Bayesian Anytime m-top Exploration

Libin, Pieter; Verstraeten, Timothy; Roijers, Diederik M; Wang, Wenjia; Theys, Kristof; Nowe, Ann

Published in:
2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)

DOI:
10.1109/ICTAI.2019.00201

Publication date:
2019

Document Version:
Accepted author manuscript

Bayesian Anytime m-top Exploration

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

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

Diederik M. Roijers
Vrije Universiteit Brussel
Computer science department
Brussels, Belgium
diederik.roijers@vub.ac.be

Wenjia Wang
Vrije Universiteit Brussel
Computer science department
Brussels, Belgium
wenjia.wang@vub.ac.be

Kristof Theys
Katholieke Universiteit Leuven
Rega Institute for Medical Research
Leuven, Belgium
kristof.theys@kuleuven.be

Ann Nowé
Vrije Universiteit Brussel
Computer science department
Brussels, Belgium
ann.nowe@vub.ac.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.**

First, we consider a set of existing benchmark problems that consider sub-Gaussian reward distributions (i.e., Gaussian with fixed variance and categorical reward). Next, we introduce a new environment inspired by a real world decision problem concerning insect control for organic agriculture. This new environment encodes a Poisson rewards distribution. For all these benchmarks, we experimentally show that BFTS consistently outperforms AT-LUCB, the current state of the art algorithm.

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

I. Introduction

The *multi-armed bandit game* [2] concerns a bandit with K stochastic arms (e.g., 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 $\mathcal{R}_k$. For each arm a_k, we have the expected reward $\mu_k = \mathbb{E}[r_k]$. 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 μ_k of the arms [3], [22]. Formally, we have $\mu_1 \geq \ldots \geq \mu_m \geq \mu_{m+1} \geq \ldots \geq \mu_K$, and the objective is to identify the set $\{\mu_1, \ldots, \mu_m\}$.

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 [8], [9], [11], [15], [16]. 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], [4], [9], [10], [16], [17]. Recently, a third setting was introduced, where the top m arms are to be recommended after every time step [14]. 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 [14].

An example of an m-top exploration problem presented in [14] is a crowd-sourcing task, i.e., the New Yorker cartoon caption contest [12]. 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 is not applicable either, 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.

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 simulation models, to evaluate a set of alternative policies that can be modelled as bandit arms [18], [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 to when a decision can be made. This is especially important when computationally intensive models are used, for which it is difficult to make a trade-off between the available budget and desired confidence.

Next to introducing the m-top exploration problem, a new algorithm is presented in [14]: AnyTime Lower and Upper Confidence Bound (AT-LUCB). This algorithm remains the state-of-the-art up until today. We discuss the algorithmic details of AT-LUCB in Section II.

While UCB algorithms, such as AT-LUCB, permit specifying tight theoretical bounds, algorithms based on Thompson Sampling (TS) typically perform better in practice [7]. 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 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 Bayesian 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.

We empirically compare the performance of BFTS to AT-LUCB. First, we evaluate the set of benchmarks settings introduced in [14], which consists out of an artificial environment (i.e., a bandit with fixed-variance Gaussian reward distributions) and a bandit that models the New York cartoon crowd sourcing task introduced earlier. Second, in order to evaluate BFTS' performance with respect to decision problems, we introduce a new benchmark environment motivated by a real-world decision problem, i.e., the **organic bandit**, where we aim to maximize the prevalence of certain insect species on farmland to support organic agriculture [26]. As this problem corresponds to maximizing the occurrence of an event, we model this setting using Poisson reward distributions. 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. We show that BFTS consistently outperforms AT-LUCB for all of the investigated environments, and show a vast improvement in performance on the organic bandit.

II. Background: AT-LUCB

AT-LUCB repeatedly invokes the fixed-confidence LUCB algorithm [15], 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. At each time step t, AT-LUCB returns the empirical m-top arms.

To provide more insight in AT-LUCB's exploration strategy, we discuss details on AT-LUCB's exploration bound [14]. Note that this bound was constructed following the assumption that reward distributions are sub-Gaussian with means in the interval $[0, 1]$.

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 confidence bound definition, it is clear that the empirical mean is the only reward distribution statistic used by AT-LUCB. We expect that such a confidence bound will be sub-optimal with respect to reward distributions with complex higher-order statistics, such as skewness or high variance. We demonstrate that this is the case in our experiments with the organic bandit, with Poisson distributed rewards, in Section V.

III. Related work

The anytime explore-m setting is a generalization of the anytime best-arm identification setting [5]. As introduced earlier, this setting is related to the fixed confidence [8], [9], [11], [15], [16] and fixed budget [2], [4], [9], [10], [16] explore-m algorithms.

As we stated in Section I, 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. In [14], another algorithm called DSAR is presented next to AT-LUCB. DSAR repeatedly invokes the fixed budget m-top algorithm Successive Accept and Reject (SAR) [4] where the budget is doubled upon each invocation. It is experimentally shown in [14] 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 therefore chose to omit the DSAR algorithm from our experiments.

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 [10] 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 [25] 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 [23].

IV. Boundary Focused TS

In this section, we propose our anytime m-top algorithm Boundary Focused Thompson sampling (BFTS). The purpose of the algorithm is to recommend the top m arms at each time step.

Consider a stochastic multi-armed bandit for which our prior belief over the means is given by a distribution $\pi(.)$. Inspired by TS, at each time step t we sample an estimate $\boldsymbol{\theta}^{(t)}$ for the means $\mu_{1..K}$ from $\pi(\cdot \mid \mathcal{H}^{(t-1)})$, i.e., the posterior over the means, given by $\pi(.)$ conditioned 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 $\Psi_\rho(\boldsymbol{\theta}^{(t)})$ to be the ρ ordered arm. In the case of vanilla TS [27], where the objective is to minimize cumulative regret, we would always play top arm $\Psi_1(\boldsymbol{\theta}^{(t)})$. However, for the anytime m-top bandit problem, where the objective is to return the top m arms at any time[1], we need to focus the exploration on the decision boundary. 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.

To implement the intuition of focussing on the decision boundary, at each time step t we play the arm ordered $\Psi_m(\boldsymbol{\theta}^{(t)})$ or $\Psi_{m+1}(\boldsymbol{\theta}^{(t)})$ 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**
    $\boldsymbol{\theta}^{(t)} \sim \pi(\cdot \mid \mathcal{H}^{(t-1)})$
    $b \sim \mathcal{B}\text{er}(0.5)$
    $a^{(t)} = \Psi_{m+b}(\boldsymbol{\theta}^{(t)})$
    $r^{(t)} \leftarrow$ Pull arm $a^{(t)}$
    $\mathcal{H}^{(t)} \leftarrow \mathcal{H}^{(t-1)} \cup \{a^{(t)}, r^{(t)}\}$
    Recommend top arms based on $\pi(\cdot \mid \mathcal{H}^{(t)})$
**end**
```

Algorithm 1: Boundary Focused TS

An important observation with respect to BFTS is that the exploration is guided by sampling from the posterior, while balancing between $\Psi_m(\boldsymbol{\theta}^{(t)})$ and $\Psi_{m+1}(\boldsymbol{\theta}^{(t)})$, 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 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.

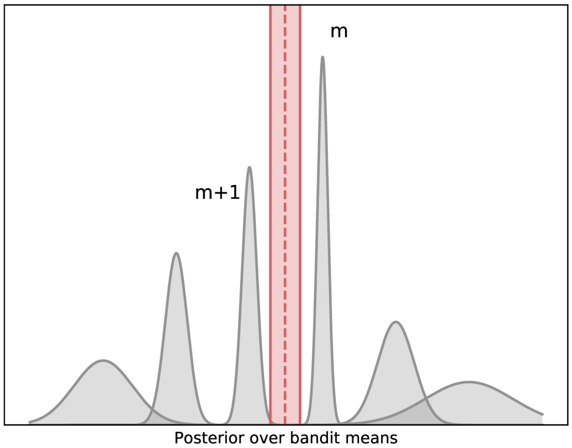

Figure 1: Posteriors for an artificial bandit ($K = 6, m = 3$) (gray) and BFTS' decision boundary with confidence bounds to demonstrate its uncertainty (red).

BFTS is thus convenient for real-world applications, as its belief-based exploration can be intuitively understood and informed by its users, without the need to specify any exploration parameters that are typically hard to choose in advance. Moreover, the anytime aspect of BFTS removes the need to decide on the computational budget or desired confidence before starting the analysis.

V. Experiments

We compare the performance of BFTS to the current state-of-the-art algorithm, i.e., AT-LUCB, and uniform sampling as a baseline. AT-LUCB operates as described in Section II, and we choose the same parameters as in [14]. 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. The use of the posterior expectation is well-grounded in our experiments, as all priors we use tend to a bell-shaped posterior, for which the expectation is a natural summary statistic.

To perform a fair and unbiased evaluation we commence with the experimental environments introduced in [14]. Then, we introduce a new environment, i.e., the organic bandit. This new settings considers a Poisson reward distribution.

AT-LUCB expects sub-Gaussian reward distributions with means in the interval $[0, 1]$. We demonstrate experimentally, using the organic 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

[1] The top m arms should be recommended, but they are not expected to be ranked.

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 [14]. 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 [14],

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

and the proportion of correctly recommended arms at time t,

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

where $J^{(t)}$ is the set of recommended arms at time t and J^* 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 obtained, while for AT-LUCB two samples per time step are used. Therefore, all figures report their results in terms of the number of samples, to allow for a fair comparison.

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 are not influenced by the prior's hyper-parameters [13].

A. Gaussian bandit with fixed variance

The first set of benchmark environments introduced in [14] concerns Gaussian reward distributions with fixed variance $\sigma^2 = 0.25$ and means in the interval $[0, 1]$. 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).

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

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

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 a uniform prior over μ. This uniform prior corresponds to the Jeffreys prior [24]. Given rewards $\mathbf{r} = \{r_1, ..., r_n\}$ we have posterior:

$$\mu \sim \mathcal{N}(\hat{\mu}_0, \hat{\sigma}_0^2), \tag{7}$$

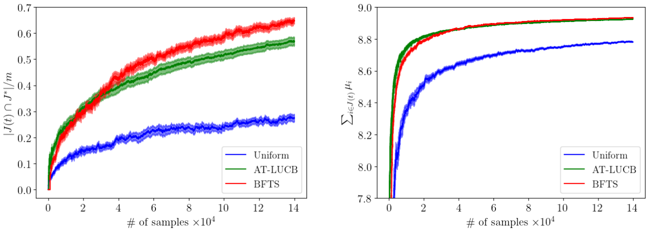

Figure 2: Results for the linear Gaussian benchmark with fixed variance ($m = 10$).

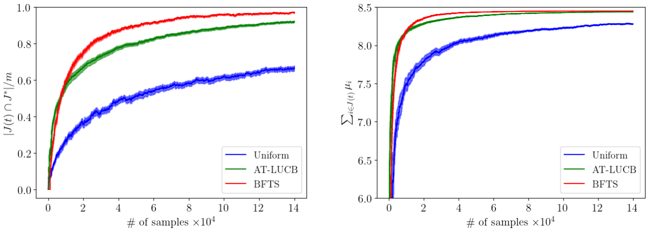

Figure 3: Results for the polynomial Gaussian benchmark with fixed variance ($m = 10$).

with,

$$\begin{aligned} \hat{\sigma}_0^2 &= \lim_{\sigma_0 \to +\infty} \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2} \right)^{-1} = \frac{\sigma^2}{n} \\ \hat{\mu}_0 &= \lim_{\sigma_0 \to +\infty} \frac{\sigma^2}{n} \left(\frac{\mu_0}{\sigma_0^2} + \frac{\sum_{i=1}^n r_i}{\sigma^2} \right) = \frac{\sum_{i=1}^n r_i}{n} \end{aligned} \tag{8}$$

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

As in [14], 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 a 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.

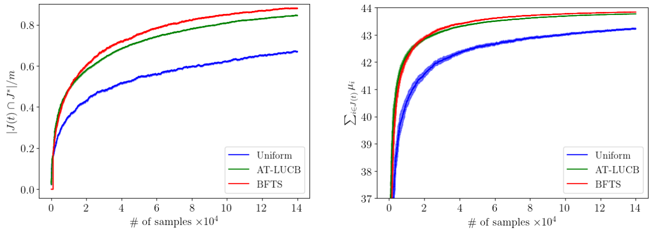

Figure 4: Results for the linear Gaussian benchmark with fixed variance ($m = 50$).

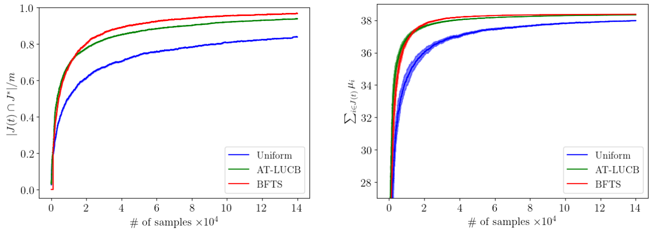

Figure 5: Results for the polynomial Gaussian benchmark with fixed variance ($m = 50$).

B. *Cartoon caption bandit*

The second benchmark environment introduced in [14] concerns the New York cartoon caption contest we described in Section I. 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 [14].

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

$$\mu \sim \mathbf{c} \cdot \mathcal{D}\text{ir}_{\mathbf{c}}(\boldsymbol{\alpha} + \mathbf{f}) \tag{9}$$

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:

$$\mathbb{E}[\mu] = \frac{\sum_{i=1}^{|\mathbf{c}|} \mathbf{c}_i (\boldsymbol{\alpha} + \mathbf{f})_i}{\sum_{i=1}^{|\mathbf{c}|} (\boldsymbol{\alpha} + \mathbf{f})_i} \tag{10}$$

As in [14], 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.

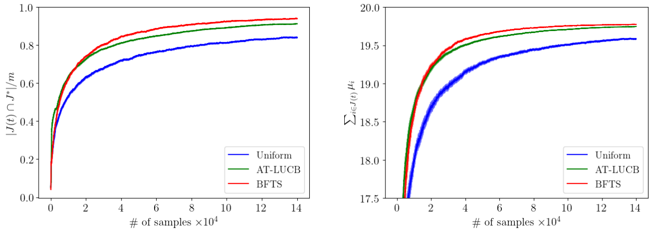

Figure 6: Results for the cartoon caption benchmark

C. *Organic bandit*

Finally, we present a new benchmark environment motivated by a research question that stems from organic agriculture, i.e., to investigate strategies that maximize the prevalence of certain insect species on farmland [26].

As we are attempting to maximize the occurrence of an event [26], we construct a benchmark environment with Poisson distributed reward distributions [6], with linearly increasing means:

$$\mu_k = \mu_{\text{min}} + \frac{k \cdot (\mu_{\text{max}} - \mu_{\text{min}})}{K - 1}, \tag{11}$$

for $\mu_{\text{min}} = 0.5$ and $\mu_{\text{max}} = 5$. The environment defines a bandit with 1000 arms.

As mentioned in the Section I, this is a particularly challenging benchmark, as for a Poisson distribution, the variance equals the mean, which complicates the m-top exploration process.

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

$$\mu \sim \mathcal{G}\text{amma}(\alpha + \sum_{i=1}^{n} r_i, \beta + n). \tag{12}$$

As $\beta = 0$, this posterior needs to be initialized one time for it be proper. The expression to compute the expectation of the posterior over μ is:

$$\mathbb{E}[\mu] = \frac{\alpha + \sum_{i=1}^{n} r_i}{\beta + n}. \tag{13}$$

We present the results for the organic bandit for $m = 10$, in Figure 7. It is clear that AT-LUCB's performance grows very slowly and is similar to random sampling, while BFTS exhibits a much steeper learning curve. We further discuss these results in Section VI.

D. *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.

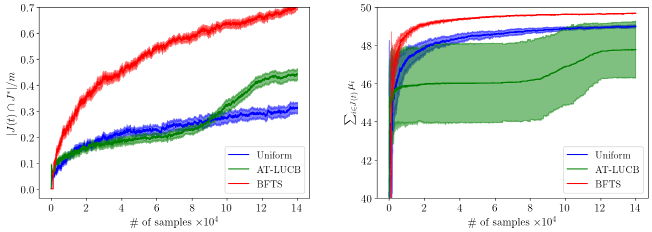

Figure 7: Results for the organic bandit benchmark

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 with the organic bandit, 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.

BFTS outperforms AT-LUCB for both of the reported statistics. For the sum of means, a proxy for the simple regret, the difference is most evident during the earlier time steps of the experiments, as the difference in performance becomes less clear when the sum of means for both BFTS and AT-LUCB converge to a similar value. Supported by this observation, we argue that the number of correctly recommend arms (i.e., proportion of success) is a better proxy for the probability of error, i.e., the quantity that we actually attempt to optimize. Given this statistic, it is immediately clear how many mistakes an algorithm makes at a certain time, and BFTS' superior performance is even more evident.

VI. Discussion

BFTS is a Bayesian algorithm, which means that prior knowledge with respect to the problem can be easily incorporated. This is important, as for many real world problems such information is available, e.g., the cartoon caption contest [12], important societal decision problems [19] and settings with correlated arms [10].

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 V. This can be explained by the symmetric bound used by AT-LUCB (see Section II), which will make bandit problems with a highly skewed reward distribution (e.g., Poisson) hard to solve.

While our experimental results are promising, a bound on the probability of error still needs to be established. For future work, we acknowledge that efforts on theoretical guarantees are warranted. 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. Even for vanilla TS, it took almost 80 years to come up with a tight bound on cumulative regret [1], [27].

VII. Conclusion

In this manuscript, we introduce BFTS, a new algorithm for the anytime explore-m problem. We empirically show that BFTS consistently outperforms the current state-of-the-art algorithm AT-LUCB, in a variety of experimental settings, even when uninformative priors are used., i.e., Gaussian with fixed variance, Categorical and Poisson reward distributions.

Reproducibility

All source code used to perform the experiments is freely available on GitHub: https://github.com/plibin-vub/bfts.

All experiments were executed on the high performance computer cluster of the University of Leuven (KU Leuven).

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. Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on Learning Theory*, pages 39–1, 2012.

[2] Jean-Yves Audibert and Sébastien Bubeck. Best arm identification in multi-armed bandits. In *COLT-23th Conference on Learning Theory*, 2010.

[3] Robert E Bechhofer. 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):408–429, 1958.

[4] Sébastian Bubeck, Tengyao Wang, and Nitin Viswanathan. Multiple identifications in multi-armed bandits. In *International Conference on Machine Learning*, pages 258–265, 2013.

[5] Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. Pure exploration in multi-armed bandits problems. In *International conference on Algorithmic learning theory*, pages 23–37. Springer, 2009.

[6] Olivier Cappé, Aurélien Garivier, Odalric-Ambrym Maillard, Rémi Munos, Gilles Stoltz, et al. Kullback–leibler upper confidence bounds for optimal sequential allocation. *The Annals of Statistics*, 41(3):1516–1541, 2013.

[7] Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. In *Advances in neural information processing systems*, pages 2249–2257, 2011.

[8] Eyal Even-Dar, Shie Mannor, and Yishay Mansour. PAC bounds for multi-armed bandit and markov decision processes. In *International Conference on Computational Learning Theory*, pages 255–270. Springer, 2002.

[9] 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*, pages 3212–3220, 2012.

[10] Matthew Hoffman, Bobak Shahriari, and Nando Freitas. On correlation and budget constraints in model-based bandit optimization with application to automatic machine learning. In *Artificial Intelligence and Statistics*, pages 365–374, 2014.

[11] Kevin Jamieson, Matthew Malloy, Robert Nowak, and Sébastien Bubeck. lil'ucb: An optimal exploration algorithm for multi-armed bandits. In *Conference on Learning Theory*, pages 423–439, 2014.

[12] Kevin G Jamieson, Lalit Jain, Chris Fernandez, Nicholas J Glattard, and Rob Nowak. Next: A system for real-world development, evaluation, and application of active learning. In *Advances in Neural Information Processing Systems*, pages 2656–2664, 2015.

[13] Edwin T Jaynes. Prior probabilities. *IEEE Transactions on systems science and cybernetics*, 4(3):227–241, 1968.

[14] Kwang-Sung Jun and Robert D Nowak. Anytime exploration for multi-armed bandits using confidence information. In *33rd International Conference on Machine Learning*, pages 974–982, 2016.

[15] Shivaram Kalyanakrishnan, Ambuj Tewari, Peter Auer, and Peter Stone. PAC subset selection in stochastic multi-armed bandits. In *ICML*, volume 12, pages 655–662, 2012.

[16] Zohar Karnin, Tomer Koren, and Oren Somekh. Almost optimal exploration in multi-armed bandits. In *International Conference on Machine Learning*, pages 1238–1246, 2013.

[17] Emilie Kaufmann and Shivaram Kalyanakrishnan. Information complexity in bandit subset selection. In *Conference on Learning Theory*, pages 228–251, 2013.

[18] Pieter Libin, Timothy Verstraeten, Kristof Theys, Diederik Roijers, Peter Vrancx, and Ann Nowe. Efficient evaluation of influenza mitigation strategies using preventive bandits. *Adaptive Learning Agents workshop, Workshop @ International Conference on Autonomous Agents and Multiagent Systems*, 2017.

[19] Pieter JK Libin, Timothy Verstraeten, Diederik M Roijers, Jelena Grujic, Kristof Theys, Philippe Lemey, and Ann Nowé. Bayesian best-arm identification for selecting influenza mitigation strategies. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 456–471. Springer, 2018.

[20] David Lunn, Chris Jackson, Nicky Best, David Spiegelhalter, and Andrew Thomas. *The BUGS book: A practical introduction to Bayesian analysis*. Chapman and Hall/CRC, 2012.

[21] Joseph Charles Mellor. *Decision Making Using Thompson Sampling*. PhD thesis, University of Manchester, 2014.

[22] Edward Paulson et al. A sequential procedure for selecting the population with the largest mean from k normal populations. *The Annals of Mathematical Statistics*, 35(1):174–180, 1964.

[23] Chao Qin, Diego Klabjan, and Daniel Russo. Improving the expected improvement algorithm. In *Advances in Neural Information Processing Systems*, pages 5381–5391, 2017.

[24] Christian Robert. *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer Science and Business Media, 2007.

[25] Daniel Russo. Simple bayesian algorithms for best arm identification. In *Conference on Learning Theory*, pages 1417–1418, 2016.

[26] Richard L Soulsby and Jeremy A Thomas. Insect population curves: modelling and application to butterfly transect data. *Methods in Ecology and Evolution*, 3(5):832–841, 2012.

[27] William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

[28] Frank Tuyl. A note on priors for the multinomial model. *The American Statistician*, 71(4):298–301, 2017.