# Extended abstract: Thompson sampling for m-top Exploration

Pieter Libin[1,2], Timothy Verstraeten[1,3], Diederik M. Roijers[1], Wenjia Wang[1], Kristof Theys[2], and Ann Nowé[1]

[1] Artificial Intelligence lab, Vrije Universiteit Brussel
[2] KU Leuven - University of Leuven, Rega Institute for Medical Research
[3] Institute of ICT, HU University of Applied Sciences

*This document is an extended abstract of the paper was accepted at the International Conference on Tools with Artificial Intelligence, that will take place in November 2019 in Portland (Oregon, United states).*

The *multi-armed bandit game* 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 $\mu_k$ of the arms. 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 $\delta$ is a failure probability that needs to be chosen up front. On the other hand, the goal for fixed budget algorithms is to recommend the top $m$ arms, within a given budget of arm pulls. Recently, a third setting was introduced, where the top $m$ arms are to be recommended after every time step [2]. 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.

An example of an $m$-top exploration problem presented in [2] is a crowdsourcing task, i.e., the New Yorker cartoon caption contest. 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. 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 [2]: AnyTime Lower and Upper Confidence Bound (AT-LUCB). This algorithm remains the state-of-the-art up until today.

While UCB algorithms, such as AT-LUCB, permit specifying tight theoretical bounds, algorithms based on Thompson Sampling (TS) typically perform better in practice [1]. 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 $m^{\text{th}}$ 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 [2], 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. 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.

## References

1. Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. In *Advances in neural information processing systems*, pages 2249–2257, 2011.
2. 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.