into three parts. The **first step** of the analysis is to prove the δ-correctness of the studied BAI strategies.

Theorem 2. *Regardless of the sampling rule, the stopping rule (4) with the threshold $c_{n,\delta}$ and the Chernoff stopping rule (5) with threshold $d_{n,\delta}$ defined in (7) satisfy $\mathbb{P}[\tau_\delta < \infty \wedge J_{\tau_\delta} \neq I^\star] \leq \delta$.*

To prove that TTTS and T3C allow to reach a β-optimal sample complexity, one needs to quantify how fast the measurement effort for each arm is concentrating to its corresponding optimal weight. For this purpose, we introduce the random variable

$$T_\beta^\varepsilon \triangleq \inf\left\{N \in \mathbb{N} : \max_{i\in\mathcal{A}} |T_{n,i}/n - \omega_i^\beta| \leq \varepsilon, \forall n \geq N\right\}.$$

The **second step** of our analysis is a sufficient condition for β-optimality, stated in Lemma 1. Its proof is given in Appendix F. The same result was proven for the Chernoff stopping rule by Qin et al. (2017).

Lemma 1. *Let $\delta, \beta \in (0,1)$. For any sampling rule which satisfies $\mathbb{E}\left[T^{\varepsilon}_{\beta}\right] < \infty$ for all $\varepsilon > 0$, we have*

$$\limsup_{\delta \to 0} \frac{\mathbb{E}\left[\tau_\delta\right]}{\log(1/\delta)} \leq \frac{1}{\Gamma^\star_\beta},$$

if the sampling rule is coupled with stopping rule (4),

Finally, it remains to show that `TTTS` and `T3C` meet the sufficient condition, and therefore the **last step**, which is the core component and the most technical part our analysis, consists of showing the following.

Theorem 3. *Under `TTTS` or `T3C`, $\mathbb{E}\left[T^{\varepsilon}_{\beta}\right] < +\infty$.*

In the rest of this section, we prove Theorem 2 and sketch the proof of Theorem 3. But we first highlight some important ingredients for these proofs.

4.1 Core ingredients

Our analysis hinges on properties of the Gaussian posteriors, in particular on the following tail bounds, which follow from Lemma 1 of Qin et al. (2017).

Lemma 2. *For any $i, j \in \mathcal{A}$, if $\mu_{n,i} \leq \mu_{n,j}$*

$$\Pi_n\left[\theta_i \geq \theta_j\right] \leq \frac{1}{2} \exp\left\{-\frac{(\mu_{n,j} - \mu_{n,i})^2}{2\sigma^2_{n,i,j}}\right\}, \tag{8}$$

$$\Pi_n\left[\theta_i \geq \theta_j\right] \geq \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(\mu_{n,j} - \mu_{n,i} + \sigma_{n,i,j})^2}{2\sigma^2_{n,i,j}}\right\}, \tag{9}$$

where $\sigma^2_{n,i,j} \triangleq \sigma^2/T_{n,i} + \sigma^2/T_{n,j}$.

This lemma is crucial to control $a_{n,i}$ and $\psi_{n,i}$, the optimal action and selection probabilities.

4.2 Proof of Theorem 2

We upper bound the desired probability as follows

$$\begin{aligned}
&\mathbb{P}\left[\tau_\delta < \infty \wedge J_{\tau_\delta} \neq I^\star\right] \leq \sum_{i \neq I^\star} \mathbb{P}\left[\exists n \in \mathbb{N} : a_{n,i} > c_{n,\delta}\right] \\
&\leq \sum_{i \neq I^\star} \mathbb{P}\left[\exists n \in \mathbb{N} : \Pi_n(\theta_i \geq \theta_{I_\star}) > c_{n,\delta}, \mu_{n,I^\star} \leq \mu_{n,i}\right] \\
&\leq \sum_{i \neq I^\star} \mathbb{P}\left[\exists n \in \mathbb{N} : 1 - c_{n,\delta} > \Pi_n(\theta_{I^\star} > \theta_i), \mu_{n,I^\star} \leq \mu_{n,i}\right].
\end{aligned}$$

The second step uses the fact that as $c_{n,\delta} \geq 1/2$, a necessary condition for $\Pi_n(\theta_i \geq \theta_{I_\star}) \geq c_{n,\delta}$ is that $\mu_{n,i} \geq \mu_{n,I_\star}$. Now using the lower bound (9), if $\mu_{n,I^\star} \leq \mu_{n,i}$, the inequality $1 - c_{n,\delta} > \Pi_n(\theta_{I^\star} > \theta_i)$ implies

$$\frac{(\mu_{n,i} - \mu_{n,I^\star})^2}{2\sigma^2_{n,i,I^\star}} \geq \left(\sqrt{\ln \frac{1}{\sqrt{2\pi}(1 - c_{n,\delta})}} - \frac{1}{\sqrt{2}}\right)^2 = d_{n,\delta},$$

where the equality follows from the expression of $c_{n,\delta}$ as function of $d_{n,\delta}$. Hence to conclude the proof it remains to check that

$$\mathbb{P}\left[\exists n \in \mathbb{N} : \mu_{n,i} \geq \mu_{n,I^\star}, \frac{(\mu_{n,i} - \mu_{n,I^\star})^2}{2\sigma^2_{n,i,I^\star}} \geq d_{n,\delta}\right] \leq \frac{\delta}{K-1}. \tag{10}$$

To prove this, we observe that for $\mu_{n,i} \geq \mu_{n,I^\star}$,

$$\begin{aligned}
\frac{(\mu_{n,i} - \mu_{n,I^\star})^2}{2\sigma^2_{n,i,I^\star}} &= \inf_{\theta_i < \theta_{I^\star}} T_{n,i} d(\mu_{n,i}; \theta_i) + T_{n,I^\star} d(\mu_{n,I^\star}; \theta_{I^\star}) \\
&\leq T_{n,i} d(\mu_{n,i}; \mu_i) + T_{n,I^\star} d(\mu_{n,I^\star}; \mu_{I^\star}).
\end{aligned}$$

Corollary 10 of Kaufmann and Koolen (2018) then allows us to upper bound the probability

$$\mathbb{P}\left[\exists n \in \mathbb{N} : T_{n,i} d(\mu_{n,i}; \mu_i) + T_{n,I^\star} d(\mu_{n,I^\star}, \mu_{I^\star}) \geq d_{n,\delta}\right]$$

by $\delta/(K-1)$ for the choice of threshold given in (7), which completes the proof that the stopping rule (4) is δ-correct. The fact that the Chernoff stopping rule with the above threshold $d_{n,\delta}$ given above is δ-correct straightforwardly follows from (10).

4.3 Sketch of the proof of Theorem 3

We present a unified proof sketch of Theorem 3 for `TTTS` and `T3C`. While the two analyses follow the same steps, some of the lemmas given below have different proofs for `TTTS` and `T3C`, which can be found in Appendix D and E respectively.

We first state two important concentration results, that hold under any sampling rule.

Lemma 3. *[Lemma 5 of Qin et al. 2017] There exists a random variable W_1, such that for all $i \in \mathcal{A}$,*

$$\forall n \in \mathbb{N}, \quad |\mu_{n,i} - \mu_i| \leq \sigma W_1 \sqrt{\frac{\log(e + T_{n,i})}{1 + T_{n,i}}} \ \textit{a.s.},$$

and $\mathbb{E}\left[e^{\lambda W_1}\right] < \infty$ for all $\lambda > 0$.

Lemma 4. *There exists a random variable W_2, such that for all $i \in \mathcal{A}$,*

$$\forall n \in \mathbb{N}, |T_{n,i} - \Psi_{n,i}| \leq W_2 \sqrt{(n+1)\log(e^2 + n)} \ \textit{a.s.},$$

and $\mathbb{E}\left[e^{\lambda W_2}\right] < \infty$ for any $\lambda > 0$.

Lemma 3 controls the concentration of the posterior means towards the true means and Lemma 4 establishes that $T_{n,i}$ and $\Psi_{n,i}$ are close. Both results rely on uniform deviation inequalities for martingales.

Our analysis uses the same principle as that of `TTEI`: We establish that T^{ε}_{β} is upper bounded by some random variable N which is a polynomial of the random variables W_1 and W_2 introduced in the above lemmas, denoted by $\text{Poly}(W_1, W_2) \triangleq \mathcal{O}(W_1^{c_1} W_2^{c_2})$, where c_1

and c_2 are two constants (that may depend on the arms' means and the constant hidden in the $\mathcal{O}$). As all exponential moments of W_1 and W_2 are finite, N has a finite expectation as well, concluding the proof.

The first step to exhibit such an upper bound N is to establish that every arm is pulled sufficiently often.

Lemma 5. *Under TTTS or T3C, there exists $N_1 = Poly(W_1, W_2)$ s.t.*

$$\forall n \geq N_1, \forall i,\ T_{n,i} \geq \sqrt{\frac{n}{K}},\ a.s..$$

Due to the randomized nature of TTTS and T3C, the proof of Lemma 5 is significantly more involved than for a deterministic rule like TTEI. Intuitively, the posterior of each arm would be well concentrated once the arm is sufficiently pulled. If the optimal arm is under-sampled, then it would be chosen as the first candidate with large probability. If a sub-optimal arm is under-sampled, then its posterior distribution would possess a relatively wide tail that overlaps with or cover the somehow narrow tails of other overly-sampled arms. The probability of that sub-optimal arm being chosen as the challenger would be large enough then.

Combining Lemma 5 with Lemma 3 straightforwardly leads to the following result.

Lemma 6. *Under TTTS or T3C, fix a constant $\varepsilon > 0$, there exists $N_2 = Poly(1/\varepsilon, W_1, W_2)$ s.t. $\forall n \geq N_2, \forall i \in \mathcal{A}, \quad |\mu_{n,i} - \mu_i| \leq \varepsilon$.*

We can then deduce a very nice property about the optimal action probability for sub-optimal arms from the previous two lemmas. Indeed, we can show that

$$\forall i \neq I^\star, \quad a_{n,i} \leq \exp\left\{-\frac{\Delta_{\min}^2}{16\sigma^2}\sqrt{\frac{n}{K}}\right\}$$

for n larger than some $\text{Poly}(W_1, W_2)$, where $\Delta_{\min}$ is the smallest mean difference among all the arms.

Plugging this in the expression of $\psi_{n,i}$, one can easily quantify how fast $\psi_{n,I^\star}$ converges to β, which eventually yields the following result.

Lemma 7. *Under TTTS or T3C, fix $\varepsilon > 0$, then there exists $N_3 = Poly(1/\varepsilon, W_1, W_2)$ s.t. $\forall n \geq N_3$,*

$$\left|\frac{T_{n,I^\star}}{n} - \beta\right| \leq \varepsilon.$$

The last, more involved, step is to establish that the fraction of measurement allocation to every sub-optimal arm i is indeed similarly close to its optimal proportion ω_i^β.

Lemma 8. *Under TTTS or T3C, fix a constant $\varepsilon > 0$, there exists $N_4 = Poly(1/\varepsilon, W_1, W_2)$ s.t. $\forall n \geq N_4$,*

$$\forall i \neq I^\star, \quad \left|\frac{T_{n,i}}{n} - \omega_i^\beta\right| \leq \varepsilon.$$

The major step in the proof of Lemma 8 for each sampling rule, is to establish that if some arm is over-sampled, then its probability to be selected is exponentially small. Formally, we show that for n larger than some $\text{Poly}(1/\varepsilon, W_1, W_2)$,

$$\frac{\Psi_{n,i}}{n} \geq \omega_i^\beta + \xi \quad \Rightarrow \quad \psi_{n,i} \leq \exp\left\{-f(n,\xi)\right\},$$

for some function $f(n,\xi)$ to be specified for each sampling rule, satisfying $f(n) \geq C_\xi\sqrt{n}$ (a.s.). This result leads to the concentration of $\Psi_{n,i}/n$, thus can be easily converted to the concentration of $T_{n,i}/n$ by Lemma 4.

Finally, Lemma 7 and Lemma 8 show that T_β^ε is upper bounded by $N \triangleq \max(N_3, N_4)$, which yields

$$\mathbb{E}[T_\beta^\varepsilon] \leq \max(\mathbb{E}\left[N_3\right], \mathbb{E}\left[N_4\right]) < \infty.$$

5 Optimal Posterior Convergence

Recall that $a_{n,I^\star}$ denotes the posterior mass assigned to the event that action $I^\star$ (i.e. the true optimal arm) is optimal at time n. As the number of observations tends to infinity, we want the posterior distribution to converge to the truth. In this section we show equivalently that the posterior mass on the complementary event, $1 - a_{n,I^\star}$, the event that arm $I^\star$ is not optimal, converges to zero at an exponential rate, and that it does so at optimal rate $\Gamma_\beta^\star$.

Russo (2016) proves a similar theorem under three confining boundedness assumptions (see Russo 2016, Assumption 1) on the parameter space, the prior density and the (first derivative of the) log-normalizer of the exponential family. Hence, the theorems in Russo (2016) do not apply to the two bandit models most used in practice, which we consider in this paper: the Gaussian and Bernoulli model.

In the first case, the parameter space is unbounded, in the latter model, the derivative of the log-normalizer (which is $e^\eta/(1+e^\eta)$) is unbounded. Here we provide a theorem, proving that under TTTS, the optimal, exponential posterior convergence rates are obtained for the Gaussian model with uninformative (improper) Gaussian priors (proof in Appendix H), and the Bernoulli model with $\mathcal{B}eta(1,1)$ priors (proof in Appendix I).

Theorem 4. *Under TTTS, for Gaussian bandits with improper Gaussian priors and for Bernoulli bandits with uniform priors, it holds almost surely that*

$$\lim_{n\to\infty} -\frac{1}{n}\log(1 - a_{n,I^\star}) = \Gamma_\beta^\star.$$

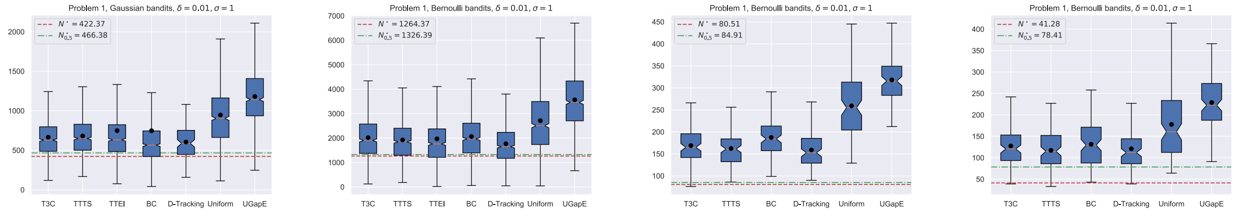

Figure 1: black dots represent means and oranges lines represent medians.

Sampling rule	T3C	TTTS	TTEI	BC	D-Tracking	Uniform	UGapE
Execution time (s)	1.6×10^{-5}	2.3×10^{-4}	1×10^{-5}	1.4×10^{-5}	1.3×10^{-3}	6×10^{-6}	5×10^{-6}

Table 1: average execution time in seconds for different sampling rules.

6 Numerical Illustrations

This section is aimed at illustrating our theoretical results and supporting the practical use of Bayesian sampling rules for fixed-confidence BAI.

We experiment with 3 Bayesian sampling rules: T3C, TTTS and TTEI with $\beta = 1/2$, against the Direct Tracking (D-Tracking) of Garivier and Kaufmann (2016) (which is adaptive to β), UGapE of Gabillon et al. (2012), and a uniform baseline. To make fair comparisons, we use the stopping rule (5) and associated recommendation rule for all of the sampling rules except for UGapE which has its own stopping rule.

We further include a top-two variant of the Best Challenger (BC) heuristic (see Ménard, 2019). BC selects the empirical best arm $\widehat{I}_n$ with probability β and the maximizer of $W_n(\widehat{I}_n, j)$ with probability $1-\beta$, but also performs forced exploration (selecting any arm sampled less than $\sqrt{n}$ times at round n). T3C can thus be viewed as a variant of BC in which no forced exploration is needed to converge to $\boldsymbol{\omega}^\beta$, due to the noise added by replacing $\widehat{I}_n$ with $I_n^{(1)}$. This randomization is crucial as BC without forced exploration can fail: we observed that on bandit instances with two identical sub-optimal arms, BC has some probability to alternate forever between these two arms and never stop.

We consider two simple instances with arm means given by $\boldsymbol{\mu}_1 = [0.5\ 0.9\ 0.4\ 0.45\ 0.44999]$, and $\boldsymbol{\mu}_2 = [1\ 0.8\ 0.75\ 0.7]$. We run simulations for both Gaussian ($\sigma = 1$) and Bernoulli bandits, with a risk parameter $\delta = 0.01$. Fig. 1 reports the empirical distribution of τ_δ under the different sampling rules, estimated over 1000 independent runs. We also indicate the values of $N^\star \triangleq \log(1/\delta)/\Gamma^\star$ (resp. $N^\star_{0.5} \triangleq \log(1/\delta)/\Gamma^\star_{0.5}$), the theoretical minimal number of samples needed for any strategy (resp. any 1/2-optimal strategy). In Appendix C, we illustrate how the empirical stopping time of T3C matches the theoretical one.

These figures provide several insights: (1) T3C is competitive with, and sometimes slightly better than TTTS/TTEI in terms of sample complexity. (2) The UGapE algorithm has a larger sample complexity than the uniform sampling rule, which highlights the importance of the stopping rule in the fixed-confidence setting. (3) The fact that D-Tracking performs best is not surprising, since it converges to $\boldsymbol{\omega}^{\beta^\star}$ and achieves minimal sample complexity. However, in terms of computation time, D-Tracking is much worse than others, as shown in Table 1, which reports the average execution time of one step of each sampling rule for $\boldsymbol{\mu}_1$ in the Gaussian case. (4) TTTS also suffers from computational costs, whose origins are explained in Sec. 2, unlike T3C or TTEI. Although TTEI is already computationally more attractive than TTTS, its practical benefits are limited to the Gaussian case, since the *Expected Improvement* (EI) does not have a closed form beyond this case and its approximation would be costly. In contrast, T3C can be applied for other distributions.

7 Conclusion

We have advocated the use of a Bayesian sampling rule for BAI. In particular, we proved that TTTS and a computationally advantageous approach T3C, are both β-optimal in the fixed-confidence setting, for Gaussian bandits. We further extended the Bayesian optimality properties (Russo, 2016) to more practical choices of models and prior distributions. In order to be optimal, these sampling rules would need the oracle tuning $\beta^\star = \arg\max_{\beta \in [0,1]} \Gamma^\star_\beta$, which is not feasible. In future work, we will investigate the efficient online tuning of β to circumvent this issue. We also wish to obtain explicit finite-time sample complexity bound for these Bayesian strategies, and justify the use of these appealing anytime sampling rules in the fixed-budget setting. The latter is often more plausible in application scenarios such as BAI for automated machine learning (Li et al., 2017; Shang et al., 2019).

Acknowledgements

The research presented was supported by European CHIST-ERA project DELTA, French National Research Agency projects BADASS (ANR-16-CE40-0002) and BOLD (ANR-19-CE23-0026-04), the Inria-CWI associate team 6PAC and Leids Universiteits Fonds / Drs. J.R.D. Kuikenga Fonds voor Mathematici travel grant number W19204-1-35.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. (2012). Online-to-confidence-set conversions and application to sparse stochastic bandits. In *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics (AIStats)*.

Audibert, J.-Y. and Bubeck, S. (2010). Best arm identification in multi-armed bandits. In *Proceedings of the 23rd Conference on Learning Theory (CoLT)*.

Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002). Finite-time analysis of the multi-armed bandit problem. *Machine Learning Journal*, 47(2-3):235–256.

Bubeck, S., Munos, R., and Stoltz, G. (2009). Pure exploration in multi-armed bandits problems. In *Proceedings of the 20th International Conference on Algorithmic Learning Theory (ALT)*.

Bubeck, S., Munos, R., and Stoltz, G. (2011). Pure exploration in finitely-armed and continuous-armed bandits. *Theoretical Computer Science*, 412(19):1832–1852.

Burnetas, A. N. and Katehakis, M. N. (1996). Optimal adaptive policies for sequential allocation problems. *Advances in Applied Mathematics*, 17(2):122–142.

Cappé, O., Garivier, A., Maillard, O. A., Munos, R., and Stoltz, G. (2013). Kullback-Leibler upper confidence bounds for optimal sequential allocation. *Annals of Statistics*, 41(3):1516–1541.

Carpentier, A. and Locatelli, A. (2016). Tight (lower) bounds for the fixed budget best arm identification bandit problem. In *Proceedings of the 29th Conference on Learning Theory (CoLT)*.

Even-dar, E., Mannor, S., and Mansour, Y. (2003). Action elimination and stopping conditions for reinforcement learning. In *Proceedings of the 20th International Conference on Machine Learning (ICML)*.

Gabillon, V., Ghavamzadeh, M., and Lazaric, A. (2012). Best arm identification: A unified approach to fixed budget and fixed confidence. In *Advances in Neural Information Processing Systems 25 (NIPS)*.

Garivier, A. and Kaufmann, E. (2016). Optimal best arm identification with fixed confidence. In *Proceedings of the 29th Conference on Learning Theory (CoLT)*.

Jamieson, K., Malloy, M., Nowak, R., and Bubeck, S. (2014). lil'UCB: An optimal exploration algorithm for multi-armed bandits. In *Proceedings of the 27th Conference on Learning Theory (CoLT)*.

Jun, K.-S. and Nowak, R. (2016). Anytime exploration for multi-armed bandits using confidence information. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*.

Kalyanakrishnan, S., Tewari, A., Auer, P., and Stone, P. (2012). PAC subset selection in stochastic multi-armed bandits. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*.

Karnin, Z., Koren, T., and Somekh, O. (2013). Almost optimal exploration in multi-armed bandits. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*.

Kaufmann, E. and Garivier, A. (2017). Learning the distribution with largest mean: two bandit frameworks. *ESAIM: Proceedings and Surveys*, 60:114–131.

Kaufmann, E. and Koolen, W. (2018). Mixture martingales revisited with applications to sequential tests and confidence intervals. *arXiv preprint arXiv:1811.11419*.

Li, L., Jamieson, K., DeSalvo, G., Talwalkar, A., and Rostamizadeh, A. (2017). Hyperband: Bandit-based configuration evaluation for hyperparameter optimization. In *Proceedings of the 5th International Conference on Learning Representations (ICLR)*.

Ménard, P. (2019). Gradient ascent for active exploration in bandit problems. *arXiv preprint arXiv:1905.08165*.

Qin, C., Klabjan, D., and Russo, D. (2017). Improving the expected improvement algorithm. In *Advances in Neural Information Processing Systems 30 (NIPS)*.

Russo, D. (2016). Simple Bayesian algorithms for best arm identification. In *Proceedings of the 29th Conference on Learning Theory (CoLT)*.

Shang, X., Kaufmann, E., and Valko, M. (2019). A simple dynamic bandit algorithm for hyper-parameter tuning. In *6th Workshop on Automated Machine Learning at International Conference on Machine Learning (ICML-AutoML)*.

Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285.