Thompson Exploration with Best Challenger Rule in Best Arm Identification

Jongyeong Lee LEE@MS.K.U-TOKYO.AC.JP
The University of Tokyo, RIKEN AIP

Junya Honda HONDA@I.KYOTO-U.AC.JP
Kyoto University, RIKEN AIP

Masashi Sugiyama SUGI@K.U-TOKYO.AC.JP
RIKEN AIP, The University of Tokyo

Editors: Berrin Yanıkoğlu and Wray Buntine

Abstract

This paper studies the fixed-confidence best arm identification (BAI) problem in the bandit framework in the canonical single-parameter exponential models. For this problem, many policies have been proposed, but most of them require solving an optimization problem at every round and/or are forced to explore an arm at least a certain number of times except those restricted to the Gaussian model. To address these limitations, we propose a novel policy that combines Thompson sampling with a computationally efficient approach known as the best challenger rule. While Thompson sampling was originally considered for maximizing the cumulative reward, we demonstrate that it can be used to naturally explore arms in BAI without forcing it. We show that our policy is asymptotically optimal for any two-armed bandit problems and achieves near optimality for general K-armed bandit problems for $K \geq 3$. Nevertheless, in numerical experiments, our policy shows competitive performance compared to asymptotically optimal policies in terms of sample complexity while requiring less computation cost. In addition, we highlight the advantages of our policy by comparing it to the concept of β-optimality, a relaxed notion of asymptotic optimality commonly considered in the analysis of a class of policies including the proposed one.

Keywords: Best arm identification, Thompson sampling, Multi-armed bandits

1. Introduction

As a formulation of reinforcement learning, multi-armed bandit (MAB) problems exemplify a trade-off between exploration and exploitation of knowledge. In traditional stochastic MAB problems, an agent plays an arm and observes a reward from the unknown but fixed distribution associated with the played arm. Although a large number of studies on MAB have been designed to maximize the cumulative rewards (Agrawal and Goyal, 2012; Slivkins et al., 2019), one might be interested only in the quality of a final decision rather than the performance of the overall plays. For example, one can consider the development of a new drug, where the researchers would aim to identify the most effective treatment from a set of alternatives before testing it on a large group of patients. When exploration and evaluation phases are separated in this way, it is known that a policy designed to maximize the cumulative rewards performs poorly (Bubeck et al., 2011). Such a setting is called pure exploration and several specialized policies have been proposed for this setting (Bubeck et al., 2009; Gabillon et al., 2012; Chen et al., 2014). In this paper, we consider the most standard fundamental formulation of the pure exploration problem, *best arm identification* (BAI), where the agent aims to identify the optimal arm that yields the largest mean reward (Maron and Moore, 1997; Even-Dar et al., 2006).

Two problem settings, the fixed-budget setting and the fixed-confidence setting, have been mainly considered in the BAI problems. In the fixed-budget setting, an agent aims to maximize the probability of successfully identifying the optimal arm within a fixed number of trials (Gabillon et al., 2012; Komiyama et al., 2022). On the other hand in the fixed-confidence setting, the agent aims to minimize the number of trials while ensuring that the probability of misidentifying the best arm is less than a fixed threshold (Kalyanakrishnan et al., 2012; Kuroki et al., 2020).

In the fixed-confidence setting, Garivier and Kaufmann (2016) provided a tight lower bound on the expected number of trials, which is also called the sample complexity, for canonical single-parameter exponential family (SPEF) bandit models including the Bernoulli distributions and Gaussian distributions with known variances. This bound represents the expected number of trials required to achieve a given level of confidence in identifying the best arm. Along with this lower bound on the sample complexity, they also proposed the Track-and-Stop (TaS) policy that tracks the optimal sampling proportion of arm plays and showed its asymptotic optimality. However, this policy requires solving a computationally expensive optimization at every round to obtain the optimal sampling proportion.

To address this limitation, several computationally efficient policies have been proposed that solve the optimization problem through a single gradient ascent in the online fashion (Ménard, 2019; Wang et al., 2021). However, most of these policies rely on forced exploration, where an arm is played a certain number of times to ensure that the empirical mean converges to its true value. While one can naturally specify the number of needed explorations for simple cases such as Bernoulli or Gaussian models, this becomes heavily nontrivial for general models where the variance of rewards may not be bounded. Recognizing the need for a more natural approach to exploration, Ménard (2019) emphasized the importance of finding policies that allow for exploration without the need for forced exploration. More recently, Barrier et al. (2022) proposed a sampling policy that naturally encourages exploration by employing an upper confidence bound. However, their algorithm is specifically designed for Gaussian bandits with known variance and exhibits slower convergence of the empirical mean compared to approaches that employ the forced exploration steps. As a result, their policy requires a larger number of samples in numerical experiments.

The BAI problems have also been considered in the Bayesian setting. Russo (2016) proposed top-two sampling rules which are adapted to solve the BAI problem. Generally in this approach, the leader (e.g., the currently best arm) is played with a fixed probability β, and the challenger (e.g., an arm selected by some randomized rule) is played with a probability of $1 - \beta$, where β is a predetermined hyperparameter. This approach allows for different configurations of the leader and the challenger in each round (Qin et al., 2017; Shang et al., 2020), for which more comprehensive examples can be found in Jourdan et al. (2022). A relaxed notion of optimality, β-optimality, has been commonly considered for top-two sampling rules. In other words, the sample complexity bounds of these β-optimal policies do not match the lower bound in general at the cost of their computational efficiency.

Contribution In this paper, we present a simple approach that combines a heuristic policy, a variant of the Best Challenger (BC) rule[1] introduced by Ménard (2019), with Thompson sampling (TS), a Bayesian policy originally introduced for cumulative reward minimization. Although it is known that a policy designed to maximize the cumulative rewards performs poorly when the exploration and evaluation phases are separated (Bubeck et al., 2011), we show that TS can still be used for the exploration part to solve the BAI problem. Our policy addresses the limitations of existing approaches, which often involve solving computationally expensive optimization problems (Garivier and Kaufmann, 2016) and/or require the forced exploration steps (Ménard, 2019; Wang et al., 2021). Therefore, our policy allows for a more computationally efficient and practical solution to the BAI problem.

It is important to note that our proposed policy does not achieve asymptotic optimality in all scenarios, similar to the β-optimal policies. Nevertheless, we prove that our policy achieves asymptotic optimality for any two-armed bandit problems, which distinguishes it from β-optimal policies. This unique characteristic of our policy offers its own advantages and strengths compared to (β-)optimal policies. The contributions of this paper are summarized as follows:

- We propose a computationally efficient policy for BAI problems in the SPEF bandits without the need for solving optimization problems, forcing explorations, and using additional hyperparameter β.
- We derive a sample complexity bound of the proposed policy for general K-armed SPEF bandits, which achieves the lower bound asymptotically for $K = 2$ and is numerically tighter than that of β-optimal policies for many instances for general K.
- We experimentally demonstrate the effectiveness of using TS as an exploration mechanism, which serves as a substitute for the forced exploration steps in the BAI problems.

Organization The rest of this paper is organized as follows. In Section 2, we formulate the BAI problems for the SPEF bandits and introduce the asymptotic optimality and TS. Next, in Section 3, we propose a simple policy called Best Challenger with Thompson Exploration (BC-TE), which is based on a variant of the best challenger policies described in previous works (Garivier and Kaufmann, 2016; Ménard, 2019). The sample complexity analysis of BC-TE is presented in Section 4, where we also compare its result with the asymptotic optimality and β-optimality. Furthermore, in Section 5, we provide simulation results that demonstrate the effectiveness of BC-TE, showing competitive performance in terms of the sample complexity and superior computational efficiency compared to other asymptotically (β-)optimal policies.

2. Preliminaries

In this section, we formulate the BAI problem for the model of SPEF and the asymptotic lower bound on the sample complexity. Then we introduce the stopping rule considered in Garivier and Kaufmann (2016).

[1] The BC rule considered in Garivier and Kaufmann (2016) and Ménard (2019) can be seen as a variant of top-two sampling since it also plays either the leader or the challenger at every round. However, the key distinction lies in the deterministic nature of BC, which is solely determined by historical information and does not involve any randomness introduced by a hyperparameter β. In this paper, the BC rule refers to a policy without hyperparameter β, while top-two sampling refers to that with β.

2.1. Notation and SPEF bandits

We consider the K-armed bandit model where each arm belongs to a canonical SPEF with a form

$$\mathcal{P} = \left\{ (\nu_{\theta_i})_{i=1}^K : \frac{\mathrm{d}\nu_{\theta_i}}{\mathrm{d}\xi}(x) = \exp(\theta_i x - A(\theta_i)), \theta_i \in \Theta, \forall i \in [K] \right\}, \tag{1}$$

where $\Theta \subset \mathbb{R}$ denotes the parameter space, ξ is some reference measure on $\mathbb{R}$, $A : \Theta \to \mathbb{R}$ is a convex and twice differentiable function, and $[K] := \{1, \ldots, K\}$. For this model, we can write the expected reward of an arm as $\mu(\theta) = A'(\theta)$ and the KL divergence between two distributions as follows (Cappé et al., 2013):

$$\mathrm{KL}(\nu_{\theta_1}, \nu_{\theta_2}) = \mu(\theta_1)(\theta_1 - \theta_2) + A(\theta_2) - A(\theta_1),$$

which induces a divergence function d on $A'(\theta)$ defined by $d(\mu(\theta), \mu(\theta')) = \mathrm{KL}(\nu_\theta, \nu_{\theta'})$. Following the notation used in Garivier and Kaufmann (2016), a bandit instance $\nu = (\nu_{\theta_1}, \ldots, \nu_{\theta_K})$ is identified with the means $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$. We denote a set of SPEF bandit models with a unique optimal arm by $\mathcal{S}$. Therefore, for any $\boldsymbol{\mu} \in \mathcal{S}$, $\arg\max_{i \in [K]} \mu_i$ is a singleton and we assume that $\mu(\theta_1) > \mu(\theta_2) \geq \cdots \geq \mu(\theta_K)$ without loss of generality. Then, we denote the current maximum likelihood estimate of $\boldsymbol{\mu}$ at round t by $\hat{\boldsymbol{\mu}}(t) = (\hat{\mu}_1(t), \ldots, \hat{\mu}_K(t))$ for $\hat{\mu}_i(t) = \frac{1}{N_i(t)} \sum_{s=1}^t x_{i,N_i(s)}$, where $N_i(t)$ denotes the number of rounds the arm i is played until round t and $x_{i,n}$ denotes the n-th observation from the arm $i \in [K]$. By abuse of notation, we sometimes denote $\hat{\mu}_i(t)$ by $\hat{\mu}_{i,N_i(t)}$ to specify the number of plays of the arm i.

In the fixed-confidence setting, a policy is said to be δ probably approximately correct (δ-PAC) when it satisfies $\mathbb{P}[i(\tau_\delta) \neq 1 \vee \tau_\delta = \infty] \leq \delta$. Here, τ_δ is the number of trials until the sampling procedure stops for a given risk parameter δ, and $i(t)$ denotes the chosen arm at round $t \in \mathbb{N}$. Thus, the agent aims to build a δ-PAC policy while minimizing the sample complexity $\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]$.

2.2. Asymptotic lower bound on the sample complexity

Garivier and Kaufmann (2016) showed that any δ-PAC policy satisfies for any $\delta \in (0, 1)$ and $\boldsymbol{\mu} \in \mathcal{S}$

$$\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta] \geq T^*(\boldsymbol{\mu}) \log\left(\frac{1}{2.4\delta}\right), \tag{2}$$

where

$$T^*(\boldsymbol{\mu}) := \left(\sup_{\boldsymbol{w} \in \Sigma_K} \min_{i \neq 1} f_i(\boldsymbol{w}; \boldsymbol{\mu}) \right)^{-1}. \tag{3}$$

Here, the function f_i is defined as

$$\begin{aligned} f_i : \Sigma_K \times \mathcal{S} &\to \mathbb{R}_+ \\ (\boldsymbol{w}; \boldsymbol{\mu}) &\mapsto w_1 d(\mu_1, \mu_{1,i}^{\boldsymbol{w}}) + w_i d(\mu_i, \mu_{1,i}^{\boldsymbol{w}}), \end{aligned} \tag{4}$$

where $\mu_{1,i}^{\boldsymbol{w}} = \frac{w_1}{w_1 + w_i}\mu_1 + \frac{w_i}{w_1 + w_i}\mu_i$ is a weighted mean and $\Sigma_K = \{\boldsymbol{w} \in [0,1]^K : \sum_{i=1}^K w_i = 1\}$ denotes the probability simplex. We define $f_i(x; \cdot) = -\infty$ for $x \notin \Sigma_K$ and $i \in [K]$ for simplicity. Through the derivation of (2), Garivier and Kaufmann (2016) also showed that the maximizer $\boldsymbol{w}^* = \boldsymbol{w}^*(\boldsymbol{\mu}) := \arg\max_{\boldsymbol{w} \in \Sigma_K} \min_{i \neq 1} f_i(\boldsymbol{w}; \boldsymbol{\mu})$ indicates the optimal sampling proportion of arm plays, that is, it is necessary to play arms to bring $\boldsymbol{w}^t := \left(\frac{N_1(t)}{t}, \dots, \frac{N_K(t)}{t}\right)$ closer to $\boldsymbol{w}^*$ for matching the lower bound. The convergence of $\boldsymbol{w}^t$ towards $\boldsymbol{w}^*$ is widely recognized as a crucial factor for achieving optimal performance in the BAI problem (Ménard, 2019; Wang et al., 2021).

Along with the lower bound in (2), a policy is said to be asymptotically optimal if it satisfies

$$\limsup_{\delta \to 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]}{\log(1/\delta)} \leq T^*(\boldsymbol{\mu}).$$

Garivier and Kaufmann (2016) proposed the Track-and-Stop (TaS) policy, which tracks the optimal proportions $\boldsymbol{w}^*$ at every round, and showed its asymptotic optimality. Since the true mean reward $\boldsymbol{\mu}$ is unknown in practice, the TaS policy tracks the plug-in estimates $\boldsymbol{w}^*(\hat{\boldsymbol{\mu}}(t))$. This means that the TaS policy essentially requires solving the minimax optimization problem at every round to find $\boldsymbol{w}^*(\hat{\boldsymbol{\mu}}(t))$. Although some computational burden can be alleviated by using the solution from the previous round as an initial solution, the TaS policy remains computationally expensive due to the presence of the inverse function of the KL divergence.

On the other hand, a relaxed optimality notion, β-optimality, has been considered in top-two sampling rules, where the leader is played with a predefined probability $\beta \in (0, 1)$ (Russo, 2016; Qin et al., 2017; Shang et al., 2020; Jourdan et al., 2022). Here, a policy is said to be asymptotically β-optimal if it satisfies

$$\lim_{t \to \infty} w_1^t \to \beta \text{ and } \limsup_{\delta \to 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]}{\log(1/\delta)} \leq T^\beta(\boldsymbol{\mu}),$$

where

$$T^\beta(\boldsymbol{\mu}) := \left(\sup_{\boldsymbol{w} \in \Sigma_K, w_1 = \beta} \min_{i \neq 1} f_i(\boldsymbol{w}; \boldsymbol{\mu}) \right)^{-1}. \tag{5}$$

From its definition, $T^*(\boldsymbol{\mu}) = \min_{\beta \in [0,1]} T^\beta(\boldsymbol{\mu})$ holds. Thus, the β-optimality does not necessarily imply the optimality in the sense of (2) unless β is equal to $w_1^*(\boldsymbol{\mu})$. Still, $\beta = 1/2$ is usually employed since $T^*(\boldsymbol{\mu}) \leq T^{1/2}(\boldsymbol{\mu}) \leq 2T^*(\boldsymbol{\mu})$ holds, that is, $T^{1/2}(\boldsymbol{\mu})$ is at most two times larger than that of optimal policies (see Russo, 2016, Lemma 3).

2.3. Stopping rule

One important question is when an agent should terminate the sampling procedure, which is usually related to a statistical test. Garivier and Kaufmann (2016) considered the generalized likelihood ratio statistic that has a closed-form expression for the exponential family. Based on this statistic, they proposed Chernoff's stopping rule which is written as

$$\tau_\delta = \inf \left\{ t \in \mathbb{N} : \max_{a \in [K]} \min_{b : \hat{\mu}_a(t) \geq \hat{\mu}_b(t)} t f_{a,b}(\boldsymbol{w}^t; \hat{\boldsymbol{\mu}}(t)) > \beta(t, \delta) \right\}, \tag{6}$$

where $f_{a,b}(\boldsymbol{w}; \boldsymbol{\mu}) := w_a d(\mu_a, \mu_{a,b}^{\boldsymbol{w}}) + w_b d(\mu_b, \mu_{a,b}^{\boldsymbol{w}})$ for $\mu_a \geq \mu_b$ and $\beta(t, \delta)$ is a threshold to be tuned appropriately. Therefore, several thresholds $\beta(t, \delta)$ have been proposed (Garivier and Kaufmann, 2016; Ménard, 2019; Jedra and Proutiere, 2020; Kaufmann and Koolen, 2021). In this paper, we simply utilize the deviational threshold $\beta(t, \delta) = \log\left(\frac{Ct^\alpha}{\delta}\right)$ for $\alpha > 1$ and some constants $C = C(\alpha, K)$ since it was shown that using Chernoff's stopping rule with this threshold ensures the δ-PAC of any policies for the SPEF (see Garivier and Kaufmann, 2016, Propostion 12).

2.4. Thompson sampling with the Jeffreys prior

In the regret minimization problem, Thompson sampling has been shown to be asymptotically optimal for various reward models (Kaufmann et al., 2012; Honda and Takemura, 2014; Riou and Honda, 2020; Lee et al., 2023). For the SPEF bandits, TS with the Jeffreys prior was shown to be asymptotically optimal (Korda et al., 2013). The Jeffreys prior is a noninformative prior that is invariant under any reparameterization (Robert et al., 2009), which is written for the model in (1) by

$$\pi_{\mathrm{j}}(\theta) \propto \sqrt{|I(\theta)|} = \sqrt{|A''(\theta)|},$$

for the Fisher information $I(\theta)$.

Under the Jeffreys prior, the posterior on θ after n observations is given by

$$\pi(\theta|x_1,\ldots,x_n) \propto \sqrt{|A''(\theta)|}\exp\left(\theta\sum_{m=1}^{n} x_m - nA(\theta)\right). \tag{7}$$

For more details on the Jeffreys prior, we recommend referring to Robert et al. (2009) and Ghosh (2011), as well as the reference therein. Additionally, one can find more specific configurations on Thompson sampling with the Jeffreys prior for SPEF bandits in Korda et al. (2013).

3. Best Challenger with Thompson Exploration

In this section, we aim to build a δ-PAC policy that does not rely on the forced exploration steps. To achieve this, we utilize TS with the Jeffreys prior as a tool to encourage the exploration of arms in a natural manner.

3.1. The use of the best challenger rule

Here, we first introduce the intuition behind the best challenger rule.

For the sake of simplicity, we define a concave objective function $g(\boldsymbol{w};\boldsymbol{\mu}) := \min_{i\neq 1} f_i(\boldsymbol{w};\boldsymbol{\mu})$ for $x \in \Sigma_K$ and $g(x;\cdot) = -\infty$ for $x \notin \Sigma_K$. Then, (3) can be rewritten as

$$(T^*(\boldsymbol{\mu}))^{-1} = \sup_{\boldsymbol{w}\in\Sigma_K} g(\boldsymbol{w};\boldsymbol{\mu}) = g(\boldsymbol{w}^*;\boldsymbol{\mu}).$$

As discussed in Section 2.1, one can achieve the asymptotic optimality by moving the empirical proportion $\boldsymbol{w}^t$ closer to the optimal proportion $\boldsymbol{w}^*$. Since the optimal proportion $\boldsymbol{w}^*$ is a point that maximizes g, moving $\boldsymbol{w}^t$ in the direction of increasing g is a reasonable idea to reduce the gap between $\boldsymbol{w}^t$ and $\boldsymbol{w}^*$. As $\boldsymbol{w}^*$ is a solution to a convex optimization problem, a natural approach is to apply a gradient method to iteratively update $\boldsymbol{w}^t$, which would bring $\boldsymbol{w}^t$ to $\boldsymbol{w}^*$ without explicitly solving complex optimization problems. Although g is not differentiable, it can be expected that playing arms to track a subgradient of g would achieve the lower bound since g is concave.[2]

Here, we say that $\boldsymbol{v}$ is a subgradient of the concave function g at the point $(\boldsymbol{w};\boldsymbol{\mu})$ if

$$\forall \boldsymbol{w}' \in \Sigma_K,\ g(\boldsymbol{w}';\boldsymbol{\mu}) \le g(\boldsymbol{w};\boldsymbol{\mu}) + \boldsymbol{v}^\top(\boldsymbol{w}'-\boldsymbol{w}).$$

[2] In the strict sense, we should use the term subgradient to minimize the convex function $-g$ or supergradient to maximize the concave function g. However, we use the term subgradient for g since the term subgradient is more popular, and the use of $-g$ needlessly degrades the readability.

The subdifferential $\partial g(\boldsymbol{w};\boldsymbol{\mu})$ is the set of all such subgradients. The following lemma shows that the subgradients of the objective function g are expressed as the sum of all-ones vector $\mathbf{1}$ and convex combinations of the gradients $\nabla_{\boldsymbol{w}} f(\boldsymbol{w};\boldsymbol{\mu})$ of f with respect to $\boldsymbol{w}$. The proofs of all lemmas and theorems are given in the supplementary material.

Lemma 1 *The subdifferential ∂g of g with respect to $\boldsymbol{w} \in \operatorname{Int}\Sigma_K$ for given $\boldsymbol{\mu} \in \mathcal{S}$ is such that*

$$\partial g(\boldsymbol{w};\boldsymbol{\mu}) = \left\{ \sum_{i \in \mathcal{J}(\boldsymbol{w};\boldsymbol{\mu})} \lambda_i \nabla_{\boldsymbol{w}} f_i(\boldsymbol{w};\boldsymbol{\mu}) + r\mathbf{1} : \sum_{i \in \mathcal{J}(\boldsymbol{w};\boldsymbol{\mu})} \lambda_i = 1, \lambda_i \geq 0, r \in \mathbb{R} \right\},$$

where $\mathcal{J}(\boldsymbol{w};\boldsymbol{\mu}) := \arg\min_{i\neq 1} f_i(\boldsymbol{w};\boldsymbol{\mu})$ denotes the set of challengers, f_i is defined in (4), and $\operatorname{Int}\Sigma_K$ denotes the interior of the probability simplex.

By letting $r = 0$ and $\lambda_i = 1/|\mathcal{J}(\boldsymbol{w};\boldsymbol{\mu})|$ for any $i \in [K]$ in Lemma 1, we can obtain a subgradient $\boldsymbol{v}$ for $\boldsymbol{\mu} \in \mathcal{S}$ satisfying

$$v_i(\boldsymbol{w};\boldsymbol{\mu}) = \begin{cases} 0 & \text{if } i \notin \{1\} \cup \mathcal{J}(\boldsymbol{w};\boldsymbol{\mu}), \\ \frac{1}{|\mathcal{J}(\boldsymbol{w};\boldsymbol{\mu})|} \sum_{j \in \mathcal{J}(\boldsymbol{w};\boldsymbol{\mu})} d(\mu_i, \mu_{i,j}^{\boldsymbol{w}}) & \text{if } i = 1, \\ \frac{1}{|\mathcal{J}(\boldsymbol{w};\boldsymbol{\mu})|} d(\mu_i, \mu_{1,i}^{\boldsymbol{w}}) & \text{if } i \in \mathcal{J}(\boldsymbol{w};\boldsymbol{\mu}). \end{cases}$$

Since our objective is to maximize the objective function g, one can easily consider a greedy approach that plays an arm with the maximum subgradient, that is

$$i(t) \in \arg\max_{i \in [K]} v_i(\boldsymbol{w}^t; \hat{\boldsymbol{\mu}}(t)),$$

which plays either the currently best arm $m(t) = \arg\max_{i\in[K]} \hat{\mu}_i(t)$ or the challenger $j(t) \in \mathcal{J}_t = \mathcal{J}(\boldsymbol{w}^t; \hat{\boldsymbol{\mu}}(t))$ at round t. For the arbitrarily chosen challenger

$$j(t) = \arg\min_{i \neq m(t)} f_i(\boldsymbol{w}^t; \hat{\boldsymbol{\mu}}(t)), \tag{8}$$

a variant of the Best Challenger (BC) rule introduced by Ménard (2019) can be expressed as

$$i(t) = \begin{cases} m(t) & \text{if } d(\hat{\mu}_{m(t)}(t), \hat{\mu}_{m(t),j(t)}(t)) \geq d(\hat{\mu}_{j(t)}(t), \hat{\mu}_{m(t),j(t)}(t)), \\ j(t) & \text{otherwise}, \end{cases}$$

where we denote $\hat{\mu}_{a,b}^{\boldsymbol{w}^t}(t) = \frac{w_a^t}{w_a^t + w_b^t} \hat{\mu}_a(t) + \frac{w_a^t}{w_a^t + w_b^t} \hat{\mu}_b(t)$ by $\hat{\mu}_{a,b}(t)$ for notational simplicity. This simple heuristic with forced exploration was shown to be computationally very efficient and showed excellent empirical performance in the BAI problems despite its lack of theoretical guarantee.

Note that the use of subgradients instead of solving the optimization problem at every round has been considered by Ménard (2019), where they applied the online mirror ascent method, and by Wang et al. (2021), where they applied the Frank-Wolfe-type algorithm to optimize the non-smooth concave objective function g. It is worth noting that both policies are shown to be asymptotically optimal for various BAI problems. Nevertheless, the families of top-two samplings (including BC rules) are especially simple, and for this reason, β-optimality is still considered despite its suboptimality (Jourdan et al., 2022, 2023; Mukherjee and Tajer, 2022).

Algorithm 1: Best challenger with Thompson Exploration (BC-TE)

```
Initialization: Play every arm twice and set $\boldsymbol{w}^{2K} = \frac{1}{K}$ and $t = 2K$.
while stopping criterion is satisfied do
    Sample $\tilde{\mu}_i(t)$ from the posterior distribution in (7).
    Set $m(t) = \arg\max_{i\in[K]} \hat{\mu}_i(t)$ and $\tilde{m}(t) = \arg\max_{i\in[K]} \tilde{\mu}_i(t)$.
    if $m(t) = \tilde{m}(t)$ then
        Find the subgradient $\boldsymbol{v}^t$ of $g(\boldsymbol{w}^t, \hat{\boldsymbol{\mu}}^t)$.
        Play $i(t+1) \in \arg\max_{i\in[K]} v_i^t$ and observe the reward.
    else
        Play $i(t+1) \in \arg\min_{i\in\{m(t),\tilde{m}(t)\}} N_i(t)$.
        Update $t = t+1$, $\hat{\boldsymbol{\mu}}^t$ and $\boldsymbol{w}^t$.
    end
end
```

3.2. The use of Thompson exploration

Although the policies using gradient methods are asymptotically optimal and/or simple, they still include the forced exploration steps to ensure that the empirical means converge to their true values. Therefore, it is worth finding a natural way to explore without forcing policies to explore. Although Barrier et al. (2022) replaced the forced exploration steps by using the upper confidence bound-based approach, their policy was restricted to the Gaussian models and exhibited large sample complexity in numerical experiments. Instead, in this paper, we employ TS as an exploration tool to eliminate the forced exploration steps, which can be applied to any SPEF bandits and performs well in practice. To be precise, we play an arm according to the BC rule only when the empirical best arm and the best arm under the posterior sample agree, that is,

$$i(t) = \begin{cases} \arg\max_{i\in[K]} v_i(\boldsymbol{w}^t; \hat{\boldsymbol{\mu}}(t)) & \text{if } m(t) = \tilde{m}(t) := \arg\max_{i\in[K]} \tilde{\mu}_i(t), \qquad \text{(BC)} \\ \arg\min_{i\in\{m(t),\tilde{m}(t)\}} N_i(t) & \text{otherwise,} \qquad \text{(Thompson exploration)} \end{cases}$$

where $\tilde{\mu}_i(t)$ denotes the posterior sample of the arm i generated by the posterior in (7). As the number of plays increases, the probability of observing a sample that deviates significantly from the current empirical mean decreases exponentially. In other words, if an arm is played only a few times, its posterior sample is more likely to deviate from its empirical mean. This discrepancy between the best arm under the posterior sample and the empirical best arm can be a guide to the policy for further exploration. By selecting an arm with a small number of plays only when the empirical best arm and the best arm under the posterior sample disagree, we can ensure the convergence of the empirical means to their true values without relying on forced exploration, which is formulated in Section 4. The proposed algorithm, called Best Challenger with Thompson Exploration (BC-TE), is described in Algorithm 1. Notice that BC-TE plays every arm twice at initialization steps to avoid an improper posterior distribution.

4. Main Theoretical Results

In this section, we show the effectiveness of TE and prove that BC-TE is nearly optimal, similar to β-optimality.

4.1. Main theorems

Firstly, let us define a random variable $T_B \in \mathbb{N}$ such that for any $\epsilon < \frac{\mu_1 - \mu_2}{2}$

$$T_B = \inf\{T \in \mathbb{N} : \forall s \geq T, \forall i \in [K], |\hat{\mu}_i(s) - \mu_i| \leq \epsilon\}. \tag{9}$$

Therefore, the empirical mean estimate $\hat{\boldsymbol{\mu}}(t)$ is sufficiently close to its true value $\boldsymbol{\mu}$ for all rounds after T_B. The theorem below shows the expected value of T_B is finite.

Theorem 2 *Under Algorithm 1, it holds that*

$$\mathbb{E}[T_B] \leq \mathcal{O}(K^2 d_\epsilon^{-2}),$$

where

$$d_\epsilon := \min_{i \in [K]} \min(d(\mu_i + \epsilon, \mu_i), d(\mu_i - \epsilon, \mu_i)). \tag{10}$$

From the definition of T_B, one can expect that the sampling rule will behave as expected after T_B rounds since the estimated means are close to the true ones. Note that T_B is not a stopping time with respect to the sequence of observations and we need a careful analysis for its expectation. The key property used in the proof is that BC-TE always plays an arm that increases the objective function $g(\boldsymbol{w}^t; \hat{\boldsymbol{\mu}}(t))$ at every round t. Since most arguments in the proof of Theorem 2 do not depend on the procedure when TE does not occur, we can expect that one can derive the same result for Theorem 2 for any policy designed to increase the objective function at every round such as Frank-Wolfe sampling (Wang et al., 2021). Then, the sample complexity of BC-TE can be upper bounded as follows.

Theorem 3 *Let $\alpha \in [1, e/2]$ and $r(t) = \mathcal{O}(t^\alpha)$. Using the Chernoff's stopping rule in (6) with $\beta(t, \delta) = \log(r(t)/\delta)$ under Algorithm 1,*

$$\limsup_{\delta \to 0} \frac{\mathbb{E}[\tau_\delta]}{\log(1/\delta)} \leq \alpha \underline{T}(\boldsymbol{\mu}),$$

where

$$\underline{T}(\boldsymbol{\mu}) := \left(\sup_{\boldsymbol{w} \in \Sigma_K, \frac{w_2}{w_1 + w_2} = \gamma} \min_{i \neq 1} f_i(\boldsymbol{w}; \boldsymbol{\mu}) \right)^{-1} \tag{11}$$

for γ satisfying

$$d(\mu_1, (1-\gamma)\mu_1 + \gamma\mu_2) = d(\mu_2, (1-\gamma)\mu_1 + \gamma\mu_2). \tag{12}$$

From the definition of $T^*(\boldsymbol{\mu})$ in (3), one can see the suboptimality of BC-TE from $\underline{T}(\boldsymbol{\mu}) \geq T^*(\boldsymbol{\mu})$, which indicates that BC-TE may be not always optimal, as it only achieves optimality when the condition $\gamma = \frac{w_2^*}{w_1^* + w_2^*}$ is true. This observation is akin to the result for β-optimality.

4.2. Comparison with β-optimality and asymptotic optimality

Recall that the quantity $T^\beta(\boldsymbol{\mu})$ in (5) demonstrates that β-optimality is achieved when the allocation of the optimal arm is β. On the other hand, $\underline{T}(\boldsymbol{\mu})$ considers the scenario where $\frac{w_2}{w_1 + w_2} = \gamma$, which is the best ratio between the best arm and the second best arm to distinguish them. Both notions are more relaxed compared to asymptotic optimality, and it is not possible to determine definitively which one is better in general.

However, it is important to note that our policy does not require prior knowledge of γ, differently from existing β-optimal policies that take β as an input to the algorithm (Russo, 2016; Shang et al., 2020; Jourdan et al., 2022; Jourdan and Degenne, 2022). Therefore, if there is no prior knowledge of β, using BC-TE would have its own advantages over β-optimal policies. In general, it is challenging to compare the quantities $\underline{T}$ and T^β for $\beta = 1/2$ analytically due to the complex formulation of KL divergence and the optimization problem in (5) and (11). For this reason, in Section 4.2.3, we provide numerical comparisons for $K \geq 2$ across various SPEF bandits.

Then, the natural question is the relationship between $T^*(\boldsymbol{\mu})$ and $\underline{T}(\boldsymbol{\mu})$. Unlike the β-optimality where β does not depend on the bandit instance, the quantity $\underline{T}(\boldsymbol{\mu})$ is problem-dependent since γ is determined by μ_1, μ_2, and $d(\cdot,\cdot)$. Here, we provide a rough comparison with the quantity $T^*(\boldsymbol{\mu})$.

4.2.1. TWO-ARMED BANDITS

When $K = 2$, (3) can be written as

$$(T^*(\boldsymbol{\mu}))^{-1} = \sup_{\alpha \in (0,1)} \alpha d(\mu_1, \mu^\alpha) + (1-\alpha) d(\mu_2, \mu^\alpha),$$

where $\mu^\alpha = (1-\alpha)\mu_1 + \alpha\mu_2$. Here, Garivier and Kaufmann (2016) showed that the maximum is reached at α^* satisfying $d(\mu_1, \mu^{\alpha^*}) = d(\mu_2, \mu^{\alpha^*})$. From (12), one can directly see that $\gamma = \alpha^*$ holds, which implies $\underline{T} = T^*(\boldsymbol{\mu})$ for any $\boldsymbol{\mu} \in \mathcal{S}$ if $K = 2$. A more detailed discussion is given in the supplementary material for the sake of completeness.

4.2.2. GAUSSIAN BANDITS

When $\boldsymbol{\mu}$ belongs to the Gaussian distributions with known variance $\sigma^2 > 0$, the KL divergence takes a simple form of $d(\mu, \mu') = \frac{(\mu-\mu')^2}{2\sigma^2}$. This allows us to derive a more explicit comparison with asymptotic optimality.

Lemma 4 *Let $\Delta_i = \mu_1 - \mu_i$ for $i \neq 1$ and $\Delta_1 = \Delta_2$. When $\boldsymbol{\mu}$ belongs to the Gaussian distributions with known variance $\sigma^2 > 0$,*

$$\underline{T}(\boldsymbol{\mu}) = \sum_{i=1}^{K} \frac{4\sigma^2}{\Delta_i^2 + (\Delta_i^2 - \Delta_2^2)}.$$

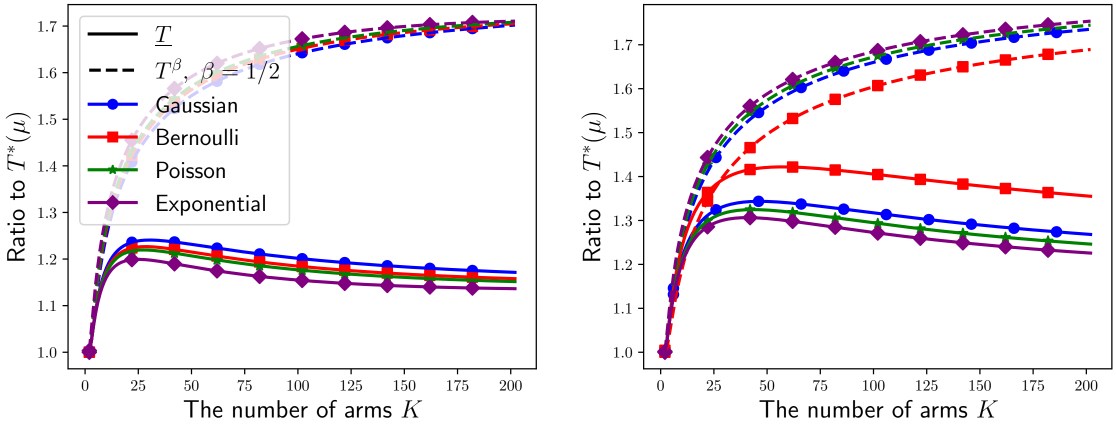
Here, Garivier and Kaufmann (2016) showed the following inequalities for the Gaussian bandits

$$\sum_{i=1}^{K} \frac{2\sigma^2}{\Delta_i^2} \leq T^*(\boldsymbol{\mu}) \leq 2\sum_{i=1}^{K} \frac{2\sigma^2}{\Delta_i^2},$$

which directly implies that

$$T^*(\boldsymbol{\mu}) \leq \underline{T}(\boldsymbol{\mu}) \leq 2T^*(\boldsymbol{\mu}), \tag{13}$$

where the left equality holds when $w_1^*(\boldsymbol{\mu}) = w_2^*(\boldsymbol{\mu})$ and the right equality holds only when $\mu_2 = \cdots = \mu_K$. Notice that the same result as (13) holds for T^β with $\beta = 1/2$ (Russo, 2016), though $T^{1/2}(\boldsymbol{\mu}) \neq \underline{T}(\boldsymbol{\mu})$ holds in general.



(a) Instance $\boldsymbol{\mu}^{(1)}$ with varying K.

(b) Instance $\boldsymbol{\mu}^{(2)}$ with varying K.

Figure 1: The ratio of $\underline{T}(\boldsymbol{\mu})$ and $T^{1/2}(\boldsymbol{\mu})$ to $T^*(\boldsymbol{\mu})$ for different reward distributions.

4.2.3. NUMERICAL COMPARISON FOR VARIOUS SPEF BANDITS

Here, we compare the quantities $\underline{T}(\boldsymbol{\mu})$, $T^*(\boldsymbol{\mu})$, and $T^\beta(\boldsymbol{\mu})$ with $\beta = 1/2$ across different bandit models and varying numbers of arms. Specifically, we consider two instances $\boldsymbol{\mu}^{(1)}$ and $\boldsymbol{\mu}^{(2)}$ for Gaussian (with unit variance), Bernoulli, Poisson, and Exponential distributions.

We consider two instances, $\boldsymbol{\mu}^{(1)} = (0.3, 0.21, 0.21 - 0.001, \ldots, 0.21 - 0.001(K-2))$ and $\boldsymbol{\mu}^{(2)} = (0.9, 0.7, 0.7 - 0.001, \ldots, 0.7 - 0.001(K-2))$. For example, when $K = 4$, $\boldsymbol{\mu}^{(1)} = (0.3, 0.21, 0.209, 0.208)$ and $\boldsymbol{\mu}^{(2)} = (0.9, 0.7, 0.699, 0.698)$ are considered. In Figure 1, the solid line represents the ratio $\underline{T}(\boldsymbol{\mu})/T^*(\boldsymbol{\mu})$, while the dashed line represents the ratio $T^{1/2}(\boldsymbol{\mu})/T^*(\boldsymbol{\mu})$. Each line corresponds to a different reward model, which is distinguished by a different color and marker. From Figure 1, we can observe that $\underline{T}(\boldsymbol{\mu})$ keeps being close to T^*, while $T^{1/2}(\boldsymbol{\mu})$ does not for large K. This contrasting behavior indicates the advantage of BC-TE over β-optimal policies, particularly for large K, as it suggests that BC-TE enjoys a much tighter upper bound on its sample complexity. Additional comparisons are provided in the supplementary material.

5. Simulation Results

In this section, we present numerical results to demonstrate the performance of BC-TE.

Compared policies We compare the performance of BC-TE with other policies, where $\diamond$ denotes that the policy requires forced exploration. For policies with † and ‡, we used the implementation by Koolen (2019) and by Wang et al. (2021), respectively.

- Track-and-Stop$^{\dagger,\diamond}$ (TaS): an asymptotically optimal policy that solves the optimization problem in (3) at every round, which is computationally costly (Garivier and Kaufmann, 2016). Here, we focus on the TaS policy with D-tracking (T-D) in our experiment.

- Lazy Mirror Ascent$^{\dagger,\diamond}$ (LMA): a computationally efficient and asymptotically optimal policy that performs a single gradient ascent in an online fashion (Ménard, 2019).

- AdaHedge vs Best Response† (AHBR): an asymptotically optimal policy that solves the optimization problem as an unknown game (Degenne et al., 2019).
- Optimistic TaS‡ (O-C): The optimistic TaS policies with C-tracking proposed by Degenne et al. (2019), which is known to be very computationally expensive.
- Frank-Wolfe Sampling$^{\ddagger,\diamond}$ (FWS): an asymptotically optimal policy that just relies on a single iteration FW algorithm instead of solving the optimization problems in (3) at every round (Wang et al., 2021).
- Round Robin (RR): a simple baseline that samples arms in a round-robin manner.
- Top-Two Transportation Cost (T3C): a computationally efficient asymptotically β-optimal top-two policy based on TS (Shang et al., 2020). Notice that its β-optimality was extended to bounded distributions by Jourdan et al. (2022) and we set $\beta = 1/2$.

In addition, we implement a modified version of FWS, called FWS-TE, where we replace the forced exploration step in FWS with our Thompson exploration step. This adaptation is based on the discussion below Theorem 2 that TE can be used for policies designed to increase the objective function g at every round.

Stopping rule Following the experiments in the previous researches (Garivier and Kaufmann, 2016; Degenne et al., 2019; Ménard, 2019; Wang et al., 2021), we considered the same threshold $\beta(t, \delta) = \log((\log(t) + 1)/\delta)$.

General setup Here, we provide the empirical sample complexities of various policies for a range of risk levels $\delta \in \{0.2, 0.1, 0.01, 0.001\}$ averaged over 3,000 independent runs. Following Degenne et al. (2019), we consider the practical version of the lower bound (PLB), which refers to the first round where $tg(\boldsymbol{w}^*; \boldsymbol{\mu}) \geq \beta(t, \delta)$ is satisfied. Hence, this practical lower bound indicates the earliest round where the generalized likelihood ratio statistic approximately crosses the threshold, and is defined as round s where $s = \beta(s, \delta)T^*(\boldsymbol{\mu})$ holds. Recall that the lower bound (LB) is given as $T^*(\boldsymbol{\mu}) \log\left(\frac{1}{2.4\delta}\right)$ according to (2).

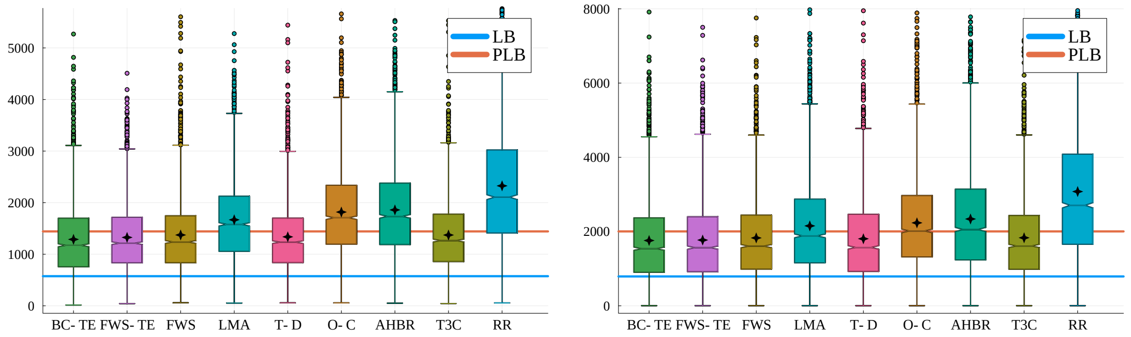
Bernoulli bandits In the first experiment, we consider the 5-armed Bernoulli bandit instance $\boldsymbol{\mu}_5^{\mathrm{B}} = (0.3, 0.21, 0.2, 0.19, 0.18)$ where $\boldsymbol{w}^*(\boldsymbol{\mu}_5^{\mathrm{B}}) = (0.43, 0.25, 0.18, 0.13,\ 0.10)$. This instance was considered in previous researches (Garivier and Kaufmann, 2016; Ménard, 2019; Wang et al., 2021).

Gaussian bandits In the second experiment, we consider the 4-armed Gaussian bandit instance $\boldsymbol{\mu}_4^{\mathrm{G}} = (1.0, 0.85, 0.8, 0.7)$ with unit variance $\sigma^2 = 1$ where $\boldsymbol{w}^*(\boldsymbol{\mu}_4^{\mathrm{G}}) = (0.41, 0.38, 0.15, 0.06)$. This instance was studied in Wang et al. (2021).

Results The overall results are presented in Table 1. Although our proposed policy BC-TE does not achieve the asymptotic optimality in general, it exhibits a better empirical performance than other optimal policies across most risk parameters, especially when large δ is considered. Interestingly, Figure 2 shows that both BC-TE and FWS-TE consistently outperform other optimal policies especially when large δ is considered, demonstrating the practical effectiveness of TE as an alternative to the forced exploration steps. Furthermore, we observe that BC-TE is more computationally efficient than other asymptotically optimal policies, and FWS-TE outperforms the original FWS in terms of efficiency, as demonstrated in Table 2.

Table 1: Sample complexity over 3,000 independent runs, where outperforming policies are highlighted in boldface using one-sided Welch's t-test with the significance level 0.05. LB denotes the lower bound in (2), and PLB denotes the practical version of LB considered in Degenne et al. (2019). $\boldsymbol{\mu}_5^{\mathrm{B}}$ denotes 5-armed Bernoulli bandit instance with means $(0.3, 0.21, 0.2, 0.19, 0.18)$ and $\boldsymbol{\mu}_4^{\mathrm{G}}$ denotes 4-armed Gaussian bandit instance with means $(1.0, 0.85, 0.8, 0.7)$ and unit variance.

$\boldsymbol{\mu}$	δ	BC-TE	FWS-TE	FWS	LMA	T-D	O-C	AHBR	T3C	RR	PLB	LB
$\boldsymbol{\mu}_5^{\mathrm{B}}$	0.2	**1065**	1077	1176	1415	1107	1545	1615	1115	1977	1208	272
	0.1	**1288**	1326	1373	1668	1337	1818	1859	1372	2326	1442	574
	0.01	**2064**	2102	2125	2509	**2066**	2706	2675	2180	3460	2211	1471
	0.001	**2849**	2870	2880	3362	**2823**	3584	3469	3011	4555	2974	2252
$\boldsymbol{\mu}_4^{\mathrm{G}}$	0.2	**1415**	**1435**	1499	1799	1472	1837	1959	1482	2555	1683	374
	0.1	**1759**	**1772**	1829	2153	**1806**	2235	2339	1833	3078	2004	791
	0.01	**2895**	**2887**	**2890**	3300	**2835**	3501	3524	2947	4730	3062	2026
	0.001	3987	**3967**	**3922**	4445	**3908**	4732	4657	4042	6349	4112	3101



(a) Bernoulli instance $\boldsymbol{\mu}_5^{\mathrm{B}}$ (b) Gaussian instance $\boldsymbol{\mu}_4^{\mathrm{G}}$

Figure 2: Stopping times of various policies for $\delta = 0.1$ over 3,000 independent runs. The black star denotes the mean of stopping times. LB denotes the lower bound given in (2), and PLB denotes the practical version of LB considered in Degenne et al. (2019).

6. Conclusion

In this paper, we introduced BC-TE, a computationally efficient approach for solving the BAI problem in SPEF bandits. By combining a gradient-based policy with Thompson sampling, BC-TE overcame the limitations of existing approaches that involve computationally expensive optimization problems, forced exploration steps, or hyperparameter tuning. Through theoretical analysis and experimental evaluation, we demonstrated that TS can serve as a substitute for the forced exploration steps in BAI problems. Although BC-TE is not universally optimal in general, we showed its optimality for the two-armed bandits setting and provided a comparison with β-optimality. Simulation results further validated the effectiveness of BC-TE, showing competitive sample complexity and improved computational efficiency compared to other optimal policies.

Table 2: Relative average time of one step of various policies.

$\boldsymbol{\mu}$	BC-TE	FWS-TE	FWS	LMA	T-D	O-C	AHBR	T3C	RR
$\boldsymbol{\mu}_5^{\mathrm{B}}$	1	35.53	40.13	1.743	43.52	448.1	2.695	0.8415	0.3246
$\boldsymbol{\mu}_4^{\mathrm{G}}$	1	80.77	96.30	3.588	582.3	4533	3.935	0.7111	0.4226

Acknowledgments

JL was supported by JST SPRING, Grant Number JPMJSP2108. JH was supported by JSPS, KAKENHI Grant Number JP21K11747, Japan. MS was supported by the Institute for AI and Beyond, UTokyo.

References

Shipra Agrawal and Navin Goyal. Analysis of Thompson sampling for the multi-armed bandit problem. In *Annual Conference on Learning Theory*. PMLR, 2012.

Antoine Barrier, Aurélien Garivier, and Tomáš Kocák. A non-asymptotic approach to best-arm identification for gaussian bandits. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 2022.

Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. Pure exploration in multi-armed bandits problems. In *International Conference on Algorithmic Learning Theory*. Springer, 2009.

Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. Pure exploration in finitely-armed and continuous-armed bandits. *Theoretical Computer Science*, 2011.

Olivier Cappé, Aurélien Garivier, Odalric-Ambrym Maillard, Rémi Munos, and Gilles Stoltz. Kullback-leibler upper confidence bounds for optimal sequential allocation. *The Annals of Statistics*, 2013.

Shouyuan Chen, Tian Lin, Irwin King, Michael R Lyu, and Wei Chen. Combinatorial pure exploration of multi-armed bandits. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2014.

Rémy Degenne, Wouter M Koolen, and Pierre Ménard. Non-asymptotic pure exploration by solving games. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2019.

Eyal Even-Dar, Shie Mannor, and Yishay Mansour. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of Machine Learning Research*, 2006.

Victor Gabillon, Mohammad Ghavamzadeh, and Alessandro Lazaric. Best arm identification: A unified approach to fixed budget and fixed confidence. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2012.

Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *Annual Conference on Learning Theory*. PMLR, 2016.

Malay Ghosh. Objective priors: An introduction for frequentists. *Statistical Science*, 2011.

Junya Honda and Akimichi Takemura. Optimality of Thompson sampling for Gaussian bandits depends on priors. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 2014.

Yassir Jedra and Alexandre Proutiere. Optimal best-arm identification in linear bandits. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2020.

Marc Jourdan and Rémy Degenne. Non-asymptotic analysis of a UCB-based top two algorithm. *arXiv preprint arXiv:2210.05431*, 2022.

Marc Jourdan, Rémy Degenne, Dorian Baudry, Rianne de Heide, and Emilie Kaufmann. Top two algorithms revisited. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2022.

Marc Jourdan, Degenne Rémy, and Kaufmann Emilie. Dealing with unknown variances in best-arm identification. In *International Conference on Algorithmic Learning Theory*. PMLR, 2023.

Shivaram Kalyanakrishnan, Ambuj Tewari, Peter Auer, and Peter Stone. PAC subset selection in stochastic multi-armed bandits. In *International Conference on Machine Learning*, 2012.

Emilie Kaufmann and Wouter M Koolen. Mixture martingales revisited with applications to sequential tests and confidence intervals. *The Journal of Machine Learning Research*, 2021.

Emilie Kaufmann, Nathaniel Korda, and Rémi Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In *International Conference on Algorithmic Learning Theory*. Springer, 2012.

Junpei Komiyama, Taira Tsuchiya, and Junya Honda. Minimax optimal algorithms for fixed-budget best arm identification. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2022.

Wouter M Koolen. tidnabbil: Julia library for bandit experiments. https://bitbucket.org/wmkoolen/tidnabbil/src/master/, 2019.

Nathaniel Korda, Emilie Kaufmann, and Remi Munos. Thompson sampling for 1-dimensional exponential family bandits. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2013.

Yuko Kuroki, Liyuan Xu, Atsushi Miyauchi, Junya Honda, and Masashi Sugiyama. Polynomial-time algorithms for multiple-arm identification with full-bandit feedback. *Neural Computation*, 2020.

Jongyeong Lee, Junya Honda, Chao-Kai Chiang, and Masashi Sugiyama. Optimality of Thompson sampling with noninformative priors for Pareto bandits. In *International Conference on Machine Learning*, 2023.

Oden Maron and Andrew W Moore. The racing algorithm: Model selection for lazy learners. *Artificial Intelligence Review*, 1997.

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

Arpan Mukherjee and Ali Tajer. SPRT-based best arm identification in stochastic bandits. In *International Symposium on Information Theory*. IEEE, 2022.

Chao Qin, Diego Klabjan, and Daniel Russo. Improving the expected improvement algorithm. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017.

Charles Riou and Junya Honda. Bandit algorithms based on Thompson sampling for bounded reward distributions. In *International Conference on Algorithmic Learning Theory*. PMLR, 2020.

Christian P Robert, Nicolas Chopin, and Judith Rousseau. Rejoinder: Harold Jeffreys's theory of probability revisited. *Statistical Science*, 2009.

Daniel Russo. Simple Bayesian algorithms for best arm identification. In *Annual Conference on Learning Theory*. PMLR, 2016.

Xuedong Shang, Rianne Heide, Pierre Menard, Emilie Kaufmann, and Michal Valko. Fixed-confidence guarantees for bayesian best-arm identification. In *International Conference on Artificial Intelligence and Statistics*, 2020.

Aleksandrs Slivkins et al. Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning*, 2019.

Po-An Wang, Ruo-Chun Tzeng, and Alexandre Proutiere. Fast pure exploration via frank-wolfe. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2021.