Boundary Focused Thompson Sampling

Pieter Libin
Vrije Universiteit Brussel
Brussels, Belgium
pieter.libin@vub.ac.be

Timothy Verstraeten
Vrije Universiteit Brussel
Brussels, Belgium
timothy.verstraeten@vub.ac.be

Diederik M. Roijers
Vrije Universiteit Amsterdam
Amsterdam, the Netherlands
d.m.roijers@vu.nl

Wenjia Wang
Vrije Universiteit Brussel
Brussels, Belgium
Wenjia.Wang@vub.ac.be

Kristof Theys
Katholieke Universiteit Leuven
Leuven, Belgium
kristof.theys@kuleuven.be

Ann Nowé
Vrije Universiteit Brussel
Brussels, Belgium
ann.nowe@kuleuven.be

ABSTRACT

We introduce Boundary Focused Thompson sampling (BFTS), a new Bayesian algorithm to solve the anytime m-top exploration problem, where the objective is to identify the m best arms in a multi-armed bandit. We consider a set of existing benchmark problems and introduce two new environments inspired by real world decision problems (i.e., scaled Gaussian and Poisson reward distributions), for which we experimentally demonstrate that BFTS consistently outperforms AT-LUCB, the current state of the art algorithm. Finally, we show that BFTS' exploration scheme is well grounded through a formal Bayesian analysis.

KEYWORDS

Thompson sampling, probability matching, m-top exploration, multi-armed bandits, anytime decision making

1 INTRODUCTION

The *multi-armed bandit game* [2] concerns a bandit with K stochastic arms (i.e., a slot machine with K levers). When an arm a_k is pulled, a reward r_k is drawn from that arm's reward distribution R_k. In this work, our aim is to solve the m-top exploration problem ($m < K$), where the objective is to identify the m best arms, with respect to the expected reward $\mathbb{E}[r_k]$ of the arms [3]. Most commonly, the m-top exploration problem is studied in a fixed confidence or fixed budget setting. On the one hand, fixed confidence algorithms attempt to recommend the m best arms with probability $1-\delta$ using a minimal number of arm pulls, where δ is a failure probability that needs to be chosen up front [9, 10, 13, 17, 18]. On the other hand, the goal for fixed budget algorithms is to recommend the top m arms, within a given budget of arm pulls [2, 6, 10, 11, 18]. Recently, a third setting was introduced, where the top m arms are to be recommended after every time step [16]. This setting, referred to as anytime explore-m, is more challenging than the fixed confidence and fixed budget setting, but offers a more realistic framework [16].

An example presented in [16] is a crowd-sourcing task, i.e., the New Yorker cartoon caption contest [14]. In this application, the aim is to collect ratings for the captions submitted for each week's cartoon, and to identify the top-m captions at a requested time. In a crowd sourcing application, the sampling budget corresponds to the number of ratings that are obtained. Therefore, as this budget is unknown a priori, the fixed-budget setting cannot be used. Moreover, the fixed-confidence setting neither is applicable, as this setting requires that an unlimited stream of samples is available until a certain confidence threshold has been reached. The crowd sourcing application is thus a natural fit for the anytime explore-m problem [16].

Apart from this example, we believe that there is a great potential for the anytime m-top exploration bandit to support decision makers with complex societal challenges such as climate issues, epidemics of infectious diseases and migration. Such decisions are often guided by intricate models, to evaluate a variety of policies, that can be represented as bandit arms [19]. Given this formulation, a learning agent can select the m policies for which it expects the highest utility, enabling the experts to inspect this small set of alternatives. The anytime component provides the decision makers with flexibility when a decision can be made, which is especially important when computationally intensive models are used.

Next to introducing the m-top exploration problem, two new algorithms are introducted in [16]: Doubling Successive Accepts and Rejects (DSAR) and AnyTime Lower and Upper Confidence Bound (AT-LUCB). These algorithms remain the state of the art up until today. On the one hand, DSAR repeatedly invokes the fixed budget m-top algorithm Successive Accept and Reject (SAR) [6] where the budget is doubled upon each invocation[1]. Each invocation thus corresponds to a stage s, for which SAR is allocated a budget $2^{s-1}K$. At each time step t, DSAR returns the empirical m-top arms considering the samples that were collected over all stages. AT-LUCB, on the other hand, repeatedly invokes the fixed-confidence LUCB algorithm [17], with a decaying failure parameter, $\delta_s = \delta_1 \alpha^{s-1}$, for each LUCB stage s, where δ_1 and α are parameters of the AT-LUCB algorithm [16]. Details on AT-LUCB's exploration bound are presented in Section 2. At each time step t, AT-LUCB returns the empirical m-top arms.

In [16], it is experimentally shown that AT-LUCB consistently outperforms DSAR, and DSAR is deemed unsuitable for anytime purposes due to fluctuations in its performance (i.e., stagnation or even decrease) when the algorithm changes from one stage to the next. We will therefore omit the DSAR algorithm from our experiments.

While UCB algorithms, such as AT-LUCB, permit specifying tight theoretical bounds, algorithms based on Thompson Sampling (TS) typically perform better in practice [8]. Furthermore, TS works for any type of reward distribution, and permits the inclusion of any form of prior knowledge. This is important, as prior knowledge can be specified for many practical settings, even if it is only in

[1] This is the so called doubling trick.

the form of basic common knowledge or even intuitions, and can greatly help to improve sample-efficiency. Therefore, we investigate the potential of TS for the m-top exploration problem, and propose the first TS-based algorithm for this setting: Boundary Focused Thompson Sampling (BFTS). BFTS is a non-parametric algorithm that focuses its exploration on the problem's decision boundary, i.e., the m^{th} and $m+1^{\text{th}}$ arm. To show that this exploration strategy is well grounded, we perform a formal Bayesian analysis in Section 6.

Furthermore, we demonstrate empirically that our algorithm outperforms the benchmark settings introduced in [16]. As the aforementioned benchmark mainly consists out of a set of artificial environments we add two new environments to evaluate BFTS' performance with respect to decision problems, as discussed earlier in this section. In the first environment the objective is to select the most promising prevention strategies in the context of pandemic influenza (i.e., scaled Gaussians reward distributions). In the second environment we aim to find strategies that maximize the occurrence of an event (i.e., Poisson reward distributions), e.g., to maximize the prevalence of certain insect species on farmland [26].

Through our additional experiments and Bayesian analysis, we demonstrate that the strict reliance of AT-LUCB on sub-Gaussian reward distributions [16], can be relaxed.

2 AT-LUCB'S EXPLORATION BOUND

To provide more insight in AT-LUCB's exploration strategy, we present details on AT-LUCB's exploration bound. Note that this bound was constructed following the assumption that reward distributions are sub-Gaussian with means in $[0, 1]$ [16].

At each stage, LUCB depends on upper confidence bound U_a^t and lower confidence bound L_a^t, where:

$$\begin{aligned} U_a^t(\delta_s) &= \hat{\mu}_a^t + \beta(n_a^t, t, \delta_s) \\ L_a^t(\delta_s) &= \hat{\mu}_a^t - \beta(n_a^t, t, \delta_s), \end{aligned} \tag{1}$$

with,

$$\beta(n_a^t, t, \delta_s) := \sqrt{\frac{1}{2n_a^t} \ln\left(\frac{5}{4}\frac{K \cdot t^4}{\delta_s}\right)}, \tag{2}$$

where $\hat{\mu}_a^t$ is the empirical mean for arm a at time t, K is the number of arms, n_a^t is the amount of times arm a was pulled at time t and δ_s is the confidence parameter at stage s.

From this bound definition, it is clear that the empirical mean is the only reward distribution statistic used by AT-LUCB. We expect that such a bound will be sub-optimal with respect to reward distributions with complex higher-order statistics, such as skewness or high variance.

3 RELATED WORK

The anytime explore-m setting is a generalization of the anytime best-arm identification setting [5].

As we stated in Section 1, the anytime explore-m setting was only recently introduced, and to our best knowledge, the AT-LUCB algorithm remains the state of the art algorithm [16]. We do however note that Bayesian exploration methods have been used in the context of best-arm identification, i.a., BayesGap, Top-Two Thompson sampling, Ordered statistic Thompson sampling, and the Top-Two Expected Improvement algorithm. BayesGap is a gap-based Bayesian algorithm [11] and requires that for each arm, a high-probability upper and lower bound is defined on the posterior of the arms' means at each time step t. These bounds are used to establish a gap quantity that the algorithm attempts to minimize. Top-Two Thompson sampling [24] uses a variant of TS that adds a re-sampling step in order to increase exploration. Ordered statistic Thompson sampling [21] ranks the samples from TS and pulls any arm randomly according to a rank distribution to add extra exploration. The Top-Two Expected Improvement algorithm enhances the Expected Improvement algorithm, by randomizing which of the two top arms to sample [22].

4 BOUNDARY FOCUSED TS

We now describe our anytime m-top algorithm Boundary Focused Thompson sampling (BFTS). The purpose of the algorithm is to recommend the top m arms, anytime. In other words, the algorithm would perform perfectly if it recommends the true top m arms at each time step.

Consider a stochastic multi-armed bandit for which our prior belief over the means is given by $\pi(.)$. Inspired by TS, at each time step t we sample an estimate $\boldsymbol{\theta}^{(t)}$ for the means $\mu_{1..K}$ from $\pi(\cdot|\mathcal{H}_{t-1})$, i.e., the posterior over the means, given by $\pi(.)$ conditional on the history of arm pulls and observed rewards $\mathcal{H}_{t-1}$. Consequently, we order the samples that comprise $\boldsymbol{\theta}^{(t)}$, and define $\sigma_i(\boldsymbol{\theta}^{(t)})$ to be the i-th ordered arm. In the case of vanilla TS [27], where the objective is to minimize cumulative regret, we would always play top arm $\sigma_1(\boldsymbol{\theta}^{(t)})$. However, for the anytime m-top bandit problem, where the objective is to return the top m arms at any time[2], the intuition is to focus the exploration on the decision boundary, more specifically to decrease the uncertainty about arm $a_m^{(t)}$ and $a_{m+1}^{(t)}$. We focus on *both sides* of the decision boundary as in a pure exploration setting, it is equally important to gain information about the arms with the potential to be optimal and sub-optimal.

In order to implement this intuition using TS, at each time step t we play the arm ordered $\sigma_m(\boldsymbol{\theta}^{(t)})$ or $\sigma_{m+1}(\boldsymbol{\theta}^{(t)})$ with equal probability. To do this, we use a Bernoulli experiment, as formalized in Algorithm 1. The reward $r^{(t)}$ of the played arm $a^{(t)}$ is observed and used to update the history $\mathcal{H}^{(t-1)}$. At the end of each step, we recommend the m-top arms based on the current belief over the bandit posterior $\pi(\cdot \mid \mathcal{H}_{t-1})$.

Given: $\pi(.)$ and $\mathcal{H}^{(0)} = \emptyset$
for $t = 1, \ldots, +\infty$ **do**
  $\boldsymbol{\theta}^{(t)} \sim \pi(\cdot \mid \mathcal{H}_{t-1})$
  $b \sim \mathcal{B}\text{er}(0.5)$
  $a^{(t)} = \sigma_{m+b}(\boldsymbol{\theta}^{(t)})$
  $r^{(t)} \leftarrow$ Pull arm $a^{(t)}$
  $\mathcal{H}^{(t)} \leftarrow \mathcal{H}^{(t-1)} \cup \{a^{(t)}, r^{(t)}\}$
  Recommend top arms based on $\pi(\cdot \mid \mathcal{H}_{t-1})$
end

Algorithm 1: Boundary Focused TS

[2]The m top arms should be recommended, they are not expected to be ranked correctly among them.

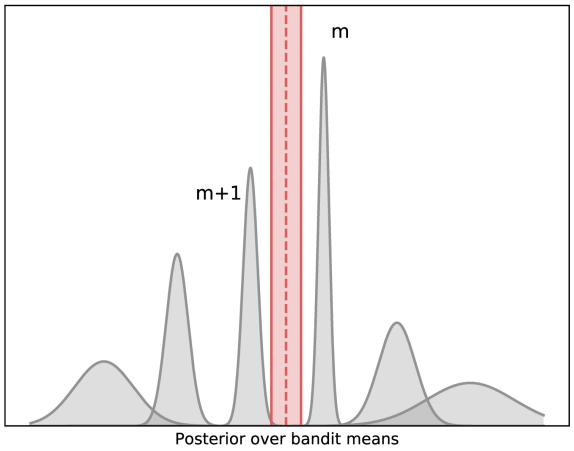

Figure 1: Posteriors for an artificial bandit ($K = 6, m = 3$) (gray) and BFTS' decision boundary with confidence bounds to demonstrate its uncertainty (red).

An important observation with respect to BFTS is that the exploration is guided by sampling from the posterior, while balancing between $\sigma_m(\boldsymbol{\theta}^{(t)})$ and $\sigma_{m+1}(\boldsymbol{\theta}^{(t)})$, i.e., our belief of the decision boundary at time t. As the posterior reflects the uncertainty with respect to the bandit problem, sampling the m^{th} or $m+1^{\text{th}}$ ordered arm will initially explore all the arms, when an uninformative prior is chosen. However, as the uncertainty of the outer extreme arms is reduced, BFTS will increase its focus on the arms near the decision boundary. In Figure 1, we visualize this process for a simple bandit setting ($K = 6$ and $m = 3$) with Gaussian posteriors.

In Section 6, we show via a formal analysis that this exploration strategy is well grounded.

5 EXPERIMENTS

We compare the performance of BFTS to the state of the art algorithm, i.e., AT-LUCB, and uniform sampling as a baseline. AT-LUCB operates as described in Section 1 and Section 2, and we choose the same parameters as in [16]. Uniform sampling pulls at each time step t the arm that was least sampled in the previous time steps, and recommends the empirical m-top arms.

For BFTS, we recommend the m-top arms with the highest posterior expectation $\mathbb{E}[\pi(\cdot \mid \mathcal{H}_{t-1})]_k$. The use of the posterior expectation is well grounded in our experiments, as all priors we use will tend to a bell-shaped posterior, for which the expectation is a natural summary statistic.

In order to perform a fair and unbiased evaluation we commence with the experimental environments introduced in [16]. We continue by introducing two more practical environments, useful to support decision makers with complex societal challenges, as introduced in Section 1. Both new settings consider interesting reward distributions: the first setting has scaled Gaussian reward distributions for which the variances are not known and the second setting has Poisson reward distributions.

AT-LUCB expects sub-Gaussian reward distributions with means in $[0, 1]$. We will demonstrate experimentally, using a bandit environment with Poisson reward distributions, that AT-LUCB indeed performs poorly when this assumption is not met.

The probability of error, i.e., the probability that all of the true best arms are recommended, does not yield a useful comparison in our experiments, as the considered environments are hard, and it takes a large amount of samples to find the true m top arms [16]. Therefore, we evaluate the algorithms' performance using two proxy statistics instead: the sum of the means of the m top arms at time t, as introduced in [16],

$$\sum_{a \in J^{(t)}} \mu_a, \tag{3}$$

and the proportion of correctly recommended arms at time t,

$$\frac{|J^{(t)} \cap J^{\text{True}}|}{m}, \tag{4}$$

where $J^{(t)}$ is the set of recommended arms at time t and J^{True} is the true set of optimal arms.

All of the algorithms were run 100 times for each of the stochastic bandit environments, as such, the average of the statistics over these runs is reported. In order to justify this number of replicates, all figures include the variance of the reported statistic, which is visualized using a lighter bound around the mean curve. In every run, each algorithm was allowed to consume 14×10^4 samples (i.e., arm pulls), a sufficient amount to discern a clear learning curve. Note that for BFTS and uniform sampling only one sample per time step is used, while for AT-LUCB two samples per time step are used. Therefore, all figures report their results in terms of the number of samples.

For all BFTS experiments, we consistently use Jeffreys' priors. Such priors are considered non-informative and objective, such that when data is observed, the posteriors will not be influenced by the prior's hyper-parameters [15].

5.1 Gaussian bandit with fixed variance

The first set of benchmark environments introduced in [16] concerns Gaussian reward distributions with fixed variance $\sigma^2 = 0.25$ and means in $[0, 1]$. The environment defines a bandit with 1000 arms. The benchmark includes two instances, one where the gap between means is increased linearly (Equation 5) and one where the gap is increased polynomially (Equation 6).

$$\forall k : \mu_k = .9\left(\frac{n-i}{n-1}\right) \tag{5}$$

$$\mu_1 = .9, \forall k \geq 2 : \mu_k = .9(1 - \sqrt{i/n}) \tag{6}$$

In this environment, as each arm a_k has a reward distribution $\mathcal{N}(\mu, \sigma^2)$ with known variance, we have a conjugate prior for the mean that is Gaussian with hyper-parameters μ_0 and σ_0^2. As the means are in $[0, 1]$, we choose this Gaussian prior to be truncated on said interval. We consider $\sigma_0^2 \to \infty$, which results in a uniform prior $\mathcal{U}(0, 1)$ over μ (full derivation in Supplementary Information [3], SI). This uniform prior corresponds to the Jeffreys prior [23].

[3] Supplementary Information available at:
https://drive.google.com/drive/folders/181CAIzCJQO5EU97irfW0zsO2yCHVFQBV

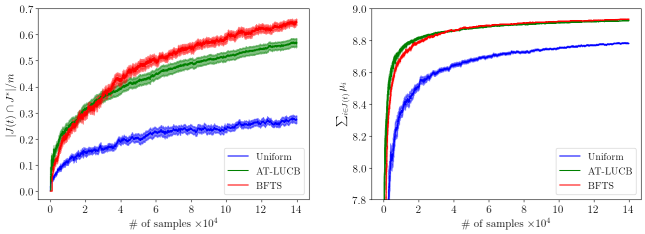

Figure 2: Results for the linear Gaussian benchmark with fixed variance ($m = 10$).

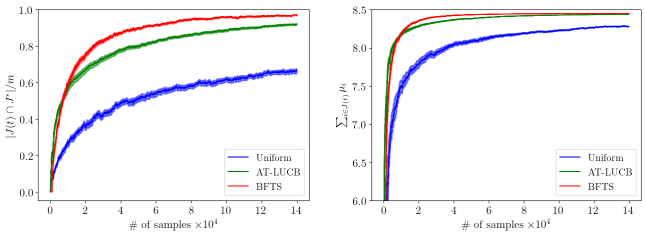

Figure 3: Results for the polynomial Gaussian benchmark with fixed variance ($m = 10$).

Given rewards $\mathbf{r} = \{r_1, ..., r_n\}$ we have posterior (full derivation in SI):

$$\mu \sim \mathcal{N}_{[0,1]}\Big(\mu_0 = \frac{\sum_{i=1}^{n} r_i}{n}, \sigma_0^2 = \frac{\sigma^2}{n}\Big). \tag{7}$$

The expectation of the posterior over μ, that is required for recommending the m top arms, is simply the mean of the truncated Gaussian in Equation 7.

As in [16], we perform the experiment with $m = 10$ and $m = 50$, for both the linear and polynomial environment. We present the results for the linear bandit in Figure 2 ($m = 10$) and Figure 4 ($m = 50$). We present the results for the polynomial bandit in Figure 3 ($m = 10$) and Figure 5 ($m = 50$). In general, BFTS needs short burn-in period to meet AT-LUCB's performance for both statistics, but then consistently outperforms AT-LUCB, most apparently with respect to the proportion of success' learning curve. On the one hand, for the linear Gaussian environment with $m = 10$, it takes BFTS the most time to meet AT-LUCB's performance. On the other hand, for the linear bandit with $m = 50$, BFTS takes the least iterations to meet the performance of AT-LUCB.

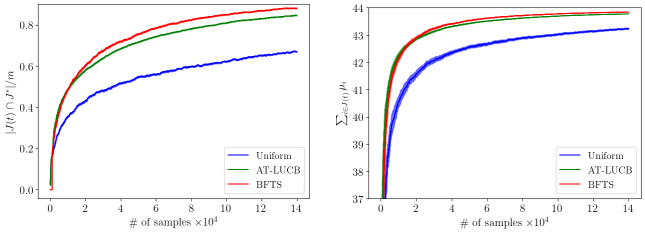

Figure 4: Results for the linear Gaussian benchmark with fixed variance ($m = 50$).

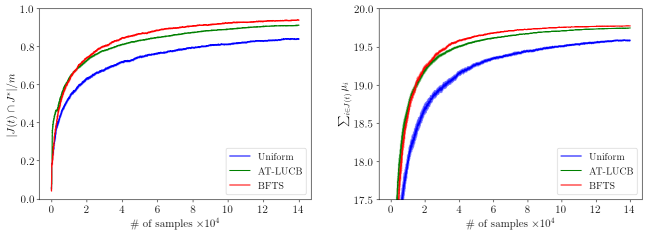

Figure 6: Results for the cartoon caption benchmark

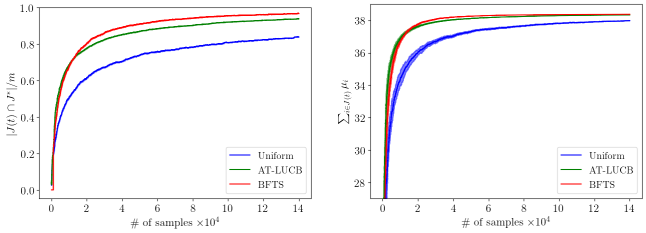

Figure 5: Results for the polynomial Gaussian benchmark with fixed variance ($m = 50$).

5.2 Cartoon caption bandit

The second benchmark environment introduced in [16] concerns the New York cartoon caption contest we described in Section 1. This benchmark simulates the caption contest by setting up a bandit with 496 arms, where each arm follows a categorical distribution $\mathcal{C}\text{at}_{\mathbf{c}}(\mathbf{p})$ on three categories $\mathbf{c} = [0, 0.5, 1]$. The distribution is parametrized with an event probability vector $\mathbf{p}$. For each arm, $\mathbf{p}$ is determined using maximum likelihood estimation, on the dataset used in [16].

For a categorical distribution $\mathcal{C}\text{at}_{\mathbf{c}}(\mathbf{p})$, the conjugate prior is a Dirichlet distribution $\mathcal{D}\text{ir}_{\mathbf{c}}(\boldsymbol{\alpha})$ with prior parameter $\boldsymbol{\alpha}$. Given rewards $\mathbf{r} = \{r_1, ..., r_n\}$, we have posterior

$$\mu \sim \mathbf{c} \cdot \mathcal{D}\text{ir}_{\mathbf{c}}(\boldsymbol{\alpha} + \mathbf{f}) \tag{8}$$

where $\mathbf{f}$ is a vector of frequencies at which the categories occur in $\mathbf{r}$. Note that this is a proper posterior if all elements in $\boldsymbol{\alpha}$ are greater than zero. For the experiment we use an uninformative Jeffreys prior $\boldsymbol{\alpha} = [.5, .5, .5]$ [28]. The expression to compute the expectation of the posterior over μ is presented in SI.

As in [16], we run the caption contest bandit experiment for $m = 50$. We present the results for this experiment in Figure 6. BFTS needs a short burn-in to meet AT-LUCB's performance for both statistics, but then consistently outperforms AT-LUCB, most significantly with respect to the proportion of success' learning curve.

5.3 Scaled Gaussian bandit

In Section 1, we addressed that there is a great potential for the anytime m-top exploration bandits, to support decision makers with complex societal challenges. Consequently, we introduce a third environment that concerns the evaluation of preventive strategies with the objective to curb epidemics (e.g., school closures, vaccine allocation) [19]. This setting is modeled as a bandit where each arm

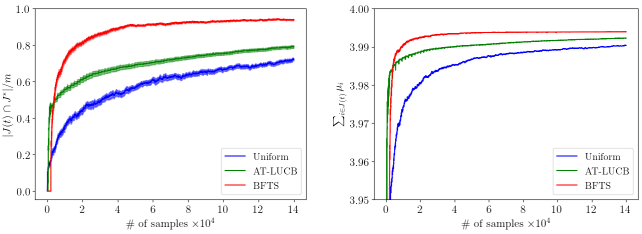

Figure 7: Results for the scaled Gaussian benchmark

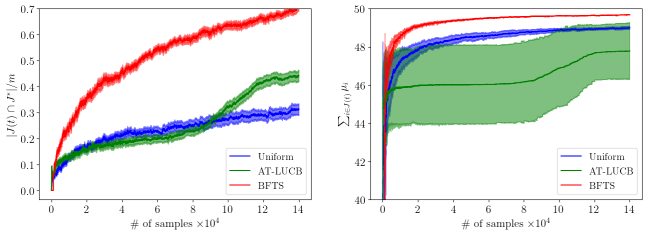

Figure 8: Results for the Poisson benchmark

corresponds to a prevention strategy, and when the arm is pulled, this policy is evaluated using a stochastic simulator. When the arm is pulled, the outcome of the simulator is returned as reward. This implies that the arm's reward distribution corresponds to the outcome distribution of the policy that is associated with that arm.

More specifically, we consider a simulator that has a Gaussian outcome distribution, as is the case for influenza pandemics [19], for which the means are distributed in a sub-interval of $[0, 1]$. The range of this sub-interval is unknown, as it depends on the simulated scenario. Additionally, for such simulations, the variance of the outcome distribution differs per prevention strategy and is unknown [19]. Inspired by one particular experiment in [19], we set up an environment where the means and variances are uniformly sampled from respectively $U(0.3, 0.4)$ and $U(1.06 \cdot 10^{-6}, 3.6 \cdot 10^{-2})$. The environment defines a bandit with 1000 arms.

As each arm has a Gaussian reward distribution with unknown variance, we assume an uninformative Jeffreys prior $(\sigma)^{-3}$ on (μ, σ^2) [12]. Given rewards $\mathbf{r} = \{r_1, ..., r_n\}$, this prior leads to the non-standardized t-distributed posterior. We truncate this prior on $[0, 1]$, given that we know that the arms' means are in this interval:

$$\mu \sim \mathcal{T}_{n,[0,1]}\left(\mu_0 = \frac{\sum_{i=1}^{n} r_i}{n}, \sigma_0^2 = \frac{\sum_{i=1}^{n}(r_i - \mu_0)^2}{n^2}\right). \tag{9}$$

This posterior needs to be initialized two times for it to be proper. The expectation of the posterior over μ is derived in the SI.

We present the results for the scaled Gaussian bandit for $m = 10$, in Figure 7. The BFTS algorithm needs two rounds of initialization per arm, but after this initialization phase, its performance surpasses AT-LUCB quickly.

5.4 Poisson bandit

Finally, we present an environment with Poisson distributed rewards [7], with means:

$$\mu = \mu_{\min} + \frac{k \cdot (\mu_{\max} - \mu_{\min})}{K - 1}, \tag{10}$$

for $\mu_{\min} = 0.5$ and $\mu_{\max} = 5$. The environment defines a bandit with 1000 arms. This is a particularly hard problem, as for Poisson distributions the variance is equal to the mean, and subsequently there is a large variance among the top arms, complicating the m-top exploration. As in the previous environment, this setting is also of interest to decision makers, i.e., for stochastic simulators with an outcome that represents the occurrence of an event. This setting has, for example, the potential to investigate strategies to maximize the prevalence of certain insect species on farmland [26].

For a Poisson distribution, the conjugate Jeffreys prior is a gamma distribution: $\mathcal{G}\text{amma}(\alpha = 0.5, \beta = 0)$ [20]. Given rewards $\mathbf{r} = \{r_1, ..., r_n\}$, this leads to posterior

$$\mu \sim \mathcal{G}\text{amma}(\alpha + \sum_{i=1}^{n} r_i, \beta + n). \tag{11}$$

As $\beta = 0$, this posterior needs to be initialized one time for it be proper.

We present the results for the Poisson bandit for $m = 10$, in Figure 8. It is clear that AT-LUCB's performance grows very slowly, while BFTS exhibits a much steeper learning curve. We discuss the reasons for AT-LUCB's poor performance for this setting in Section 7.

5.5 Conclusion

In our experiments, BFTS consistently outperforms AT-LUCB , for all reported statistics. We do identify that BFTS needs an initialization period to meet AT-LUCB's performance, but we do not deem the lower performance during the first iterations of the algorithm problematic, as at these times both algorithms perform poorly, and a fair amount of exploration is required to improve this.

Interestingly, BFTS also exhibits a significant performance improvement compared to AT-LUCB for the new settings we introduce. These additional experiments show that AT-LUCB struggles in the Poisson environment, while BFTS performs much better. This demonstrates that BFTS has a great potential to be used with reward distributions that are not sub-Gaussian and non-symmetric. This is an important result, as we are unaware of any algorithms able to solve such problems efficiently.

6 BAYESIAN ANALYSIS OF BFTS

To provide more insight in BFTS' sampling strategy we perform a formal Bayesian analysis. In this Bayesian framework, we reason about the full distribution over bandits. Consequently, the actual means $\boldsymbol{\mu}$ are unknown, and we assert our belief over $\boldsymbol{\mu}$ given

$$\pi(\cdot|\mathcal{H}^{(t-1)}), \tag{12}$$

i.e., the prior belief over the means $\pi(.)$ conditional on the observed history

$$\mathcal{H}^{(t-1)} = \left\{a^{(i)}, r^{(i)}\right\}_{i=1}^{t-1} \tag{13}$$

at time t.

We define the random variables A_o^{π} as the o-ranked arms according to the prior belief, and A_o^{TS} as the o-ranked arm according to

TS:

$$A_o^{\pi} = \sigma_o(\boldsymbol{\mu})$$
$$A_o^{TS} = \sigma_o(\boldsymbol{\theta}^{(t)}) \tag{14}$$

TS is a *probability matching* algorithm [1, 25] and therefore it samples directly from the belief asserted in Equation 12. Formally, this is defined as:

$$P(A_o^{TS} = \cdot \mid \mathcal{H}^{(t-1)}) = P(A_o^{\pi} = \cdot \mid \mathcal{H}^{(t-1)}) \tag{15}$$

We define $o^+ \in [1, \ldots, m]$ and $o^- \in [m+1, \ldots, K]$. Using this notation, we can express the true optimal arm set J^* and recommended arm set $J^{(t)}$ as:

$$J^* = \{A_{o^+}^{\pi} \mid \forall o^+\}$$
$$J^{TS} = \{A_{o^+}^{TS} \mid \forall o^+\}, \tag{16}$$

we will refer to $\overline{J^*}$ as the complement of J^*, i.e., the set of all arms excluding J^*. Note that, both J^* and J^{TS} are random variables, as they are expressed as a union of random variables.

Given this framework, we identify two heuristics that underlie BFTS' sampling strategy.

For the remainder of this Section, we will use $P_t(.)$ to denote a probability that is conditional on the observed history $\mathcal{H}^{(t-1)}$ at time t:

$$P_t(.) = P(.|\mathcal{H}^{(t-1)}) \tag{17}$$

Heuristic 1. *The expectation that BFTS wrongly ranks an arm that is believed to be optimal is bounded by the probability that BFTS wrongly ranks the arm on the sub-optimal side of the decision boundary:*

$$\mathop{\mathbb{E}}_{o^-}[P_t(A_{o^-}^{TS} \in J^*)] \leq P_t(A_{m+1}^{TS} \in J^*) \tag{18}$$

Given this inequality, we expect that sampling the $m+1$-th arm will reduce $\mathbb{E}_{o^-}[P_t(A_{o^-}^{TS} \in J^*)]$.

Heuristic 2. *The expectation that BFTS wrongly ranks an arm that is believed to be sub-optimal is bounded by the probability that BFTS wrongly ranks the arm ranked on the optimal side of the decision boundary.*

$$\mathop{\mathbb{E}}_{o^+}[P_t(A_{o^+}^{TS} \in \overline{J^*})] \leq P_t(A_m^{TS} \in \overline{J^*}) \tag{19}$$

Given this inequality, we expect that sampling the m-th arm will reduce $\mathbb{E}_{o^+}[P_t(A_{o^+}^{TS} \in \overline{J^*})]$.

These heuristics stem from the fact that it is counter-intuitive for TS to order an arm often as optimal when it is *believed* to be sub-optimal. However, due to the stochastic nature of both the bandit and TS, it is possible to end up with a posterior for which the heuristics do not hold. Notwithstanding, we argue that given the intuition behind probability matching, such events become unlikely when reasonable priors are chosen and BFTS' exploration strategy is followed. We show this empirically in Section 6.1 for a diverse set of environments and their corresponding posteriors.

We will now show how the expectations in the heuristics relate to the probability of error. As such, given the heuristics, we can bound the probability of error with respect to both sides of the decision boundary (i.e., A_{m+1}^{TS} and A_m^{TS}), demonstrating that BFTS' exploration strategy is well grounded.

First, we derive the bound in terms of A_{m+1}^{TS}:

$$\begin{aligned}
&P_t\left(J^* \neq J^{TS}\right) \\
&= P_t\left(\bigvee_{o^-} A_{o^-}^{TS} \in J^*\right) \\
&\leq \sum_{o^-} P_t\left(A_{o^-}^{TS} \in J^*\right) \\
&= \frac{\sum_{o^-} P_t\left(A_{o^-}^{TS} \in J^*\right) \cdot (K-m)}{(K-m)} \\
&= \mathop{\mathbb{E}}_{o^-}\left[P_t\left(A_{o^-}^{TS} \in J^*\right)\right] \cdot (K-m) \\
&\overset{(H1)}{\leq} P_t\left(A_{m+1}^{TS} \in J^*\right) \cdot (K-m)
\end{aligned} \tag{20}$$

In the first step, we express the probability of error in terms of the arms that are ranked as sub-optimal by TS. In the second step, we apply a union bound. In the third and fourth step, we transform the sum to an expected value. In the final step, we apply Heuristic 1 ($H1$).

Second, we derive the bound in terms of A_m^{TS}:

$$\begin{aligned}
&P_t\left(J^* \neq J^{TS}\right) \\
&= P_t\left(\bigvee_{o^+} A_{o^+}^{TS} \in \overline{J^*}\right) \\
&\leq \sum_{o^+} P_t\left(A_{o^+}^{TS} \in \overline{J^*}\right) \\
&= \frac{\sum_{o^+} P_t\left(A_{o^+}^{TS} \in \overline{J^*}\right) \cdot m}{m} \\
&= \mathop{\mathbb{E}}_{o^+}\left[P_t\left(A_{o^+}^{TS} \in \overline{J^*}\right)\right] \cdot m \\
&\overset{(H2)}{\leq} P_t\left(A_m^{TS} \in \overline{J^*}\right) \cdot m
\end{aligned} \tag{21}$$

In the first step, we express the probability of error in terms of the arms that are ranked as optimal by TS. In the second step, we apply a union bound. In the third and fourth step, we transform the sum to an expected value. In the final step, we apply Heuristic 2 ($H2$).

These insights motivate a uniform selection of the decision boundary, as is reflected in BFTS (see Algorithm 1, line 2 and 3 in the for loop).

The BFTS algorithm is constructed such that its sampling strategy is completely independent of its recommendation strategy. Likewise, in this analysis, we consider the belief that BFTS maintains over the problem, in terms of the random variable J^{TS} (Equation 16), rather than the statistic that is used to make recommendations (e.g., the mean of the posterior in our experiments). This observation shows that our analysis is independent from the statistic used to make recommendations with BFTS.

When inspecting other algorithms for the m top setting, we observe that the decision boundary between the m^{th} and $m+1^{\text{th}}$ arms also play an important role. For example, the frequentist algorithm AT-LUCB samples two arms each step; the one with the smallest lower-bound among the top m arms, and the one with the greatest upper-bound among the rest. This is analogous to choosing

the optimal and sub-optimal arms that are closest to the decision boundary.

6.1 Empirical validation of the heuristics

To experimentally validate Heuristic 1 and Heuristic 2 we will express the inequalities in terms of sums over all arms.

For Heuristic 1, we have on the left hand side:

$$P_t(A^{TS}_{o^-} \in J^*) \overset{(I)}{=} \sum_a^K P_t(A^{TS}_{o^-} = a) P_t(a \in J^*) \overset{(Eq.\ 15)}{=} \sum_a^K P_t(A^{\pi}_{o^-} = a) P_t(a \in J^*), \tag{22}$$

and on the right hand side:

$$P_t(A^{TS}_{m+1} \in J^*) \overset{(I)}{=} \sum_a^K P_t(A^{TS}_{m+1} = a) P_t(a \in J^*) \overset{(Eq.\ 15)}{=} \sum_a^K P_t(A^{\pi}_{m+1} = a) P_t(a \in J^*), \tag{23}$$

where we rely on the independence of TS rankings with respect to the optimal set (I) and probability matching (*Eq.* 15).

For Heuristic 2, following analogous arguments, we have on the left hand side:

$$P_t(A^{TS}_{o^+} \in \overline{J^*}) = \sum_a^K P_t(A^{\pi}_{o^+} = a) P_t(a \in \overline{J^*}), \tag{24}$$

and on the right hand side:

$$P_t(A^{TS}_{m} \in \overline{J^*}) = \sum_a^K P_t(A^{\pi}_{m} = a) P_t(a \in \overline{J^*}) \tag{25}$$

All ranking probabilities in Equations 22 to 25 are in terms of the belief over the means. Therefore, these ranking probabilities can be estimated at each step of BFTS by obtaining a set of samples from the posterior. Each element in this set represents a sample from our belief over the means, and by ordering the entries in this sample, we obtain a ranking. By doing this over all samples in the set, we obtain a frequency distribution over rankings.

Given that both heuristics consider an expected value in terms of K on the left hand side and that the equations above denote a sum over all arms, estimating the heuristics experimentally has a computational complexity that is quadratic in K. Therefore, in our experiments, we use bandit environments with $K = 100$ and $m = 5$, instead of the 1000-armed bandit environments that were covered in Section 5. We evaluate the heuristics for the same environments as in Section 5. For each environment, we run 100 BFTS replicates for $3 \cdot 10^4$ time steps[4]. While running BFTS, we compute the probabilities that make up the heuristics at every 100-th time step. Consequently, for each heuristic a total of $3 \cdot 10^4$ samples were collected per environment.

Our experiments show that for the two Gaussian environments with fixed variance (i.e., linear and polynomial) both heuristics hold for all measurements. For the scaled Gaussian environment, Heuristic 1 holds for all measurements. For Heuristic 2, we recorded 9 failures (= 0.03%), of which most (8 out of 9) occurred during the initial time steps ($t \leq 2000$) of BFTS. For the Poisson environment, Heuristic 1 holds for all measurements. For Heuristic 2, we recorded 46 failures (< 0.154%), of which most (40 out of 46) occurred during the initial time steps ($t \leq 2000$) of BFTS. For the Categorical environment, Heuristic 1 holds for all measurements. For Heuristic 2, we recorded 139 failures (< 0.467%), of which most (111 out of 139) occurred during the initial time steps ($t \leq 2000$) of BFTS.

[4]We present the results, with respect to proportion of success and sum of the means, for these environments in SI.

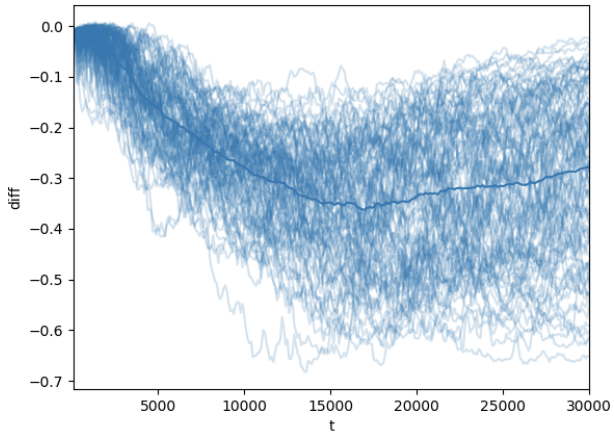

Figure 9: Trajectories of probability differences for Heuristic 2 in the Categorical environment (traces: light blue lines, mean: thick blue line).

These results justify our choice of heuristics, as failures are rare, especially after the initial phase of the algorithm. The observation that failures correlate strongly with the earlier time step of BFTS can be explained by the fact that the posteriors have not yet converged to a symmetric bell shape at this time. We would indeed expect to see less failures when the posteriors are bell-shaped, and this intuition is supported by the experiments with Gaussian posteriors.

Furthermore, as Heuristic 2 is to be interpreted as a zero-bounded difference in probabilities:

$$\mathop{\mathbb{E}}_{o^+}[P_t(A^{TS}_{o^+} \in \overline{J^*})] - P_t(A^{TS}_{m} \in \overline{J^*}) \leq 0, \tag{26}$$

we note that all failures are caused by differences that are close to zero. Furthermore, we observe that the general trend is well below the zero-bound, as shown in Figure 9 for the Categorical environment. We show plots for the other environments in SI.

7 DISCUSSION

BFTS is a Bayesian algorithm, which means that prior knowledge with respect to the problem can be incorporated. This is important, as for many real world problems such information is available, e.g., the cartoon caption contest [14], challenging decision problems [19] and settings with correlated arms [11].

As expected from its assumptions imposed on the reward distribution, AT-LUCB performs poorly in non sub-Gaussian settings, as we experimentally confirm in Section 5. This can be explained by the symmetric bound used by AT-LUCB (see Section 2), which will make bandit problems with a highly skewed reward distribution (e.g., Poisson), hard to solve. From our Bayesian analysis, it is clear

that BFTS is not bound by such restrictions, and only relies on two heuristics, for which we argue that they are sensible in the context of probability matching when reasonable priors are chosen. This is an interesting observation, as this moves the assumption away from the reward distribution, which is inherently problem specific, to the posterior distribution, that represents the belief over the bandit's arms' means. While reward distributions are static, posteriors evolve when rewards are observed, and due to the central limit theorem, we expect that any specified prior will eventually tend towards a Gaussian [4]. This is important, as we expect both of the heuristics to hold well for bell-shaped posteriors in general, which was empirically supported through our experiments in Section 6.

While we performed a Bayesian analysis that provides important insights in BFTS' sampling strategy, a tight bound on the probability of error still needs to be established. We want to assert that, to our best knowledge, no such proofs have been established with respect to TS in the context of pure exploration. Furthermore, even for vanilla TS, it took almost 80 years to come up with a tight bound on cumulative regret [1, 27].

For future work, we acknowledge that additional efforts on theoretical guarantees are warranted, and we believe the heuristics proposed in Section 6 could provide an interesting starting point, when additional constraints are imposed. Additionally, for decision making settings, it is important that the uncertainty of a decision can be quantified. In the context of TS, the state of the posteriors represents the actual belief over the bandit problem, and can therefore be used to compute such statistics [19].

8 CONCLUSION

In this manuscript, we introduce BFTS, a new algorithm for the anytime explore-m problem. We show that BFTS' exploration strategy is well grounded using a formal Bayesian analysis. We empirically show that BFTS consistently outperforms the current state of the art algorithm AT-LUCB, in a variety of experimental settings, i.e., Gaussian with fixed and unknown variance, Categorical and Poisson reward distributions.

ACKNOWLEDGEMENTS

Pieter Libin and Timothy Verstraeten were supported by a PhD grant of the FWO (Fonds Wetenschappelijk Onderzoek - Vlaanderen). Kristof Theys was supported by a postdoctoral grant of the FWO. The computational resources were provided by an EWI-FWO grant (Theys, KAN2012 1.5.249.12.).

REFERENCES

[1] Shipra Agrawal and Navin Goyal. 2012. Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on Learning Theory*. 39–1.
[2] Jean-Yves Audibert and Sébastien Bubeck. 2010. Best arm identification in multi-armed bandits. In *COLT-23th Conference on Learning Theory*.
[3] Robert E Bechhofer. 1958. A sequential multiple-decision procedure for selecting the best one of several normal populations with a common unknown variance, and its use with various experimental designs. *Biometrics* 14, 3 (1958), 408–429.
[4] Patrick Billingsley. 2008. *Probability and measure*. John Wiley and Sons.
[5] Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. 2009. Pure exploration in multi-armed bandits problems. In *International conference on Algorithmic learning theory*. Springer, 23–37.
[6] Sébastian Bubeck, Tengyao Wang, and Nitin Viswanathan. 2013. Multiple identifications in multi-armed bandits. In *International Conference on Machine Learning*. 258–265.
[7] Olivier Cappé, Aurélien Garivier, Odalric-Ambrym Maillard, Rémi Munos, Gilles Stoltz, et al. 2013. Kullback–leibler upper confidence bounds for optimal sequential allocation. *The Annals of Statistics* 41, 3 (2013), 1516–1541.
[8] Olivier Chapelle and Lihong Li. 2011. An empirical evaluation of thompson sampling. In *Advances in neural information processing systems*. 2249–2257.
[9] Eyal Even-Dar, Shie Mannor, and Yishay Mansour. 2002. PAC bounds for multi-armed bandit and Markov decision processes. In *International Conference on Computational Learning Theory*. Springer, 255–270.
[10] Victor Gabillon, Mohammad Ghavamzadeh, and Alessandro Lazaric. 2012. Best arm identification: A unified approach to fixed budget and fixed confidence. In *Advances in Neural Information Processing Systems*. 3212–3220.
[11] Matthew Hoffman, Bobak Shahriari, and Nando Freitas. 2014. On correlation and budget constraints in model-based bandit optimization with application to automatic machine learning. In *Artificial Intelligence and Statistics*. 365–374.
[12] Junya Honda and Akimichi Takemura. 2014. Optimality of Thompson Sampling for Gaussian Bandits Depends on Priors.. In *AISTATS*. 375–383.
[13] Kevin Jamieson, Matthew Malloy, Robert Nowak, and Sébastien Bubeck. 2014. lil'ucb: An optimal exploration algorithm for multi-armed bandits. In *Conference on Learning Theory*. 423–439.
[14] Kevin G Jamieson, Lalit Jain, Chris Fernandez, Nicholas J Glattard, and Rob Nowak. 2015. Next: A system for real-world development, evaluation, and application of active learning. In *Advances in Neural Information Processing Systems*. 2656–2664.
[15] Edwin T Jaynes. 1968. Prior probabilities. *IEEE Transactions on systems science and cybernetics* 4, 3 (1968), 227–241.
[16] Kwang-Sung Jun and Robert D Nowak. 2016. Anytime Exploration for Multi-armed Bandits using Confidence Information.. In *33rd International Conference on Machine Learning*. 974–982.
[17] Shivaram Kalyanakrishnan, Ambuj Tewari, Peter Auer, and Peter Stone. 2012. PAC Subset Selection in Stochastic Multi-armed Bandits.. In *ICML*, Vol. 12. 655–662.
[18] Zohar Karnin, Tomer Koren, and Oren Somekh. 2013. Almost optimal exploration in multi-armed bandits. In *International Conference on Machine Learning*. 1238–1246.
[19] Pieter JK Libin, Timothy Verstraeten, Diederik M Roijers, Jelena Grujic, Kristof Theys, Philippe Lemey, and Ann Nowé. 2018. Bayesian Best-Arm Identification for Selecting Influenza Mitigation Strategies. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 456–471.
[20] David Lunn, Chris Jackson, Nicky Best, David Spiegelhalter, and Andrew Thomas. 2012. *The BUGS book: A practical introduction to Bayesian analysis*. Chapman and Hall/CRC.
[21] Joseph Charles Mellor. 2014. *Decision Making Using Thompson Sampling*. Ph.D. Dissertation. University of Manchester.
[22] Chao Qin, Diego Klabjan, and Daniel Russo. 2017. Improving the Expected Improvement Algorithm. In *Advances in Neural Information Processing Systems*. 5381–5391.
[23] Christian Robert. 2007. *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer Science and Business Media.
[24] Daniel Russo. 2016. Simple bayesian algorithms for best arm identification. In *Conference on Learning Theory*. 1417–1418.
[25] Daniel Russo and Benjamin Van Roy. 2016. An information-theoretic analysis of Thompson sampling. *The Journal of Machine Learning Research* 17, 1 (2016), 2442–2471.
[26] Richard L Soulsby and Jeremy A Thomas. 2012. Insect population curves: modelling and application to butterfly transect data. *Methods in Ecology and Evolution* 3, 5 (2012), 832–841.
[27] William R Thompson. 1933. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika* 25, 3/4 (1933), 285–294.
[28] Frank Tuyl. 2017. A note on priors for the multinomial model. *The American Statistician* 71, 4 (2017), 298–301.