ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Simple Bayesian Algorithm for Top-m Arm Identification

Master Thesis

Kevin Klein

November 16, 2019

Advisors: Johannes Kirschner, Mojmír Mutný, Prof. Dr. Andreas Krause
Department of Computer Science, ETH Zürich

Abstract

This example thesis briefly shows the main features of our thesis style, and how to use it for your purposes.

Contents

Contents **iii**

1 Introduction **1**

2 Background **3**
2.1 Chernoff's 2-player game 3
2.2 Bandits 3
2.3 Notation 4
2.4 Thompson Sampling for Bandits 5
2.5 Best Arm Identification 6
2.5.1 Model 6
2.5.2 Optimal allocation 7
2.5.3 A Constrained Optimal Allocation 8
2.5.4 Top-Two Thompson Sampling algorithm 9
2.5.5 Alternative approaches 10
2.6 Top-m Arm Identification 10
2.6.1 Current approaches: LUCB 10

3 Charactarizing the Constrained Optimal Top-m Allocation **13**
3.1 Statements 14
3.2 A Concrete Optimal Allocation 17
3.3 Proofs 18

4 Top-2m XOR Thompson Sampling **25**
4.1 Algorithm 25
4.2 Analysis 27
4.3 Empirical behaviour 27
4.4 Proofs 27

5 Conclusion **29**

A Appendix **31**
A.1 Computing $C_{j,i}$ for Bernoulli means 31
A.2 Facts about the exponential family 32
A.3 Useful technical statements 32

Bibliography **35**

Chapter 1

Introduction

What is the relationship between the optimal allocation and the algorithm? How does this statement translate to real world applications? What is particular about the algorithm being Bayesian? What are constaints on prior and posterior distributions? Are we in a frequentist or Bayesian setting?

Chapter 2

Background

2.1 Chernoff's 2-player game

Tbd whether relevant.

2.2 Bandits

The so-called stochastic multi-armed bandit is a general model to simulate decision-making in uncertain environments. In particular, one assumes a set of options or arms to choose from, each choice leading to an outcome, also referred to as reward. In general, one assumes many sequential selections among the set of arms. The outcome of the selection of an individual arm typically follows a fixed but unknown probability distribution.

Within the bandits model, the concern revolves around which arm to choose next. Allocation strategies tackling this question address either of two problems: explore-exploit or pure-explore.

The former concerns itself with maximizing the *cummulative reward*. This means that one attempts to maximize the *sum of the rewards obtained* through all arm selections. Naturally, starting off without any knowledge about the underlying distributions, this involves both exploration of arm qualities as well as exploitation of knowledge obtained so far.

The latter revolves around *simple reward*. This problem consists of seeking to maximize the reward one obtains if one leverages the current knowledge for *another* draw. This implies that the focus lies on *identification* of high-quality candidates, quality usually being assessed by high means. It is therefore it is part of the task to align estimation of the best arm with the underlying truth as well as reaching a high *confidence* in this alignment.

The bandits model can be used for all processes involving sequential decision making. Yet, an important simplification is the assumption of the distri-

butions being fixed over time. In how far this simplification is representative of the to-be-modelled process varies.

The explore-exploit problem is applied in Recommender Systems, ad campaigns, in the context of Reinforcement Learning and Robotics.

For real-world applications of the pure-exploration bandit please refer to 2.5.1.

There exist many specialized versions of the Bandits model, such as contextual bandits or adversarial bandits.

2.3 Notation

We assume k arms to choose from and denote the set of possible arms as $[k]$. Subsets $S \subset [k]$ are assumed to be of size $m < k$.

A possible set of means for the arms is denoted as the k-dimensional vector θ, where every mean can lie between 0 and 1. We also refer to such a θ as parameter.

Finding ourselves in a frequentist setting, we assume an underlying true mean of the arms refer to this as θ^*. Moreover, this ground truth implies a true best arm l^* and true top-m arms S^*. When referring to an individual arm in the top-m case, we proceed to use j for arms in S^*, i for arms not in S^* and l for arms of which this knowledge does not exist.

Given the constraint that the means lie between 0 and 1, there are infinitely many possible parameter vectors which make arm $l \in [k]$ the best one under θ or $S \subset [k]$ top-m under θ. Hence we group such θ in the following way:

$$\Theta := [0,1]^k \tag{2.1}$$

$$\Theta_{1,l} := \{\theta \in \Theta | l = \arg\max_{l' \in [k]} \theta_{l'}\} \tag{2.2}$$

$$\Theta_{m,S} := \{\theta \in \Theta | S = \text{top-}m(\theta)\} \tag{2.3}$$

$$= \{\theta \in \Theta | \min_{j_1 \in S} \ \theta_{j_1} > \max_{j_2 \notin S} \ \theta_{j_2}\} \tag{2.4}$$

$$\Theta_{m,l} = \{\theta \in \Theta | l \in \ \text{top-}m(\theta)\} \tag{2.5}$$

where top-m returns the m highest values, i.e. means, from its argument θ.

After having made n observations of rewards, bundled in D_n, Π_n expresses a posterior distribution with density π_n over elements of Θ. E.g. for sets $\Theta_{m,S}$, we have:

$$\Pi_n(\Theta_{m,S}) := \int_{\theta \in \Theta_{m,S}} \pi_n(\theta) d\theta \tag{2.6}$$

$$= \Pr[S \text{ is top-}m | D_n] \tag{2.7}$$

Clearly, it holds that $\Pi_n(\Theta) = 1$. As a shorthand for the top-m case, we will use:

$$\alpha_{n,l} := \Pi_n(\Theta_{m,l}) \tag{2.8}$$

$$\alpha_{n,S} := \Pi_n(\Theta_{m,S}) \tag{2.9}$$

We are mostly interested in strategies that allocate measurement effort over the arms in a randomized fashion, i.e. according to a probability distribution. We refer to such an allocation or distribution as ψ, defining the probability of each arm being sampled. Naturally, it holds that $\sum_{l \in [k]} \psi_l = 1$. We also define the allocation property of a set by the sum of its arms: $\psi_S = \sum_{l \in S} \psi_l$.

The n-th sampled arm is denoted as l_n. For adaptive strategies, the allocation can change after every sample. After n samples we refer to the allocation as ψ_n and therefore we have $\psi_{n,l} = \Pr[l_n = l]$ for arm l. We might as well be interested in the average allocation up to a certain sample n. We write:

$$\bar{\psi}_{n,l} := \frac{\sum_{n'=1}^{n} \psi_{n',l}}{n} \tag{2.10}$$

We use the notation $d(\theta_1 || \theta_2)$ to represent the Kullback-Leibler divergence between θ_1 and θ_2.

We employ Russo's notation $a_n \doteq b_n$ if the the following relationship holds:

$$\lim_{n \to \infty} \frac{1}{n} \log \frac{a_n}{b_n} \to 0 \tag{2.11}$$

Hence $\doteq$ can be vaguely thought of as asymptotic relationship discounting for logarithmic factors. Note the asymmetry of the relationship: it rather represents an inequality than an equality.

2.4 Thompson Sampling for Bandits

Thompson sampling is a sampling strategy often employed for selecting arms in the bandits scenario. By default, it particularly lends itself to the cummulative reward setting. In the following, we will discuss Thompson sampling in the context of bandits.

The foundation of Thompson sampling is a Bayesian approach: instead of 'only' possessing estimates on the means/distributions of arms, it entails a distribution *over* arms, indicating the likelihood of a random variable over the arms. In the bandits scenario said random variable tends to be the mean of an arm. Hence, instead point estimate of the means, Thompson sampling requires have distributions over the means per arm.

Naturally, this distribution over the arms should also leverage the knowledge aquired throughout the sampling process. Hence, per round, we can

talk about *prior* and *posterior* distributions. This posterior distribution of a step n answers the question: 'How likely are the arms $1, 2, 3$ to have means $[.193, .254, .1]$ after having observed samples 1 to n. As this is an iterative process, the posterior after observing sample n will serve as the prior for selecting sample $n + 1$. Hence the posterior can be thought of as the the prior 'updated' with the knowledge acquired of a specific round. Thanks to this duality we will continue to only talk about posteriors.

The attentive reader might wonder how exactly posteriors are updated. The update procedure is dependent of the type of utilized class of distribution for prior and posteriors. An example of this can be found in 4.3.

In contrast to greedily selecting the arm that empirically maximizes the metric of desire, e.g. the mean, Thompson sampling suggests to randomly draw beliefs on arms weighted by the posterior distribution. Note that said posterior distribution is itself a function of the empirical means. Subsequently, it selects the arm which maximizes the metric of desire on the sampled belief. With many repetitions, this leads to each arm being sampled proportionally to its likelihood of maximizing the metric of desire according to current belief. Algorithm 1 illustrates this process more explicitly.

Algorithm 1 Given a posterior Π_n in step $n + 1$

$\hat{\theta} \sim \Pi_n$
$l_{n+1} = \arg\max_{l \in [k]} \hat{\theta}_l$
Play l_{n+1}
Observe reward r_{n+1}
$\Pi_{n+1} = update(\Pi_n, l_{n+1}, r_{n+1})$

Intuitively, this is very appealing for the cummulative reward scenario as it creates a natural balance between exploration and exploitation.

2.5 Best Arm Identification

2.5.1 Model

Best Arm identification implies disregarding the sum of the rewards encountered while sampling. Rather, the goal is to efficiently gather information maximizing the confidence of the suggestion made *after* the sampling phase. In other words, there is a purely explorative phase, also referred to as 'experiment' which is then typically followed by a purely exploitative phase, having committed to an option. For this reason, Best Arm Identification is a kind of 'Experiment Design'.

Quite naturally, all other aspects being controlled for, requiring fewer samples is preferable. Hence one can formulate the goal in two ways: maximize confidence for a given amount of samples or minimize amount of samples for a given confidence level.

Confidence can mean and be evaluated in different ways. For Bayesian algorithms, it is possible to quantify the confidence the model has in the currently most favourable looking candidate. This quantity corresponds to the mass a posterior puts on parameters favouring said candidate. Another approach to measure confidence stems from the realm of Probably approximately correct learning, or PAC in short. In the PAC context, one desires to make a statement of the sort $\Pr[\text{output of algoritm is } \varepsilon\text{-correct}] \geq 1 - \delta$, where ε-correctness requires an explicit definition. In this case, under the toleration of ε, $1 - \delta$ can be thought of as confidence.

Real-world applications of Best Arm identification include:

- Physical simulations: Given a collection of designs, e.g. the bodywork of a car, determine aerodynamic properties of the designs. Elaborate simulations can come with significant resource demands, such as compute power as well as the direct and indirect costs of time. Arriving at the same conclusion of the superiority of a certain designs, with fewer simulations can be extremely valuable.
- Crop selection: Experimenting different crop types in a given growing environment measuring yields, generate a recommendation for that same growing environment.

”ML: drive generation of own data instead of” -¿ it is about acquisition

2.5.2 Optimal allocation

We start off by describing what it means for an allocation to be optimal and by enumerating some of its properties. We do not present a constructive allocation, rather we assume knowledge about the underlying truth to provide a tight bound on the best possible allocation. Intuitively it is not possible to match the performance of that optimal allocation without the knowledge of the underlying truth. Hence a very desirable statement to show that a proposed, constructive adaptive allocation *converges* to the optimal allocation.

As mentioned in 2.5.1, the goal consists of maximizing confidence, i.e. the mass the posterior lays onto the true best arm. Oberve that for any allocation sampling each arm infinitely many times, this quantity will tend towards 1. What makes the optimal allocation optimal is the rate at which the posterior of the true arm convergence towards 1.

Note that the optimal allocation assumes knowledge about the underlying true value and therefore does not need to be adaptive. It can be framed as

the following thought experiment: Assume you know the true underlying means. Yet your adversary doesn't trust your 'knowledge'. He only trusts the sample rewards that he can observe for himself. Now it is your task to leverage your knowledge about the true means to sample arms in a fashion convincing the adversary as quickly as possible.

Rate of convergence

Russo [3] shows that the rate at which the posterior Π_n of the set parameters under which the true best arm l^* is optimal Θ^* cannot be faster than the following:

$$\Pi_n(\Theta_{l^*}) = 1 - \Pi_n(\Theta_{l^*}^c) = 1 - \exp\{-n\Gamma^*\} \tag{2.12}$$

$$\Gamma^* = \max_{\psi} \min_{\theta \in \Theta_{l^*}^c} \sum_{l \in [k]} \psi_l d(\theta_l^* || \theta_l) \tag{2.13}$$

where n corresponds to the number of samples acquired.

The allocation ψ maximizing this quantity is what we refer to as optimal allocation.

Defining properties

Russo's underlying idea is that the optimal allocation gathers equal *evidence*, e.g. compared to having equal effort, as for a uniform distribution. This notion of evidence relies on the comparison between the true best against all other arms. It takes both their respective true means as well as sampling frequencies into consideration. For an allocation ψ, he defines

$$C_i(\psi_{l^*}, \psi_i) := \min_{x \in \mathbb{R}} \psi_{l^*} d(\theta_{l^*}^* || x) + \psi_i d(\theta_i^* || x) \tag{2.14}$$

and goes on to show that the optimal allocation $\psi^{\frac{1}{2}*}$ is identified by fulfilling the condition

$$\forall i_1, i_2 \neq l^* : C_{i_1}(\psi_{l^*}, \psi_{i_1}) = C_{i_2}(\psi_{l^*}, \psi_{i_2}) \tag{2.15}$$

Intuitively, this expresses that for every suboptimal arm, an equal amount of evidence of suboptimality has been gathered.

Moreover, Russo goes on to show the uniqueness of the optimal allocation.

2.5.3 A Constrained Optimal Allocation

In order to bridge the gap between algorithm and optimal allocation, Russo introduces the concept of *constraining* the optimal allocation. His algorithm naturally implies the constraint that in the limit, β of the measurement effort is allocated to the true best arm.

He is able to show that his algorithm's allocation converges to the overall optimal allocation by showing that

- The algorithm's average allocation $\bar{\psi}_n$ converges to the optimal allocation under the constraint $\psi^*_{l^*} = \beta$
- The hyperparameter β can be tuned to equal the overall optimal value.

The optimal convergence exponent under said constraint becomes:

$$\Gamma^*_\beta = \max_{\psi:\psi_{l^*}=\beta} \min_{\theta\in\Theta^c_{l^*}} \sum_{l\in[k]} \psi_l d(\theta^*_l||\theta_l) \tag{2.16}$$

2.5.4 Top-Two Thompson Sampling algorithm

Russo proposes different algorithms that satisfy aforementioned theoretical results, yet suggests that one of them outperfoms the other empirically: Top-Two Thompson Sampling (TTTS) algorithm.

The main idea behind TTTS is to repeat obtaining candidates through Thompson sampling until two different candidates have been proposed. This illustrates how Thompson sampling, usually only truly useful for exploit-explore settings can be used for pure-exploit: some of its focus is shifted towards inferior-looking candidates. Among those two candidates, the former is picked with probability β, explaining the provenance of the hyperparameter. Note the difference between Algorithm 1 and Algorithm 2: an additional level of 'randomization'.

Algorithm 2 Given a posterior Π_n in step $n+1$

$\hat{\theta} \sim \Pi_n$
$l_1 := \arg\max(\hat{\theta})$
repeat
  $\hat{\theta} \sim \Pi_n$
  $l_2 := \arg\max(\hat{\theta})$
until $l_1 \neq l_2$
$B \sim Bernoulli(\beta)$
if $B = 1$ **then**
  $l_{n+1} := l_1$
else
  $l_{n+1} := l_2$
Play l_{n+1}, observe reward and update priors

Russo's formal treatment of this algorithm relies on the assumption that observations follow 1-dimensional distributions belogning to the family of

exponential distributions. Moreover, he allows priors that are non-conjugate to the posteriors.

2.5.5 Alternative approaches

What are alternative approaches to Russo's? How do they compare against Russo's?

2.6 Top-m Arm Identification

Top-m Arm Identification is a generalization of Best Arm Identification. The objective is to identify the set of arms, with cardinality m, which contains the m best arms.

Applications of top-m arm identification are similar to those mentioned in 2.5.1. Natural explications for desiring the identification m instead of a single high-quality option can be diversification or regulation.

2.6.1 Current approaches: LUCB

How are those methods evaluated? What are possible qualitative shortcomings of those methods? What are possible quantitative shortcomings of those methods?

The general LUCB algorihm can be seen in Algorithm 3. For every arm l, an empirical mean $\hat{\mu}_l^t$ is kept and updated after each sample. The overall idea is to compute a confidence bound for every arm, in every step. Per round, arms are separated into two sets: the arms with the m best empirical means, $Top(t)$ and the rest, $Bottom(t)$. The arm with the lowest lower confidence bound from $Top(t)$ and the arm with the hightest upper confidence bound from $Bottom(t)$ are sampled and all information updated. The algorithm stops once its stopping criterion is met. The confidence bound of an arm l equals $(\hat{\mu}_l^t - \beta(l), \hat{\mu}_l^t + \beta(l))$. Kalyanakrishnan et al. [2] propose a concrete instantiation:

$$\beta(l) =$$

Algorithm 3 Given a prior Π_n in step n

Play l_n, observe reward and update priors
Sample each arm once
repeat
  Compute $Top(t), Bottom(t), h_*^t, l_*^t$
  Sample h_*^t and l_*^t
  Update $\hat{\mu}_t$ and confidence bounds
until $\mu < \varepsilon$

Confidence estimation 1

Confidence estimation 2

Confidence estimation 3

Chapter 3

Charactarizing the Constrained Optimal Top-m Allocation

TODO Talk about limitations on distributions.

Analogously to 2.5.2, we define the optimal top-m allocation by its convergence rate of the posterior mass put on parameters reflecting the true top-m arms $\Pi_n(\Theta_{m,S^*})$. Moreover, just as in Russo's top-1 case, the algorithm we will propose introduces a constraint. We will also put the optimal allocation under that constraint with the idea being that this is but a hyperparameter that can be optimized over.

In Section 3.1 we first introduce some of Russo's results that also apply in our scenario. Then we will introduce some general properties, followed by a complete characterization of the optimal allocation. Moreover, we will show an example of a concrete optimal allocation in Section 3.2. Proofs of the statements are from Section 3.1 are provided in Section 3.3.

Overall we hope to convince the reader that the characterization results portrayed in this chapter make for a natrual generalization of Russo's top-1 case. The guiding theme will be that instead of comparing the best arm l^* to a suboptimal arm in order to identify the one it is hardest to distinguish from it, we will compare the a pairs of optimal and suboptimal arms in order to find to the pair that is the hardest to distinguish. Note that again, distinction revolves around two factors: the frequency with which an arm has been sampled, $\bar{\psi}_{n,l}$, as well as the proximity of its true mean, i.e. comparing θ_l^*.

Throughout this whole chapter we will assume that every true mean is unique, i.e.

$$\forall l_1, l_2 \in [k] : l_1 \neq l_2 \Rightarrow \theta_{l_1}^* \neq \theta_{l_2}^* \tag{3.1}$$

TODO: How would the results look like without the relaxation from $theta_i$ to bar_theta_i? TODO: Could everything be done with j instead of i?

3.1 Statements

Russo proves a proposition about the posterior convergence rate of general parameter sets $\tilde{\Theta}$.

Proposition 3.1 (Russo: Proposition 5) *For any open set $\tilde{\Theta} \subset \Theta$ and average allocation $\bar{\psi}_n$*

$$\Pi_n(\tilde{\Theta}) \doteq \exp\{-n \inf_{\theta \in \tilde{\Theta}} \sum_{l \in [k]} \bar{\psi}_{n,l} d(\theta_l^* || \theta_l)\}$$

This proposition already sets the tone by expressing that the posterior of a set depends on a property of a single contained element θ. The property in question is how hard it is to distinguish θ from the truth θ^*.

As mentioned before, we seek to analyze how fast Θ_{S^*} converges to 1. Clearly we have $\Theta_{S^*} = \Theta - \Theta^c_{S^*}$. Hence instead of analyzing the rate of convergence of $\Pi_n(\Theta_{S^*})$ to 1, we can analyze the rate of convergence of $\Pi(\Theta^c_{S^*})$ to 0.

Instead of analyzing $\Pi_n(\Theta^c_{m,S^*})$ directly, we express it with via $\Theta_{m,l}$ and its relaxation $\bar{\Theta}_i$. Intuitively, $\bar{\Theta}_i$ is the set of parameters under which i proves that S^* is not optimal.

$$\bar{\Theta}_i = \{\theta \in \Theta | \text{top-}m(\theta, S^* \cup \{i\}) \neq S^*\} \tag{3.2}$$

Observe that we have $\bar{\Theta}_i \supsetneq \Theta_i$. Moreover we present a useful relationship between Θ^c_{m,S^*} and $\bar{\Theta}_i$:

Lemma 3.2

$$\Theta^c_{m,S^*} = \bigcup_{i \notin S^*} \bar{\Theta}_i = \Theta - \bigcap_{j \in S^*} \Theta_j$$

Leveraging this relationship of the sets allows us to bridge the gap between the posterior of Θ^c_{m,S^*} and the posterior of individual sets $\bar{\Theta}_i$. Note the transition from a union of sets to a minimum over sets permitted by the usage of the $\doteq$ relation, as shown in Lemma 3.3.

Lemma 3.3 *If $\alpha_{n,S^*} \to 1$, then*

$$\Pi_n(\Theta^c_{m,S^*}) \doteq \max_{i \notin S^*} \Pi_n(\bar{\Theta}_i) \doteq 1 - \min_{j \in S^*} \Pi_n(\Theta_j) \tag{3.3}$$

Plugging $\bar{\Theta}_i$ into Proposition 3.1 leaves us with a sum of KL divergences. We seek to simplify this sum to individual terms just after defining:

$$C_{j,i}(\psi_j, \psi_i) = \min_{x \in \mathbb{R}} \psi_j d(\theta_j^* || x) + \psi_i d(\theta_i^* || x) \tag{3.4}$$

TODO: Talk about player scenario?

In particular, $C_{j,i}$ can be thought of as evidence that j is distinct from i. Quite naturally, we want this evidence to be as large as possible for every pair $(j, i) \in S^* \times S^{*c}$.

Lemma 3.4 *For any $i \notin S^*$ and any allocation ψ,*

$$\min_{\theta \in \bar{\Theta}_i} \sum_{l \in [k]} \psi_l d(\theta_l^* || \theta_l) = \min_{j \in S^*} C_{j,i}(\psi_j, \psi_i) \tag{3.5}$$

Intuitively, this lemma tells us that for parameters in $\bar{\Theta}_i$ the weighted sum of KL divergences can be by simplified to only two arms. One of those arms will be i. The other has arm to stem from the true set of arms S^* in order to satisfy $\bar{\Theta}_i$'s requirement of 'disproving' S^*. This arm $j \in S^*$ is chosen as to seem the 'least distinctive' from i. Distinction is made up of two aspects, as seen in the definition of $C_{j,i}$ in (3.4): how much this arm j is sampled, i.e. ψ_j, and how different its true mean θ_j^* is from the true mean θ_i^* of i. The latter difference is captured by the KL divergence between both arms and a minimal x, individually.

We can apply those statements to the quantity we care about: the mass the posterior puts on the complement of parameters reflecting the true top-m arms, $\Theta_{S^*}^c$.

For a given allocation ψ, we have:

$$\begin{aligned}
\Pi_n(\Theta_{S^*}^c) &\doteq \max_{i \notin S^*} \Pi_n(\bar{\Theta}_i) \text{ (Lemma 3.3)} && (3.6)\\
&\doteq \max_{i \notin S^*} \exp\{-n \inf_{\theta \in \bar{\Theta}_i} \sum_{l \in [k]} \bar{\psi}_{n,l} d(\theta_l^* || \theta_l)\} \text{ (Proposition 3.1)} && (3.7)\\
&= \max_{i \notin S^*} \exp\{-n \min_{j \in S^*} C_{j,i}(\psi_j, \psi_i)\} \text{ (Lemma 3.4)} && (3.8)\\
&= \exp\{-n \min_{i \notin S^*} \min_{j \in S^*} C_{j,i}(\psi_j, \psi_i)\} && (3.9)
\end{aligned}$$

In other words, the rate at which the posterior converges to the truth with increasing number of samples is defined by the pair of optimal and suboptimal arms for which the least evidence exists.

As we've described, the optimal allocation $\psi^{\frac{1}{2}*}$ is the one which makes the quantity from Equation (3.9) as small as possible. We have:

$$\begin{aligned}
\psi^{\frac{1}{2}*} &= \arg\min_{\psi} \exp\{-n \min_{i \notin S^*} \min_{j \in S^*} C_{j,i}(\psi_j, \psi_i)\} && (3.10)\\
&= \arg\max_{\psi} \min_{i \notin S^*} \min_{j \in S^*} C_{j,i}(\psi_j, \psi_i) && (3.11)
\end{aligned}$$

For the sake of convenience we define the optimal exponent:

$$\Gamma^* = \max_{\psi} \min_{i \notin S^*} \min_{j \in S^*} C_{j,i}(\psi_j, \psi_i) \tag{3.12}$$

An Optimal Constrained Allocation Our proposed algorithm will always allocate $\frac{1}{2}$ of its samples to arms in S^* in the long run. As a consequence it may not attain the overall optimal exponent without hyperparameter tuning. Hence we consider a modified, constrained setting under which the algorithm performs optimally. By adapting (3.11) and (3.12) we obtain:

$$\psi^{\frac{1}{2}*} = \arg\max_{\psi:\psi_{S^*}=\frac{1}{2}} \min_{i\notin S^*} \min_{j\in S^*} C_{j,i}(\psi_j, \psi_i) \tag{3.13}$$

$$\Gamma^{\frac{1}{2}*} = \max_{\psi,\psi_{S^*}=\frac{1}{2}} \min_{i\notin S^*} \min_{j\in S^*} C_{j,i}(\psi_j, \psi_i) \tag{3.14}$$

Hence we have established the optimal constrained rate of convergence $\Gamma^{\frac{1}{2}*}$ of the optimal constrained allocation $\psi^{\frac{1}{2}*}$ by making a link to the minimization over $C_{j,i}$. We seek to further characterize $\psi^{\frac{1}{2}*}$ by relying on the same maxim as in Russo's scenario: we expect the measurement plan to collect *equal evidence*, not *equal measurement* for every arm. Hence in Proposition 3.8 we will show that the optimal measurement plan will fulfill Equation (3.15). Before doing so, we set the stage by enounciating two properties of $C_{j,i}$.

Lemma 3.5 *Each* $\min_{j\in S^*} C_{j,i}(\psi_j, \psi_i)$ *is a concave function.*

Lemma 3.6 *Given a* $j \in S^*$*, the solution to the minimization problem* (3.4) *in* x *is* $\bar{\theta} \in \mathbb{R}$*, satisfying:*

$$A'(\bar{\theta}) = \frac{\psi_j A'(\theta_j^*) + \psi_i A'(\theta_i^*)}{\psi_j + \psi_i}$$

where $A'(\theta)$ *is the mean observation under* θ*. Therefore*

$$C_{j,i}(\psi_j, \psi_i) = \psi_j d(\theta_j^* || \bar{\theta}) + \psi_i d(\theta_i^* || \bar{\theta})$$

Lemma 3.7 *For fixed* $i \notin S^*$ *and* $j \in S^*$*,* $C_{j,i}$ *is strictly increasing in both of its arguments.*

Proposition 3.8 *The solution to the optimzation problem 3.13 is the allocation* $\psi^{\frac{1}{2}*}$*, which is unique and satisfies*

$$\forall j_1, j_2 \in S^*, \forall i_1, i_2 \notin S^* : C_{j,i}(\psi_{j_1}^{\frac{1}{2}*}, \psi_{i_1}^{\frac{1}{2}*}) = C_{j,i}(\psi_{j_2}^{\frac{1}{2}*}, \psi_{i_2}^{\frac{1}{2}*}) \tag{3.15}$$

If $\psi_n = \psi^{\frac{1}{2}*}$ *for all* n*, then*

$$\Pi_n(\Theta_{S^*}^c) \doteq \exp\{-n\Gamma_{\frac{1}{2}}^*\}.$$

Moreover, under any other adaptive allocation rule, if $\bar{\psi}_{n,S^*} \to \frac{1}{2}$ *then*

$$\lim_{n\to\infty} \sup -\frac{1}{n} \log \Pi_n(\Theta_{S^*}^c) \leq \Gamma_{\frac{1}{2}}^*$$

almost surely.

3.2 A Concrete Optimal Allocation

For the sake of concreteness, we present numeric values of optimal allocations, both constrained and unconstrained for two top-m identification scenarios. Hence given, the true means θ^*, we seek to determine $\psi^{\frac{1}{2}*}$ and $\psi^{\frac{1}{2}*}$. Relying on (3.15), we have a massively over-determined system of equations. The latter can be approximately solved by numerical methods, in our case least squares minimization.

For this concrete example, we assume arm rewards to follow Bernoulli distributions with means θ_1 and θ_2 respetively. Thanks to this assumption, the KL divergences can be minimized analytically with great convenience. We refer to Appendix A.1 for greater details. Figure 3.1 indicates the probability mass put onto each arm for each scenario. The results were produced for a top-4 scenario with

- $\theta_1 = [.1,\ .2,\ .3,\ .4,\ .5,\ \underbrace{.6,\ .7,\ .8,\ .9}_{S^*}]$
- $\theta_2 = [.4,\ .425,\ .45,\ .475,\ .5,\ \underbrace{.525,\ .55,\ .575,\ .6}_{S^*}]$

For further details regrading the simulation implementation we refer to the code in our repository [1].

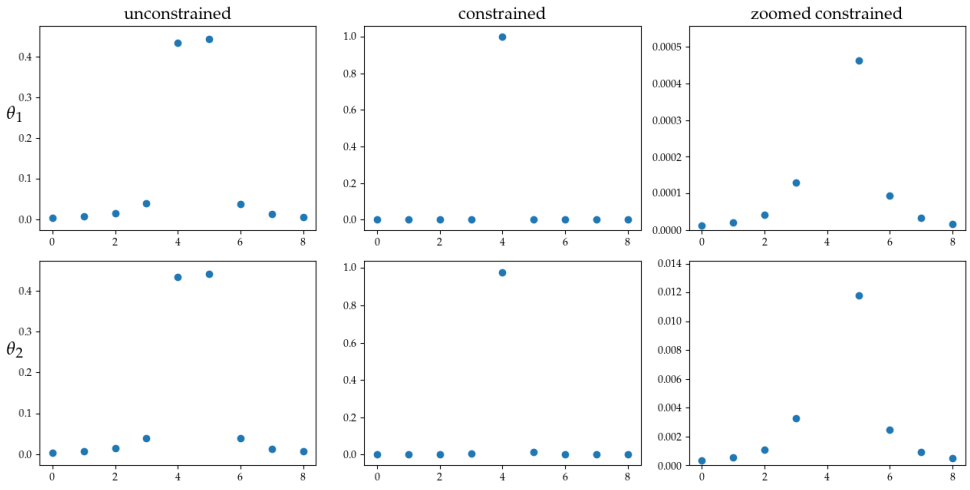

Figure 3.1: Unconstrained and constrained optimal allocation for θ_1 and θ_2, top-4

TODO: Propose an intuition why the concentration is so high in unconstrained scenario.

[1] https://github.com/kkleindev/ttts/compute_optimal_allocation.py

3.3 Proofs

Proof (Lemma 3.2) We first show the relationship between $\Theta^c_{S^*}$ and sets $\bar{\Theta}_i$ and then show the relationship between $\Theta^c_{S^*}$ and Θ_j.

$$\Theta^c_{m,S^*} = \{\theta \in \Theta | \min_{j \in S^*} \theta_j > \max_{i \notin S^*} \theta_i\}^c \tag{3.16}$$
$$= \{\theta \in \Theta | \max_{i \notin S^*} \theta_i \geq \min_{j \in S^*} \theta_j\} \tag{3.17}$$
$$= \{\theta \in \Theta | \exists i \notin S^* : \theta_i \geq \min_{j \in S^*} \theta_j\} \tag{3.18}$$
$$= \{\theta \in \Theta | \exists i \notin S^* : \text{top-}m(\theta, S^* \cup \{i\}) \neq S^*\} \tag{3.19}$$
$$= \bigcup_{i \notin S^*} \{\theta \in \Theta | \text{top-}m(\theta, S^* \cup \{i\}) \neq S^*\} \tag{3.20}$$
$$= \bigcup_{i \notin S^*} \bar{\Theta}_i \tag{3.21}$$
$$\Theta_{m,S^*} = \{\theta \in \Theta | \min_{j \in S^*} \theta_j > \max_{j \in S^*} \theta_j\} \tag{3.22}$$
$$= \{\theta \in \Theta | \bigwedge_{j \in S^*} \theta_j > \max_{j \in S^*} \theta_j\} \tag{3.23}$$
$$= \bigcap_{j \in S^*} \{\theta \in \Theta | \theta_j > \max_{j \in S^*} \theta_j\} \tag{3.24}$$
$$= \bigcap_{j \in S^*} \{\theta \in \Theta | j \in \text{top-}m(\theta, [k])\} \tag{3.25}$$
$$= \bigcap_{j \in S^*} \Theta_j \tag{3.26}$$
$$\Theta^c_{m,S^*} = \Theta - \Theta_{m,S^*} \tag{3.27}$$
$$\Theta^c_{m,S^*} = \Theta - \bigcap_{j \in S^*} \Theta_j \tag{3.28}$$

□

Proof (Lemma 3.3) First, we prove the equality for $i \notin S^*$, followed by the equality for $j \in S^*$.

The union from Lemma 3.2 has an additive effect with respect to the probability distribution Π_n. There are $k - m$ possible is and each single one leads to a set, whose density is bounded by the maximal density of all such sets. This gives us:

$$\max_{i \notin S^*} \Pi_n(\bar{\Theta}_i) \leq \Pi_n(\Theta^c_{m,S^*}) \leq (k-m) \max_{i \notin S^*} \Pi_n(\bar{\Theta}_i) \tag{3.29}$$

We use the definition of $\doteq$ to compare lower bound and upper bound.

$$\lim_{n \to \infty} \frac{1}{n} \log \frac{\max_{i \notin S^*} \Pi_n(\bar{\Theta}_i)}{(k-m) \max_{i \notin S^*} \Pi_n(\bar{\Theta}_i)} = \lim_{n \to \infty} \frac{1}{n} \log \frac{1}{k-m} \to 0 \tag{3.30}$$

Thanks to the lower and upper bound being 'equal' in the $\doteq$ sense, we can apply the squeeze theorem. We obtain the desired result:

$$\Pi_n(\Theta^c_{m,S^*}) \doteq \max_{i \notin S^*} \Pi_n(\bar{\Theta}_i) \tag{3.31}$$

For $j \in S^*$, we follow a very similary path by first leveraging the set equality from Lemma 3.2. Instead of a union, as was the case for $i \in S^*$, we are now confronted with an intersection. This implies a mulitplicative effect on the distribution Π_n instead of an additive one.

$$1 - \min_{j \in S^*} \Pi(\Theta_j) \leq \Pi(\Theta^c_{m,S^*}) \leq 1 - \min_{j \in S^*} (\Pi(\Theta_j))^m \tag{3.32}$$

$$\tag{3.33}$$

Again, we will compare upper and lower bound in the $\doteq$ sense.

$$\lim_{n\to\infty} \frac{1}{n} \log\left(\frac{\min_{j\in S^*} \Pi(\Theta_j)}{\min_{j\in S^*} (\Pi(\Theta_j))^m}\right) = \lim_{n\to\infty} \frac{-(m-1)}{n} \log(\min_{j \in S^*} \Pi(\Theta_j)) \tag{3.34}$$

Observe that $1 \geq \min_{j\in S^*} \Pi(\Theta_j) \geq \alpha_{n,S^*} \to 1$. Hence the limit of the fraction goes to 0 and we have $\min_{j\in S^*} \Pi(\Theta_j) \doteq \min_{j\in S^*} (\Pi(\Theta_j))^m$. Applying, by the Squeeze theorem it follows that

$$\Pi_n(\Theta^c_{m,S^*}) \doteq 1 - \min_{j \in S^*} \Pi_n(\Theta_j) \tag{3.35}$$

□

Proof (Lemma 3.4)

$$\begin{aligned}
\min_{\theta \in \bar{\Theta}_i} D_\psi(\theta^* || \theta) &= \min_{\theta \in \bar{\Theta}_i} \sum_{j=1}^{k} \psi_j d(\theta^*_j || \theta_j) && (3.36)\\
&= \min_{\theta \in \bar{\Theta}_i} \sum_{j \in S^*} \psi_j d(\theta^*_j || \theta_j) + \psi_i d(\theta^*_i || \theta_i) + \sum_{j \notin S^* \cup \{i\}} \psi_j d(\theta^*_j || \theta_j) && (3.37)\\
&= \min_{\theta \in \bar{\Theta}_i} \sum_{j \in S^*} \psi_j d(\theta^*_j || \theta_j) + \psi_i d(\theta^*_i || \theta_i) && (3.38)\\
&= \min_{j \in S^*} \min_{\theta \in \bar{\Theta}_i} \psi_j d(\theta^*_j || \theta_j) + \psi_i d(\theta^*_i || \theta_i) && (3.39)\\
&= \min_{j \in S^*} \min_{\theta \in \bar{\Theta}_i} \psi_j d(\theta^*_j || \theta_j) + \psi_i d(\theta^*_i || \theta_j) && (3.40)\\
&= \min_{j \in S^*} \min_{x \in \mathbb{R}} \psi_j d(\theta^*_j || x) + \psi_i d(\theta^*_i || x) && (3.41)\\
&= \min_{j \in S^*} C_{j,i}(\psi_j, \psi_i) && (3.42)
\end{aligned}$$

(3.38) follows from the fact that for any feasible θ, we can define an alternative θ' s.t. $\theta'_i = \theta_i$, $\theta'_j = \theta_j$ for all $j \in S^*$ and $\theta'_{i_x} = \theta^*_{i_x}$ for all $i_x \notin S^* \cup \{i\}$.

For such a θ', all terms involving $i_x \notin S^* \cup \{i\}$ are zero by the definition of the KL divergence while all others terms remain unchanged. Hence the minimum occurs with such a θ'. Importantly, θ' remains feasible according to current definitions of $\bar{\Theta}_i$, i.e. $\theta' \in \bar{\Theta}_i$.

(3.39) follows from a similar observation: only a single arm from S^* needs to be inferior to arm i under θ. Recall that the terms of the individual arms do not influence each other. This implies that the minimization will gravitate towards setting all but one arm from S^* in θ to the their true value - as the KL divergence is minimized for the true values. Hence the terms of all but one arm from S^* will be cancelled out by the minimization. As i remains superior to one arm in S^*, we have top-$m(\theta, S^* \cup \{i\}) \neq S^*$. Thereby such a θ is feasible according to $\bar{\Theta}_i$.

(3.40) follows from monotonicity of the KL divergence, as displayed in Equation (A.10), combined with the possibility of $\theta_i = \theta_j$ tells us that the minimum will be reached in the case of equality.

(3.41) follows from observing that our minimization over θ has reduced to a minimization over θ_j. The latter is a one-dimensional real. □

Proof (Lemma 3.5) For the sake of clarity, let us define

$$f(x, (\psi_j, \psi_i)) = \psi_j d(\theta_i^* || x) + \psi_i d(\theta_i^* || x) \tag{3.43}$$

Note that $C_{j,i}(\psi_j, \psi_i) = min_{x \in \mathbb{R}} f(x, (\psi_j, \psi_i))$. Clearly, f is linear in (ψ_j, ψ_i). According to Boyd and Vandenberghe (3.2.5) [1], the minimum over a family of linear functions is concave. □

Proof (Lemma 3.6) By Equation (A.9) we know that for the exponential family of probability distributions it holds that:

$$d(\theta || \theta') = (\theta - \theta')A'(\theta) - A(\theta) + A(\theta')$$

Appyling this identity to (3.41) for a given j gives us:

$$\psi_j d(\theta_j^* || x) + \psi_i d(\theta_i^* || x) \tag{3.44}$$

$$= \psi_j((\theta_j^* - x)A'(\theta_j^*) - A(\theta_j^*) + A(x)) + \psi_i((\theta_i^* - x)A'(\theta_i^*) - A(\theta_i^*) + A(x)) \tag{3.45}$$

$$= -x(\psi_j A'(\theta_j^*) + \psi_i A'(\theta_i^*)) + A(x)(\psi_j + \psi_i) + c \tag{3.46}$$

Where c is independent of x. As we seeks to minimize this quantity with respect to x, we differentiate it with respect to x and set it to 0. This yields:

$$A'(x) = \frac{\psi_j A'(\theta_j^*) + \psi_i A'(\theta_i^*)}{\psi_j + \psi_i}$$

□

Proof (Lemma 3.7) We will proceed to show for fixed $j \in S^*$ and $i \notin S^*$, $C_{j,i}$ is increasing in its first argument ψ_j as its second argument ψ_i follows by symmetry.

Let us define $f(x, \psi_j) = \alpha d(\theta_i^* || x) + \psi_j d(\theta_j^* || x) +$ which implies that $C_{j,i}(\psi_j, \alpha) = \min_{x \in \mathbb{R}} f(x, \psi_j)$. As the KL divergence is non-negative we have:

$$\begin{aligned} f(x, \psi_j + \varepsilon) &= \alpha d(\theta_i^* || x) + \psi_j d(\theta_j^* || x) + \varepsilon d(\theta_j^* || x) && (3.47) \\ &> \alpha d(\theta_i^* || x) + \psi_j d(\theta_j^* || x) && (3.48) \\ &= f(x, \psi_j) && (3.49) \end{aligned}$$

And therefore f is stricly increasing in ψ_j.

We fix two reals $\psi_{j1} < \psi_{j2}$ from $[0, 1]$ as well as their counterparts

$$x_1 = \arg\min_{x \in \mathbb{R}} f(x, \psi_{j1}) \qquad x_2 = \arg\min_{x \in \mathbb{R}} f(x, \psi_{j2}) \tag{3.50}$$

Hence our goal is to show that $f(x_1, \psi_{j1}) < f(x_1, \psi_{j2})$ for minimizing x_1. By Lemma 3.6 both x_1 and x_2 are unique. As x_1 is unique, we have

$$\begin{aligned} f(x_1, \psi_{j1}) &< f(x_2, \psi_{j1}) && (3.51) \\ f(x_2, \psi_{j2}) &< f(x_1, \psi_{j2}) && (3.52) \end{aligned}$$

As f is strictly increasing in its second argument, it holds that $f(x_2, \psi_{j1}) < f(x_2, \psi_{j2})$. Chaining those inequalities together we obtain:

$$f(x_1, \psi_{j1}) < f(x_2, \psi_{j1}) < f(x_2, \psi_{j2}) < f(x_1, \psi_{j2}) \tag{3.53}$$

□

Proof (Proposition 3.8) We prove in the following order: (i) (3.15) must hold for an optimal allocation, (ii) an optimal allocation is unique. After this, the remaining claim, namely that no other constrained allocation can be better, follows directly.

(i) Suppose that $\psi^{\frac{1}{2}*}$ is optimal but does not satisfy (3.15). Hence for some $i_1, i_2 \notin S^*, j_1, j_2 \in S^*$ with $(j_1, i_1) \neq (j_2, i_2)$:

$$C_{j_1,i_1}(\psi_{j_1}^{\frac{1}{2}*}, \psi_{i_1}^{\frac{1}{2}*}) > C_{j_2,i_2}(\psi_{j_2}^{\frac{1}{2}*}, \psi_{i_2}^{\frac{1}{2}*})$$

This implies:

$$C_{j_1,i_1}(\psi_{j_1}^{\frac{1}{2}*}, \psi_{i_1}^{\frac{1}{2}*}) > \min_{i \notin S^*} \min_{j \in S^*} C_{j,i}(\psi_j^{\frac{1}{2}*}, \psi_i^{\frac{1}{2}*})$$

Consider the the measurement plan ψ^{ε} with

- $\psi^{\varepsilon}_{i_1} = \psi^{\frac{1}{2}*}_{i_1} - \varepsilon$
- $\psi^{\varepsilon}_{j_1} = \psi^{\frac{1}{2}*}_{j_1} - \varepsilon$
- $\forall l \notin \{i_1, j_1\} : \psi^{\varepsilon}_{l} = \psi^{\frac{1}{2}*}_{\gamma} + \frac{2\varepsilon}{k-2}$

We can choose ε sufficiently small such that we preserve the inequality

$$C_{j_1,i_1}(\psi^{\varepsilon}_{j_1}, \psi^{\varepsilon}_{i_1}) > C_{j_2,i_2}(\psi^{\varepsilon}_{j_2}, \psi^{\varepsilon}_{i_2}) \tag{3.54}$$

while shifting enough measurement to all other arms such that

$$\min_{i \notin S^*} \min_{j \in S^*} C_{j,i}(\psi^{\varepsilon}_{j}, \psi^{\varepsilon}_{i}) > \min_{i \notin S^*} \min_{j \in S^*} C_{j,i}(\psi^{\frac{1}{2}*}_{j}, \psi^{\frac{1}{2}*}_{i}) \tag{3.55}$$

Hence ψ^{ε} obtains more evidence on the worst possible pair than $\psi^{\frac{1}{2}*}$ and thereby achieves a better convergence rate (3.14). In other words, $\psi^{\frac{1}{2}*}$ is not optimal, which is a contradiction.

(ii) Suppose that there are optimal ψ^1, ψ^2, therefore both satisfying (3.15) with the exact same value C^*. It follows that there is at least one l s.t. $\psi^1_l \neq \psi^2_l$. W.l.o.g. assume $l = i_x \notin S^*$ and $\psi^1_{i_x} > \psi^2_{i_x}$.

We proceed by case distinction and show that each leads to a contradiction.

- Only one arm is disinct.

  For some $\varepsilon > 0$ and any $j \in S^*$ have

$$\begin{aligned} C_{j,i_x}(\psi^2_j, \psi^2_{i_x} + \varepsilon) &= C_{j,i_x}(\psi^1_j, \psi^2_{i_x} + \varepsilon) && (3.56)\\ &= C_{j,i_x}(\psi^1_j, \psi^1_{i_x}) && (3.57)\\ &= C^* && (3.58)\\ &= C_{j,i_x}(\psi^2_j, \psi^2_{i_x}) && (3.59) \end{aligned}$$

  Which is a contradiction as ε is positive C_{j,i_x} strictly increasing by Lemma 3.7.

- More that one arm is distinct, but they all belong to either S^* or S^{*c}.

  As our distinct value so far comes from S^{*c}, let's also assume, w.l.o.g., $i_y \notin S^*$ with $i_y \neq i_x$.

  Note that indepdently of whether $\psi^1_{i_y} \geq \psi^2_{i_y}$ or $\psi^1_{i_y} < \psi^2_{i_y}$ holds, the previous argument can be applied for any $j \in S^*$.

- At least one arm is distinct in both S^* and S^{*c}.

  Let's assume first that the distinctive optimal arm is $j_x \in S^*$. Observe that $\psi^1_{j_x} < \psi^2_{j_x}$ has to hold, otherwise both i_x and j_x were allocated more weight in ψ^1 than in ψ^2. By Lemma 3.7 we recall the increase of C_{j_x,i_x} in both of its arguments. This implies greater evidence for ψ^1 than for ψ^2, a contradiction to the assumption that both are optimal.

  Recall our constraint $\frac{1}{2} = \sum_{j \in S^*} \psi_l = \sum_{i \notin S^*} \psi_i$, which has to hold for both ψ^1 and ψ^2. Hence for psi^1's over-allocation on i_x, there has to be an $i_y \notin S^*$ for which ψ^1 under-allocates, compared to psi^2. Summarizing, we have:

  - $\psi^1_{i_x} = \psi^2_{i_x} + \varepsilon$
  - $\psi^1_{j_x} = \psi^2_{j_x} - \varepsilon'$
  - $\psi^1_{i_y} = \psi^2_{i_y} - \varepsilon''$

  Combining this with the fact that all C values from both ψ^1 and ψ^2 have to equal one another, we obtain:

$$C_{j_x,i_y}(\psi^2_{j_x} - \varepsilon', \psi^2_{i_y} - \varepsilon'') = C_{j_x,i_y}(\psi^1_{j_x}, \psi^1_{i_y}) \tag{3.60}$$
$$= C^* \tag{3.61}$$
$$= C_{j_x,i_y}(\psi^2_{j_x}, \psi^2_{i_y}) \tag{3.62}$$

  which, again, is a contradiction by the strictly increaseing nature of C_{j_x,i_y}. □

Chapter 4

Top-2m XOR Thompson Sampling

In this chapter we will both present and analyze our suggested generalization of Russo's Top-Two Thompson sampling: Top-2m XOR Thompson sampling. The underlying idea still revolves around repeatedly apllying Thompson sampling until two different candidates are obtained.

Section 4.1 introduces the algorithm and provides some explanations with regards to generalization decisions. It also introduces the constraint $\psi_{S^*} = \frac{1}{2}$. Subsequently, Section 4.2 analyzes properties of the algorithm. In particular it presents bounds on the measurement plan, states general consequences of finite measurement and discusses implications of under- and over-allocation. We firmly believe that all of those are very useful to show that this algorithm's measurement plan converges to the optimal constrained measurement plan $\psi^{\frac{1}{2}*}$. Proofs for those statements are provided in Section 4.4. Additionaly, we provide some empirial results in Section 4.3.

4.1 Algorithm

As a generalization of Russo's Top-Two, the main differences lie in the fact that candidates are sets. Hence Thompson sampling is repeated until set inequality is reached. A greater difference is due to fact that we can only ever sample individual arms. Hence in this set-scenario, we still need a mechanism to select a single arm from a set. Therefore, when generalizing Russo's approach, the central and unavoidable question arises: 'How to select a single arm from two unequal set candidates?'

According to a suggestion Russo gives in his outlook, we decided to tackle this question by splitting up in two steps.

First, given candidates S_1 and S_2 with $S_1 \neq S_2$, we compute the set of elements which are contained in exactly one of both sets, short the XOR of both sets. Naturally, this restricted set is of cardinality at least 2. Intuitively

it points us towards arms which are 'uncertain'. Observe that arms which are clearly suboptimal are very unlikely to appear in either S_1 or S_2 through Thompson sampling. At the same time, arms that are clearly optimal are very likely to appear in both S_1 and S_2 through Thompson sampling. Hence arms that only appear in either of them, are neither clearly optimal nor clearly suboptimal.

As mentioned before, the XOR of S_1 and S_2 will a priori not be a singleton. Hence we need to define an approach on how to select from the XOR. According to both Russo's suggestion as well as Occam's razor, we opted for uniform selection.

We believe that the XOR approach is essential for the correctness of the algorithm, whereas

We present the appraoch to sample an arm in step $n+1$ in Algorithm 4. This appraoch can be repeated either for a fixed amount of samples or until a specific confidence level is reached. Note that the confidence level can be approximated in every step. The confidence level is equal to $\Pi_n(\Theta_S)$ where $S = \arg\max_{S'} \Pi_n(\Theta'_S)$. Concretely, this quantity can be estimated by drawing samples from Π_n and identifying for which set S the fraction $\frac{\#\text{ samples in which } S \text{ is optimal}}{\#samples}$ is largest. The fraction of this S approximates the confidence level.

Algorithm 4 Given a posterior Π_n in step $n+1$

$\hat{\theta} \sim \Pi_n$
$S_1 = \text{top-}m(\theta)$
repeat
  $\hat{\theta} \sim \Pi_n$
  $S_2 = \text{top-}m(\hat{\theta})$
until $S_1 \neq S_2$
$I_{n+1} \sim \mathcal{U}(S_1 \oplus S_2)$
Play I_{n+1}, observe reward and update posterior

We expect this algorithm to induce a measurement plan ψ such that $\psi_{S^*} = \frac{1}{2}$. This expectation gives rise to the constrained optimization from the previous chapter. Note that this constraint is a consequence of the design decision of which operation to apply to candidates S_1 and S_2. For instance, substituting the uniform distribution by another distribution should alter the hyperparamter $\beta = \psi_{S^*}$.

4.2 Analysis

4.3 Empirical behaviour

What true distributions are assumed? What prior and posterior distributions are assumed? How is C computed? How is alpha computed, as it is defined via a huge integral? How is psi computed, as there is no closed form?

4.4 Proofs

Chapter 5

Conclusion

Appendix A

Appendix

A.1 Computing $C_{j,i}$ for Bernoulli means

We assume that every arm l follows a Bernoulli distribution with parameter θ_l. The option space for rewards being $\{0, 1\}$, they are discretely distributed. Let us reiterate the definition of the KL divergence for discrtete distributions:

$$d(p||q) = \sum_{y \in Y} p(y) \log(\frac{p(y)}{q(y)})$$

where Y corresponds the option space for the outcome. Instantiating this definition with our scenario, we observe that $Y = \{0, 1\}$, $p(y = 1) = \theta_l$ as well as $p(y = 0) = 1 - \theta_l$, for a given arm l.

This yields:

$$d(\theta_l||x) = \theta_l \log(\frac{\theta_l}{x}) + (1 - \theta_l) \log(\frac{1 - \theta_l}{1 - x}) \tag{A.1}$$

Recall that for computing $C_{j,i}$, we seek to minimize the expression from (3.4) $x \in \mathbb{R}$. Hence we are interested in the derivative of (A.1) with respect to x.

$$\frac{\mathrm{d}(d(\theta_l||x)}{\mathrm{d}x} = -\frac{\theta_l}{x} + \frac{(1 - \theta_l)}{1 - x} \tag{A.2}$$

Drawing from the minimization problem of $C_{j,i}$ for given j, i, ψ_j and ψ_i, we define $f(x) = \psi_j d(\theta_j^*||x) + \psi_i d(\theta_i^*||x)$. We proceed by deriving f with

respect to x and setting it to 0.

$$\frac{\mathrm{d}f(x)}{\mathrm{d}x} = \psi_j(\frac{\theta_j}{x} + \frac{(1-\theta_j)}{1-x}) + \psi_i(\frac{\theta_i}{x} + \frac{(1-\theta_i)}{1-x}) \tag{A.3}$$

$$= -\frac{1}{x}(\psi_j\theta_j + \psi_i\theta_i) + \frac{1}{1-x}(\psi_j(1-\theta_j) + \psi_i(1-\theta_i)) \tag{A.4}$$

$$\frac{\mathrm{d}f(x)}{\mathrm{d}x} = 0 \Rightarrow (1-x_0)(\psi_j\theta_j + \psi_i\theta_i) = x_0(\psi_j(1-\theta_j) + \psi_i(1-\theta_i)) \tag{A.5}$$

$$\Rightarrow x_0((\psi_j\theta_j + \psi_i\theta_i) + (\psi_j(1-\theta_j) + \psi_i(1-\theta_i))) = (\psi_j\theta_j + \psi_i\theta_i) \tag{A.6}$$

$$\Rightarrow x_0 = \frac{\psi_j\theta_j + \psi_i\theta_i}{(\psi_j\theta_j + \psi_i\theta_i) + (\psi_j(1-\theta_j) + \psi_i(1-\theta_i))} \tag{A.7}$$

$$\Rightarrow x_0 = \frac{\psi_j\theta_j + \psi_i\theta_i}{\psi_j + \psi_i} \tag{A.8}$$

Hence we have a very intuitive analytical solution for x: it is an average of the means of j and j, weighted by their respective measurement allocation.

A.2 Facts about the exponential family

Russo presents the following insight about the exponential family of probability distributions.

- Its log-partition-function $A(\theta)$ is strictly convex and differentiable.
- Its mean equals $A'(\theta) = \int T(y)p(y|\theta)d\nu(y)$.
- The KL divergence equals:

$$d(\theta||\theta') = (\theta - \theta')A'(\theta) - A(\theta) + A(\theta') \tag{A.9}$$

- The KL divergence satisfies:

$$\theta'' > \theta' \geq \theta \Rightarrow d(\theta||\theta'') > d(\theta||\theta') \tag{A.10}$$

$$\theta'' < \theta' \leq \theta \Rightarrow d(\theta||\theta'') < d(\theta||\theta') \tag{A.11}$$

A.3 Useful technical statements

Lemma A.1

$$\limsup_{n\to\infty} f(n) \leq \Gamma \Rightarrow \exists \varepsilon_n > 0, s.t. \varepsilon_n \to 0, \forall n\ f(n) \leq \Gamma + \varepsilon_n$$

Proof Given the assumption, we can do a case distinction:

- $g(n) = \Gamma$ and $\sup f(n)$ intersect in n_0
  For $n < n_0$, we have $\sup f(n) > \Gamma$. As $\sup f(n)$ is a decreasing function, there are positive, decreasing ε_n such that $\sup f(n) \leq \Gamma + \varepsilon_n$.
  For $n \geq n_0$, we have $\sup f(n) \leq \Gamma$. Hence for such n, $\varepsilon_n = 0$ would already satisfies the constraint. Hence any $\varepsilon_n \to 0$, fulfills the inequality.

  Consequently, there is a positive $\varepsilon_n \to 0$, s.t. $\forall n : f(n) \leq \sup f(n) \leq \Gamma + \varepsilon_n$.

- $g(n) = \Gamma$ and $\sup f(n)$ do not intersect
  By our assumption we have $\forall n \ \sup f(n) \leq \Gamma$. Hence, by setting ε_n to equal 0 for all n, we have a positive $\varepsilon_n \to 0$, such that $f(n) \leq \sup f(n) \leq \Gamma + \varepsilon_n$. □

Bibliography

[1] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, New York, NY, USA, 2004.

[2] Shivaram Kalyanakrishnan, Ambuj Tewari, Peter Auer, and Peter Stone. PAC subset selection in stochastic multi-armed bandits. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012*, 2012.

[3] Daniel Russo. Simple bayesian algorithms for best arm identification. *CoRR*, abs/1602.08448, 2016.

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Declaration of originality

The signed declaration of originality is a component of every semester paper, Bachelor's thesis, Master's thesis and any other degree paper undertaken during the course of studies, including the respective electronic versions.

Lecturers may also require a declaration of originality for other written papers compiled for their courses.

I hereby confirm that I am the sole author of the written work here enclosed and that I have compiled it in my own words. Parts excepted are corrections of form and content by the supervisor.

Title of work (in block letters):

Authored by (in block letters):
For papers written by groups the names of all authors are required.

Name(s):	**First name(s):**

With my signature I confirm that

- I have committed none of the forms of plagiarism described in the 'Citation etiquette' information sheet.
- I have documented all methods, data and processes truthfully.
- I have not manipulated any data.
- I have mentioned all persons who were significant facilitators of the work.

I am aware that the work may be screened electronically for plagiarism.

Place, date	**Signature(s)**

For papers written by groups the names of all authors are required. Their signatures collectively guarantee the entire content of the written paper.