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

150 words

1 Top-m arm identification comprehensible for any scientist

2 Top-m arm identification comprehensible for scientist from the field

1 The problem addressed by this paper

1 Summary of main result

2 How the main result changes previous knowledge

1 Result in general context

2 Broader perspective, understandable by any scientist

Acknowledgements

Many thanks to ...

Contents

Chapter 1

Introduction

Much of Machine Learning revolves around how to learn from observations. Technological advances allow for an extensive gathering of data in a plurality of domains. Notwithstanding, many domains are still, and likely to remain, areas of expensive data acquisition. Due to immense complexity, human behavior as well as many physical processes can only be simulated at very high costs or not at all. As a consequence, data is gathered by experiments that can be long-lasting, strictly constrained in quantity and precious. In light of that, we focus on the following aspect of Machine Learning: how to acquire observations.

In order to capture the complexity of real-world processes, we model outcomes of decisions according to probability distributions. More precisely, we make use of the well-established multi-armed bandits model for that exact purpose. As we've mentioned, we seek to describe a manner of acquiring, or sampling, observations. Naturally, one must ask the question: 'With what exact goal?'

We assume the goal of the data generation and acquisition to be the identification of the m best out of k options. Within the realm of this goal, we attempt to tackle the following problems:

- What is, in theory, the best possible acquisition strategy?
- What is an algorithm that can, under some constraints, mimic this best possible acquisition strategy?

In particular, we seek to address these challenges by generalizing work from Russo [8] on identification of the single best option. Consistent with existing literature, we will continue to refer to options as arms. This terminology stems from the image of a many-armed bandit slot machine.

We assume a frequentist setting. This implies that the options follow fixed distributions, which are unknown to us. Yet, we can learn about those un-

known distributions by observing the samples the environment draws. This can be thought of as being confronted with a noise-free signal that is then polluted with noise from a sensor. In this spirit, we seek to identify the true best options with as few samples as possible and be as confident of our recommendation as possible, which will be formalized.

Our first contribution is the characterization of the optimal acquisition strategy, also referred to as optimal measurement plan or allocation. What makes the optimal allocation optimal is the rate of increase in confidence in the true best options per sample. This optimal strategy is based on a thought experiment of knowing the underlying truth and acquiring samples validating the knowledge as well as possible. As a consequence, it serves as a bound for practical algorithms outside of thought experiments. Intuitively, the guiding theme of the characterization is that the allocation seeks to gather equal *evidence* on every option. We will provide a definition and intuition of evidence closely following Russo's work.

Our second contribution is the definition of a concrete and simple adaptive algorithm. As compared to the optimal allocation, the algorithm changes its measurement plan after every sample, based on the observations it has made. The algorithm, being Bayesian, comes with the upside of confidence estimation and allows for leveraging and expressing domain knowledge through the use of priors. We analyze properties of the algorithm and highlight insights that we deem very useful for proving asymptotic convergence of this algorithm's allocation to the optimal allocation.

In those undertakings we make very few assumptions on the underlying conditions. Concretely, we expect the options to follow distributions from the one-dimensional canonical exponential family. This includes common distributions such as the Bernoulli, binomial with known number of trials, Poisson, exponential, Pareto with known minimal value, chi-squared or the normal distribution with known variance [1].

Chapter 2 introduces the general problem context as well as Russo's work, which we attempt to generalize from top-1 to top-m selection. We characterize the optimal allocation in Chapter 3. Our proposed algorithm, Top-2m XOR Thompson sampling, is presented and analyzed in Chapter 4 before concluding our work in Chapter 5.

[1] https://en.wikipedia.org/wiki/Exponential_family

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 arms to choose from, each choice leading to an outcome, 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.

More concretely, the former concerns itself with maximizing the *cumulative 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. The main task is therefore to estimate the true best arms with high *confidence*.

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-modeled process varies.

The explore-exploit problem can be found in Recommender Systems [7], ad campaigns [1], in the context of Reinforcement Learning [4] [9] and Robotics [2].

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

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.

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. Finding ourselves in a frequentist setting, we assume an underlying true mean of the arms. We refer to this as θ^*. Moreover, this ground truth implies a true best arm l^* and true top-m arms S^*.

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:

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

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

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

$$\Theta_S := \{\theta \in \Theta | S = \text{top-}m(\theta)\} \tag{2.4}$$

$$= \{\theta \in \Theta | \min_{j_1 \in S} \theta_{j_1} > \max_{j_2 \notin S} \theta_{j_2}\} \tag{2.5}$$

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 Θ_S, we have:

$$\Pi_n(\Theta_S) := \int_{\theta \in \Theta_S} \pi_n(\theta) d\theta \tag{2.6}$$

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

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

$$\alpha_{n,S} := \Pi_n(\Theta_S) \tag{2.8}$$

$$\alpha_{n,l} := \sum_{S: l \in S} \Pi_n(\Theta_S) \tag{2.9}$$

$$\tag{2.10}$$

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 measurement plan as $\psi \in [0,1]^k$, 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:

$$\bar{\psi}_{n,l} := \frac{\sum_{n'=1}^{n} \psi_{n',l}}{n} \tag{2.11}$$

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:

$$\lim_{n \to \infty} \frac{1}{n} \log \frac{a_n}{b_n} \to 0 \tag{2.12}$$

Hence $\doteq$ can be vaguely thought of as asymptotic relationship indicating similar growth. Note the asymmetry of the relationship: it rather represents an inequality than an equality. Russo points refers to this relation as *logarithmic equivalence* and points out that a_n and b_n are *equal up to first order in the exponent*. Clearly, it holds that $a_n + b_n \doteq \max\{a_n, b_n\}$ and $ca_n \doteq a_n$ for positive constant c.

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 cumulative 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 of only point estimates for per arm, Thompson sampling requires a distribution over the arms.

Naturally, this distribution over the arms should also leverage the knowledge acquired 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 is the the prior *updated* with the knowledge acquired of a specific round. Thanks to this duality we will continue to only talk about posteriors.

One 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 draw from posterior. 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-1} in step n

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

Intuitively, this is very appealing for the cumulative 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 closely related to Optimal 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.

Up to this point we vaguely talked about confidence. As a matter of fact, different approaches for best arm identification rely on different approaches to express their confidence in the estimation. Herein lies a substantial differentiation and therefore, naturally, it is not trivial to compare different approaches other than by investigating what their best arm output is, i.e. disregarding the confidence. For Bayesian algorithms, it is possible to quantify the confidence the model has in the currently most favorable looking candidate. This quantity corresponds to the mass a posterior puts on parameters favoring 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 $\Pr[\text{output of algorithm is } \varepsilon\text{-correct}] \geq 1 - \delta$, 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.

2.5.2 Optimal fixed 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 fixed 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

is 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. Observe that for any allocation sampling each arm with at least some probability, this quantity will tend towards 1. What makes the optimal allocation optimal is the rate at which the posterior of the true arm converges 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

Recall that Θ_{1,l^*} is the set of parameters under which the true best arm l^* is optimal. Russo [8] shows that the rate at which the posterior Π_n of Θ_{1,l^*} converges towards 1 cannot be faster than the following:

$$\Pi_n(\Theta_{1,l^*}) = 1 - \Pi_n(\Theta^c_{l^*}) = 1 - \exp\{-n\Gamma^*\} \tag{2.13}$$

$$\Gamma^* = \max_{\psi} \min_{\theta \in \Theta^c_{l^*}} \sum_{l \in [k]} \psi_l d(\theta^*_l || \theta_l) \tag{2.14}$$

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. He defines

$$C_i(p_1, p_2) := \min_{x \in \mathbb{R}} p_1 d(\theta^*_{l^*} || x) + p_2 d(\theta^*_i || x) \tag{2.15}$$

Where p_1 and p_2 will typically correspond to the measurement allocations ψ_{l^*} and ψ_i:

$$C_i(\psi_{l^*}, \psi_i) := \min_{x \in \mathbb{R}} \psi_{l^*} d(\theta^*_{l^*} || x) + \psi_i d(\theta^*_i || x) \tag{2.16}$$

He goes on to show that the optimal allocation ψ^* is identified by fulfilling the condition

$$\forall i_1, i_2 \neq l^* : C_{i_1}(\psi_{l^*}, \psi_{i_1}) = C_{i_2}(\psi_{l^*}, \psi_{i_2}) \tag{2.17}$$

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:

$$\Gamma^*_\beta = \max_{\psi:\psi_{l^*}=\beta} \min_{\theta \in \Theta^c_{l^*}} \sum_{l \in [k]} \psi_l d(\theta^*_l || \theta_l) \tag{2.18}$$

2.5.4 Top-Two Thompson sampling algorithm

Russo proposes different algorithms that satisfy aforementioned theoretical results, yet suggests that one of them outperforms 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, most commonly used 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-1} in step n

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

Russo's formal treatment of this algorithm relies on the assumption that observations follow 1-dimensional distributions belonging to the family of exponential distributions. Moreover, he allows priors that are non-conjugate to the posteriors, though does not discuss how to update or sample from them.

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 A common approach: LUCB

The general Lower Upper Confidence Bound algorithm is a common approach for top-m arm identification [6] and typically evaluated in the PAC setting. It can be seen in Algorithm 3.

For every arm l, an empirical mean $\hat{\mu}_l$ 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. We refer to those arms as tl and bu respectively. The algorithm

stops once its stopping criterion is met. The confidence bound of an arm l in a given step equals $(\hat{\mu}_l - \beta(l), \hat{\mu}_l + \beta(l))$. Kalyanakrishnan et al. [5] propose a concrete instantiation by defining the estimation of the confidence bound and stopping crierion:

$$\beta(l) = \sqrt{\frac{1}{2c_l} \ln \frac{knt^4}{\delta}} \tag{2.19}$$

$$\xi = (\hat{\mu}_{bu} + \beta(bu) - (\hat{\mu}_{tl} - \beta(tl)) \tag{2.20}$$

with constant $k = \frac{5}{4}$, c_l the number of times arm l has been sampled and t the number of steps completed so far.

Algorithm 3 Given prior means

```
Sample each arm once
Update μ̂ and confidence bounds
repeat
    Compute Top(t), Bottom(t), tl, bu
    Sample tl and bu
    Update μ̂ and confidence bounds
until ξ < ε
```

Chapter 3

The Constrained Optimal Top-m Allocation

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 $\Pi_n(\Theta_{S^*})$. This, too can be thought of as a thought experiment of convincing an adversary by relying on knowledge of the true means. 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 other words: the best unconstrained allocation is the best constrained allocation with optimal hyperparameter.

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 from Section 3.1 are provided in Section 3.4. We finish the chapter with Section 3.3, a presenting a condition on adaptive allocations, which is sufficient to show convergence to the optimal allocation.

Overall we hope to convince the reader that the characterization results portrayed in this chapter make for a natural generalization of Russo's top-1 case. In the top-1 case, Russo implied comparing the best arm l^* to a suboptimal arm in order to identify the one it is hardest to distinguish from it. Our guiding theme, on the other hand, will be to 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.

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

and as previously mentioned the rewards follow distributions belonging to the family of exponential distributions.

3.1 Characterization 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 $\bar{\psi}_n$*

$$\Pi_n(\tilde{\Theta}) \doteq \exp\{-n \inf_{\theta \in \tilde{\Theta}} \sum_{l \in [k]} \bar{\psi}_{n,l} d(\theta^*_l || \theta_l)\} \tag{3.2}$$

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 we've stressed, the optimal allocation ψ^* is non-adaptive and fixed over time. Thereby the average allocation $\bar{\psi}^*$ corresponds to ψ^*.

As mentioned before, we seek to analyze how fast $\Pi_n(\Theta_{S^*})$ 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 $\Pi_n(\Theta^c_{S^*})$ directly, we express it 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.

$$\bar{\Theta}_i = \{\theta \in \Theta | \text{top-}m(\theta, S^* \cup \{i\}) \neq S^*\} \tag{3.3}$$

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

Lemma 3.2

$$\Theta^c_{S^*} = \bigcup_{i \notin S^*} \bar{\Theta}_i = \Theta - \bigcap_{j \in S^*} \Theta_{m,j} \tag{3.4}$$

Leveraging this relationship of the sets allows us to bridge the gap between the posterior of $\Theta^c_{S^*}$ 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*

$$\Pi_n(\Theta^c_{S^*}) \doteq \max_{i \notin S^*} \Pi_n(\bar{\Theta}_i) \doteq 1 - \min_{j \in S^*} \Pi_n(\Theta_{m,j}) \tag{3.5}$$

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:

$$C_{j,i}(p_1, p_2) = \min_{x \in \mathbb{R}} p_1 d(\theta^*_j || x) + p_2 d(\theta^*_i || x) \tag{3.6}$$

Usually we will encounter $C_{j,i}$ with its firs argument being the measurement plan for arm j, ψ_j and its second argument being the measurement plan for arm i, ψ_i.

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,*

$$\min_{\theta \in \bar{\Theta}_i} \sum_{l \in [k]} \psi_l d(\theta^*_l || \theta_l) = \min_{j \in S^*} C_{j,i}(\psi_j, \psi_i) \tag{3.7}$$

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 arm has 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 under ψ. Distinction is made up of two aspects, as seen in the definition of $C_{j,i}$ in (3.6): 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, $\Theta^c_{S^*}$.

For a given fixed allocation ψ, we have:

$$\begin{aligned}
\Pi_n(\Theta^c_{S^*}) &\doteq \max_{i \notin S^*} \Pi_n(\bar{\Theta}_i) \text{ (Lemma 3.3)} && (3.8)\\
&\doteq \max_{i \notin S^*} \exp\{-n \inf_{\theta \in \bar{\Theta}_i} \sum_{l \in [k]} \psi_{n,l} d(\theta^*_l || \theta_l)\} \text{ (Proposition 3.1)} && (3.9)\\
&= \max_{i \notin S^*} \exp\{-n \min_{j \in S^*} C_{j,i}(\bar{\psi}_j, \bar{\psi}_i)\} \text{ (Lemma 3.4)} && (3.10)\\
&= \exp\{-n \min_{i \notin S^*} \min_{j \in S^*} C_{j,i}(\bar{\psi}_j, \bar{\psi}_i)\} && (3.11)
\end{aligned}$$

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 ψ^* is the one which makes the quantity from Equation (3.11) as small as possible. We have:

$$\psi^* = \arg\min_{\psi} \exp\{-n \min_{i \notin S^*} \min_{j \in S^*} C_{j,i}(\psi_j, \psi_i)\} \tag{3.12}$$

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

For the sake of convenience we define the optimal exponent:

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

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 (3.13) and (3.14) we obtain:

$$\psi^{\frac{1}{2}*} = \arg\max_{\psi:\psi_{S^*}=\frac{1}{2}} \min_{i \notin S^*} \min_{j \in S^*} C_{j,i}(\psi_j, \psi_i) \tag{3.15}$$

$$\Gamma^{\frac{1}{2}*} = \max_{\psi,\psi_{S^*}=\frac{1}{2}} \min_{i \notin S^*} \min_{j \in S^*} C_{j,i}(\psi_j, \psi_i) \tag{3.16}$$

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 (3.19). Before doing so, we set the stage by enunciating three properties of $C_{j,i}$: concavity, existence of a unique solution x and strict monotonicity.

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

Lemma 3.6 *Assume* $A'(\theta)$ *to be the mean observation under* θ*. Given a* $j \in S^*$*, the solution to the minimization problem* (3.6) *in* x *is* $\bar{\theta} \in \mathbb{R}$*, satisfying:*

$$A'(\bar{\theta}) = \frac{\psi_j A'(\theta_j^*) + \psi_i A'(\theta_i^*)}{\psi_j + \psi_i} \tag{3.17}$$

Therefore

$$C_{j,i}(\psi_j, \psi_i) = \psi_j d(\theta_j^* || \bar{\theta}) + \psi_i d(\theta_i^* || \bar{\theta}) \tag{3.18}$$

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 optimization problem 3.15 is the allocation $\psi^{\frac{1}{2}*}$, which is unique and satisfies*

$$\forall j_1, j_2 \in S^*, \forall i_1, i_2 \notin S^* : C_{j,i}(\psi_{j_1}^{\frac{1}{2}*}, \psi_{i_1}^{\frac{1}{2}*}) = C_{j,i}(\psi_{j_2}^{\frac{1}{2}*}, \psi_{i_2}^{\frac{1}{2}*}) \tag{3.19}$$

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.

The attentive reader might have noticed that we started off with relations between Θ_{S^*} and both suboptimal arms i and optimal j, such as in Lemma 3.2 and Lemma 3.3. We stopped doing so in Lemma 3.4 and proceeded to rely on i. We conjecture that the same results could have been reached by focusing on j, through symmetry.

3.2 An exemplary 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 ψ^* and $\psi^{\frac{1}{2}*}$. Relying on (3.19), we have a over-determined system of equations. The latter can be approximately solved by numerical methods, in our case non-linear least squares minimization.

For this concrete example, we assume arm rewards to follow Bernoulli distributions with means θ^1 and θ^2 respectively. 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

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

The unconstrained case on a linear scale of Figure 3.1 might surprise with a lack of symmetry between the measurement effort put on the best suboptimal arm and the worst optimal arm. The log scales paint a more intuitive, symmetric picture.

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

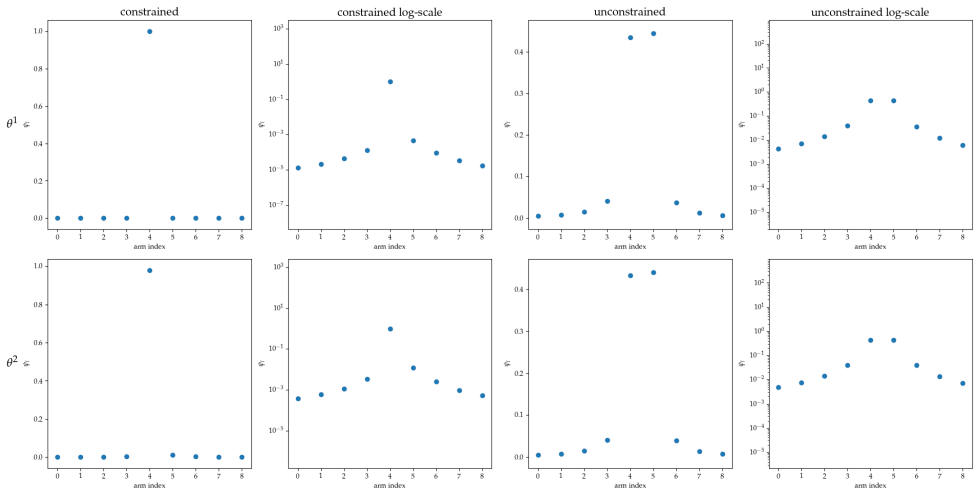

Figure 3.1: Unconstrained and constrained optimal allocation for θ_1 and θ_2, top-4.

3.3 A sufficient condition for optimality

This sufficient condition for convergence towards the optimal allocation is centered around observing consequences of overallocation. Note that whenever an allocation is different from the optimal allocation, this implies both over- and underallocation. We focus on the statement for overallocation, yet an analogous statement for underallocation should be possible.

In particular, the statement expresses that if an arm l has been over-allocated so far, indicated by a $\bar{\psi}_{n,l}$ larger than $\psi_l^{\frac{1}{2}*}$, this over-allocation will be corrected. In this case, correction means that its likelihood of being sampled in the next and round, $\psi_{n,l}$ is very low. The intuition of the whole statement being that if all overallocation is corrected for, then consequently all underallocation is corrected for as well, by the fact that ψ is a probability distribution.

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

Proposition 3.9 *If*

$$\forall l \in [k], \delta > 0 : \sum_{n \in \mathbb{N}} \psi_{n,l} \mathbb{I}[\bar{\psi}_{n,l} \geq \psi_l^* + \delta] < \infty \tag{3.20}$$

then $\bar{\psi}_n \to \psi^*$.

3.4 Proofs

Proof (Lemma 3.2) We first show the relationship between $\Theta_{S^*}^c$ and sets $\bar{\Theta}_i$ and then show the relationship between $\Theta_{S^*}^c$ and $\Theta_{m,j}$.

$$\begin{aligned}
\Theta_{S^*}^c &= \{\theta \in \Theta | \min_{j \in S^*} \theta_j > \max_{i \notin S^*} \theta_i\}^c && (3.21)\\
&= \{\theta \in \Theta | \max_{i \notin S^*} \theta_i \geq \min_{j \in S^*} \theta_j\} && (3.22)\\
&= \{\theta \in \Theta | \exists i \notin S^* : \theta_i \geq \min_{j \in S^*} \theta_j\} && (3.23)\\
&= \{\theta \in \Theta | \exists i \notin S^* : \text{top-}m(\theta, S^* \cup \{i\}) \neq S^*\} && (3.24)\\
&= \bigcup_{i \notin S^*} \{\theta \in \Theta | \text{top-}m(\theta, S^* \cup \{i\}) \neq S^*\} && (3.25)\\
&= \bigcup_{i \notin S^*} \bar{\Theta}_i && (3.26)\\
\Theta_{S^*} &= \{\theta \in \Theta | \min_{j \in S^*} \theta_j > \max_{i \notin S^*} \theta_i\} && (3.27)\\
&= \{\theta \in \Theta | \bigwedge_{j \in S^*} \theta_j > \max_{i \notin S^*} \theta_i\} && (3.28)\\
&= \bigcap_{j \in S^*} \{\theta \in \Theta | \theta_j > \max_{i \notin S^*} \theta_i\} && (3.29)\\
&= \bigcap_{j \in S^*} \{\theta \in \Theta | j \in \text{top-}m(\theta, [k])\} && (3.30)\\
&= \bigcap_{j \in S^*} \Theta_{m,j} && (3.31)\\
\Theta_{S^*}^c &= \Theta - \Theta_{S^*} && (3.32)\\
&= \Theta - \bigcap_{j \in S^*} \Theta_{m,j} && (3.33)
\end{aligned}$$

□

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:

$$\max_{i \notin S^*} \Pi_n(\bar{\Theta}_i) \leq \Pi_n(\Theta^c_{S^*}) \leq (k-m) \max_{i \notin S^*} \Pi_n(\bar{\Theta}_i) \tag{3.34}$$

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

$$\lim_{n\to\infty} \frac{1}{n} \log \frac{\max_{i \notin S^*} \Pi_n(\bar{\Theta}_i)}{(k-m) \max_{i \notin S^*} \Pi_n(\bar{\Theta}_i)} = \lim_{n\to\infty} \frac{1}{n} \log \frac{1}{k-m} \to 0 \tag{3.35}$$

Thanks to the lower and upper bound being 'equal' in the $\doteq$ sense, we can apply the Squeeze theorem[2]. We obtain the desired result:

$$\Pi_n(\Theta^c_{S^*}) \doteq \max_{i \notin S^*} \Pi_n(\bar{\Theta}_i) \tag{3.36}$$

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 multiplicative effect on the distribution Π_n instead of an additive one.

$$1 - \min_{j \in S^*} \Pi(\Theta_{m,j}) \leq \Pi(\Theta^c_{S^*}) \leq 1 - \min_{j \in S^*} (\Pi(\Theta_{m,j}))^m \tag{3.37}$$

$$\tag{3.38}$$

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

$$\lim_{n\to\infty} \frac{1}{n} \log \left(\frac{\min_{j \in S^*} \Pi(\Theta_{m,j})}{\min_{j \in S^*} (\Pi(\Theta_{m,j}))^m} \right) = \lim_{n\to\infty} \frac{-(m-1)}{n} \log \left(\min_{j \in S^*} \Pi(\Theta_{m,j}) \right) \tag{3.39}$$

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

$$\Pi_n(\Theta^c_{S^*}) \doteq 1 - \min_{j \in S^*} \Pi_n(\Theta_{m,j}) \tag{3.40}$$

□

Proof (Lemma 3.4)

$$\min_{\theta \in \bar{\Theta}_i} \sum_{j=1}^{k} \psi_j d(\theta^*_j || \theta_j) = \min_{\theta \in \bar{\Theta}_i} \sum_{j \in S^*} \psi_j d(\theta^*_j || \theta_j) + \psi_i d(\theta^*_i || \theta_i) + \sum_{j \notin S^* \cup \{i\}} \psi_j d(\theta^*_j || \theta_j) \tag{3.41}$$

$$= \min_{\theta \in \bar{\Theta}_i} \sum_{j \in S^*} \psi_j d(\theta^*_j || \theta_j) + \psi_i d(\theta^*_i || \theta_i) \tag{3.42}$$

$$= \min_{j \in S^*} \min_{\theta \in \bar{\Theta}_i} \psi_j d(\theta^*_j || \theta_j) + \psi_i d(\theta^*_i || \theta_i) \tag{3.43}$$

$$= \min_{j \in S^*} \min_{\theta \in \bar{\Theta}_i} \psi_j d(\theta^*_j || \theta_j) + \psi_i d(\theta^*_i || \theta_j) \tag{3.44}$$

$$= \min_{j \in S^*} \min_{x \in \mathbb{R}} \psi_j d(\theta^*_j || x) + \psi_i d(\theta^*_i || x) \tag{3.45}$$

$$= \min_{j \in S^*} C_{j,i}(\psi_j, \psi_i) \tag{3.46}$$

[2] Also referred to as Sandwich teorem

(3.42) 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. $\theta' \in \bar{\Theta}_i$.

(3.43) 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 $\bar{\Theta}_i$.

(3.44) 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.

(3.45) follows from observing that our minimization over θ has reduced to a minimization over θ_j. The latter is a one-dimensional real as we are in the case of distributions belonging to the one-dimensional exponential family.□

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

$$f(x, (\psi_j, \psi_i)) = \psi_j d(\theta^*_j || x) + \psi_i d(\theta^*_i || x) \tag{3.47}$$

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) [3], the minimum over a family of linear functions is concave. □

Lemma 3.6 Assume $A'(\theta)$ to be the mean observation under θ. Given a $j \in S^*$, the solution to the minimization problem (3.6) in x is $\bar{\theta} \in \mathbb{R}$, satisfying:

$$A'(\bar{\theta}) = \frac{\psi_j A'(\theta^*_j) + \psi_i A'(\theta^*_i)}{\psi_j + \psi_i} \tag{3.48}$$

Therefore

$$C_{j,i}(\psi_j, \psi_i) = \psi_j d(\theta^*_j || \bar{\theta}) + \psi_i d(\theta^*_i || \bar{\theta}) \tag{3.49}$$

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

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

Applying this identity to the definition of $C_{j,i}$ Equation (3.6) for given j, i, ψ_j, ψ_i gives us:

$$\psi_j d(\theta_j^*||x) + \psi_i d(\theta_i^*||x) \tag{3.50}$$
$$= \psi_j((\theta_j^* - x)A'(\theta_j^*) - A(\theta_j^*) + A(x)) + \psi_i((\theta_i^* - x)A'(\theta_i^*) - A(\theta_i^*) + A(x)) \tag{3.51}$$
$$= -x(\psi_j A'(\theta_j^*) + \psi_i A'(\theta_i^*)) + A(x)(\psi_j + \psi_i) + c \tag{3.52}$$

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 $f(x, \psi_j, p) = \psi_j d(\theta_j^*||x) + p d(\theta_i^*||x)$ which implies that $C_{j,i}(\psi_j, p) = \min_{x \in \mathbb{R}} f(x, \psi_j, p)$. We first show that f is strictly increasing in ψ_j and will use that to show that the initial claim.

As the KL divergence is non-negative we have:

$$f(x, \psi_j + \varepsilon, p) = \psi_j d(\theta_j^*||x) + \varepsilon d(\theta_j^*||x) + p d(\theta_i^*||x) \tag{3.53}$$
$$\geq p d(\theta_i^*||x) + \psi_j d(\theta_j^*||x) \tag{3.54}$$
$$= f(x, \psi_j, p) \tag{3.55}$$

And therefore f is increasing in ψ_j.

For a given p, we fix $\psi_{j1} < \psi_{j2}$ from $[0, 1]$ as well as their counterparts

$$x_1 = \arg\min_{x \in \mathbb{R}} f(x, \psi_{j1}, p) \qquad x_2 = \arg\min_{x \in \mathbb{R}} f(x, \psi_{j2}, p) \tag{3.56}$$

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

$$f(x_1, \psi_{j1}, p) < f(x_2, \psi_{j1}, p) \tag{3.57}$$
$$f(x_2, \psi_{j2}, p) < f(x_1, \psi_{j2}, p) \tag{3.58}$$

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

$$f(x_1, \psi_{j1}, p) < f(x_2, \psi_{j1}, p) \leq f(x_2, \psi_{j2}, p) < f(x_1, \psi_{j2}, p) \tag{3.59}$$
□

Proof (Proposition 3.8) We prove in the following order: (i) (3.19) 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 (3.19). 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

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

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

$$C_{j_1,i_1}(\psi_{j_1}^\varepsilon, \psi_{i_1}^\varepsilon) > C_{j_2,i_2}(\psi_{j_2}^\varepsilon, \psi_{i_2}^\varepsilon) \tag{3.60}$$

while shifting enough measurement to all other arms such that

$$\min_{i \notin S^*} \min_{j \in S^*} C_{j,i}(\psi_j^\varepsilon, \psi_i^\varepsilon) > \min_{i \notin S^*} \min_{j \in S^*} C_{j,i}(\psi_j^{\frac{1}{2}*}, \psi_i^{\frac{1}{2}*}) \tag{3.61}$$

Hence ψ^ε obtains more evidence on the worst possible pair than $\psi^{\frac{1}{2}*}$ and thereby achieves a better convergence rate (3.16). 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 (3.19) with the exact same value C^*. It follows that there is at least one l s.t. $\psi_l^1 \neq \psi_l^2$. W.l.o.g. assume $l = i_x \notin S^*$ and $\psi_{i_x}^1 > \psi_{i_x}^2$.

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

$$\begin{aligned} C_{j,i_x}(\psi_j^2, \psi_{i_x}^2 + \varepsilon) &= C_{j,i_x}(\psi_j^1, \psi_{i_x}^2 + \varepsilon) && (3.62)\\ &= C_{j,i_x}(\psi_j^1, \psi_{i_x}^1) && (3.63)\\ &= C^* && (3.64)\\ &= C_{j,i_x}(\psi_j^2, \psi_{i_x}^2) && (3.65) \end{aligned}$$

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 independently 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 underallocates, 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:

$$C_{j_x,i_y}(\psi^2_{j_x} - \varepsilon', \psi^2_{i_y} - \varepsilon'') = C_{j_x,i_y}(\psi^1_{j_x}, \psi^1_{i_y}) \tag{3.66}$$
$$= C^* \tag{3.67}$$
$$= C_{j_x,i_y}(\psi^2_{j_x}, \psi^2_{i_y}) \tag{3.68}$$

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

Proof (Proposition 3.9) First, we prove that $\liminf_{n\to\infty} \bar{\psi}_{n,l} \leq \psi^*_l$ by contradiction. Assume otherwise. Then there are n_0 and δ such that for all $n > n_0$: $\bar{\psi}_{n,l} \geq \psi^*_l + \delta$.

$$\sum_{n\in\mathbb{N}} \psi_{n,i} = \sum_{n\in[0,n_0]} \psi_{n,l} + \sum_{n\in[n_0+1,\infty]} \psi_{n,l} \tag{3.69}$$
$$= \sum_{n\in[0,n_0]} \psi_{n,l} + \sum_{n\in[n_0+1,\infty]} \psi_{n,l} \mathbb{I}[\bar{\psi}_{n,l} \geq \psi^*_l + \delta] \tag{3.70}$$

The first term is finite as it consists of finitely many terms that are each finite themselves, in particular bounded from above by 1. The second term is finite by the assumption of the Lemma and the knowledge that ψ is positive. Hence $\sum_{n\in\mathbb{N}} \psi_{n,l} < \infty$.

$$\lim_{n\to\infty} \sum_{l\in[k]} \bar{\psi}_{n,l} = \lim_{n\to\infty} \sum_{l\in[k]} \frac{\sum_{n'\in[n]} \psi_{n',l}}{n} \to 0 \tag{3.71}$$

However, we know by the definition of ψ that for any n, $\sum_{n'\in[n]} \sum_{l\in[k]} \bar{\psi}_{n,l} = n$, which is a contradiction.

Second, we will prove that $\limsup_{n\to\infty} \bar{\psi}_{n,l} \leq \psi_l^*$ by contradiction. Assume otherwise. Combined with the previous point, this implies that for infinitely many n, $\bar{\psi}_{n,l}$ is above and for infinitely many it is below ψ_l^*. As n belongs to a countable set, those points below and above have to alternate. This implies that our indicator function is 'triggered' infinitely many times. Observe that if $\psi_{n,l}$ is 0 once, it will never increase again as no posterior updates will be made. Hence every time the indicator function is triggered, $\psi_{n,l}$ has to be stricly positive; otherwise l would never be sampled again and $\bar{\psi}_{n,l}$ monotonically decrease, even though it oscillates. As a consequence, we have an infinite amount of non-zero, positive values and hence

$$\sum_{n\in\mathbb{N}} \psi_{n,l} \mathbb{I}[\bar{\psi}_{n,l} \geq \psi_l^* + \delta] = \infty$$

which violates the assumption of the lemma and thereby is a contradiction.

With the help of our initial assumption we know that for all l it holds that $\limsup \bar{\psi}_{n,l} \to \psi_l^*$. This, combined with the fact that $\sum_{l\in[k]} \bar{\psi}_{n,l} = \sum_{l\in[k]} \psi_l^*$ allows us to conclude that $\bar{\psi}_n \to \psi^*$. □

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 applying Thompson sampling until two different candidates are at hand.

Section 4.1 introduces the algorithm and provides some explanations on how the algorithm generalizes Russo's TTTS for best arm identification. It also introduces the constraint $\psi_{S^*} = \frac{1}{2}$. Subsequently, Section 4.2 analyzes properties of the algorithm. In particular it states general consequences of finite measurement and presents bounds on the measurement plan. In Section 4.4 we discuss properties of underallocation and overallocation. We conjecture that all of those can be very useful to show that this algorithm's measurement plan converges to the optimal constrained measurement plan $\psi^{\frac{1}{2}*}$. We emphasize that this is an *adaptive* algorithm, in stark contrast to the fixed, optimal allocation $\psi^{\frac{1}{2}*}$. In other words, instead of only ψ influencing Π_n, they now both influence each other. Proofs for those statements are provided in Section 4.5. Additionally, we provide some empirical results in Section 4.3.

4.1 Algorithm

As a generalization of Russo's TTTS, the main difference lies in the fact that candidates are sets. Hence Thompson sampling is repeated until set inequality is reached. However, we can only ever sample individual arms, i.e. not sets. 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 it up into 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, i.e. 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 appear in only either of them are neither clearly optimal nor clearly suboptimal.

As mentioned before, the XOR of S_1 and S_2 will always contain at least two elements. 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 application of the XOR operation to both candidates is essential for the correctness of the algorithm, whereas the uniform sampling from the XOR set of both candidates could possibly be substituted by other distributions. We would expect such a change to preserve correctness, as long as every arm of the XOR set is sampled with strictly positive probability. Yet, it would likely alter the hyperparamter $\beta = \psi_{S^*}$.

We present the approach of sampling the nth arm in Algorithm 4. This approach 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-1} in step n.

$\hat{\theta} \sim \Pi_{n-1}$
$S_1 = \text{top-}m(\hat{\theta})$
repeat
  $\hat{\theta} \sim \Pi_{n-1}$
  $S_2 = \text{top-}m(\hat{\theta})$
until $S_1 \neq S_2$
$I_n \sim \mathcal{U}(S_1 \oplus S_2)$
Play I_n, 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.

4.2 Analysis

Starting off by repeating a general proposition from Russo, we continue by a statement about the implications of finite measurement. Later establish bounds and equalities on the measurement plan ψ of the algorithm.

Proposition 4.1 (Russo: Proposition 4) *Assuming $\theta \in [\underline{\theta}, \bar{\theta}]^k$, for any $l \in [k]$ if $\psi_{n,l} \to \infty$ then for all $\varepsilon > 0$:*

$$\Pi_n(\{\theta \in \Theta | \theta_l \notin (\theta_l^* - \varepsilon, \theta_l^* + \varepsilon)\}) \to 0 \tag{4.1}$$

with probability 1. If $\mathcal{I} = \{l : \sum_{n=1}^{\infty} \psi_{n,l} < \infty\} \neq \emptyset$ then

$$\inf_{n \in \mathbb{N}} \Pi_n(\{\theta \in \Theta | \theta_i \in (\theta_l', \theta_l'') \forall l \in \mathcal{I}\}) > 0 \tag{4.2}$$

for any collections of open intervals $(\theta_l', \theta_l'') \subset (\underline{\theta}, \bar{\theta})$ ranging over $l \in \mathcal{I}$.

Directly using Proposition 4.1, the following lemma formalizes three seemingly intuitive statements. First, it shows that for each arm that is sampled infinitely many times, the estimated mean will converge to its true mean. Second, it demonstrates that if every arm is sampled infinitely often, the posterior will put all of its mass on parameters with S^* as top-m arms. Third, it argues that as long as some arms have only been granted finite measurement, they can't be ruled out from being optimal.

Lemma 4.2 *For $\mathcal{I} = \{l : \sum_{n=1}^{\infty} \psi_{n,l} < \infty\}$ it holds that:*

- $\forall l \notin \mathcal{I}, \forall \varepsilon : \Pi_n(\{\theta : \theta_l \in (\theta_l^* - \varepsilon, \theta_l^* + \varepsilon)\}) \to 1$
- $\mathcal{I} = \emptyset \Rightarrow \alpha_{n,S} \to \begin{cases} 1 & \text{if } S = S^* \\ 0 & \text{if } S \neq S^* \end{cases}$
- $\forall S \subset \mathcal{I} : \liminf_{n \to \infty} \alpha_{n,S} > 0$

Going on to expressing the measurement plan, we observe that a closed-form equality is not easy to obtain. In comparison to Russo's case, it is much harder to construct a closed-form probability of arm l being sampled in a given step as there are many more different scenarios. Vaguely speaking, the XORs can be of very different kinds.

In light of this, we offer both lower and upper bounds on the measurement plan of a single arm. The overall idea is to case-distinguish if l lies set S_1 or S_2. Facing the probability of l being selected given its belonging to

the XOR, we leverage the knowledge of the uniform draw. Hence we have $\Pr[I_n = l | l \in S_1 \oplus S_2] = \frac{1}{|S_1 \oplus S_2|}$. The cardinality of the XOR can be trivially bounded from below by 2 or from above by $2m$. Note that these inequalities are tight with respect to n as m can be seen as a constant.

Proposition 4.3

$$\psi_{n,l} \leq \frac{1}{2}((1-\alpha_{n,l}) \sum_{S: l \in S} \frac{\alpha_{n,S}}{1-\alpha_{n,S}} + \alpha_{n,l} \sum_{S': l \notin S'} \frac{\alpha_{n,S'}}{1-\alpha_{n,S'}}) \tag{4.3}$$

$$\psi_{n,l} \geq \frac{1}{m}((1-\alpha_{n,l}) \sum_{S: l \in S} \frac{\alpha_{n,S}}{1-\alpha_{n,S}} + \alpha_{n,l} \sum_{S': l \notin S'} \frac{\alpha_{n,S'}}{1-\alpha_{n,S'}}) \tag{4.4}$$

$$\tag{4.5}$$

Having established bounds on the measurement plan for individual arms, we proceed to establish two equalities on the measurement plan for sets. In this case we neither propose a closed-form solution nor a bound. Rather, we veil some of the uncertainty with a probability term, which turns out to be sufficiently expressive for some of the later applications.

Proposition 4.4

$$\psi_{n,S} = \frac{1}{2}\alpha_{n,S} + \Pr[I_n \in S | S_1 \neq S](1-\alpha_{n,S}) \tag{4.6}$$

$$\psi_{n,S} = \frac{\alpha_{n,S}}{2} + \frac{\alpha_{n,S}}{2} \sum_{S' \neq S} \frac{\alpha_{n,S'}}{1-\alpha_{n,S'}} + \Pr[S_1, S_2 \neq S \wedge I_n \in S] \tag{4.7}$$

This form of $\psi_{n,S}$ allows us to deduct that the algorithm collects infinite measurement on every arm given infinite samples.

Lemma 4.5

$$\sum_{n \in \mathbb{N}} \psi_{n,S} \to \infty \tag{4.8}$$

Recalling Lemma 4.2, this is particularly interesting as it signals that for $n \to \infty$ we have that $\mathcal{I} = \emptyset$. Therefore $\alpha_{n,S^*} \to 1$ holds true, too. Having established $\psi_{n,S}$ as function of $\alpha_{n,S}$, we continue by analyzing the effect of $\alpha_{n,S} \to 1$ on $\psi_{n,S}$, for general S, followed by the application on S^*.

Lemma 4.6

$$\alpha_{n,S} \to 1 \Rightarrow \psi_{n,S} \to \frac{1}{2} \tag{4.9}$$

This result confirms our hunch about the constraint induced of the algorithm: once it becomes 'sure' of the optimality of a set S, it will sample arms from it with probability $\frac{1}{2}$. As we have previously argued, we have that $\alpha_{n,S^*} \to 1$ and therefore $\psi_{n,S^*} \to \frac{1}{2}$.

Moreover, once we know that $\alpha_{n,S^*} \to 1$, we can simplify the bound from Proposition 4.3.

Lemma 4.7 *If $\alpha_{n,S^*} \to 1$, then for all $i \notin S^*$ and for all $j \in S^*$:*

$$\psi_{n,i} \leq \frac{\alpha_{n,i}}{\max_{S' \neq S^*} \alpha_{n,S'}} \tag{4.10}$$

$$\psi_{n,j} \leq \frac{1 - \alpha_{n,j}}{\max_{S' \neq S^*} \alpha_{n,S'}} \tag{4.11}$$

Lemma 4.8 *If $\bar{\psi}_{n,S^*} \to 1/2$, then there exists a sequence $\varepsilon_n \to 0$ such that $\forall i \notin S^*, \forall n \in \mathbb{N}$:*

$$\psi_{n,i} \leq \exp\{-n(\min_{j \in S^*} C_{j,i}(\bar{\psi}_j, \bar{\psi}_i) - \Gamma^*_{\frac{1}{2}} - \varepsilon_n)\} \tag{4.12}$$

4.3 Empirical behavior

For the following empirical results we assumed the rewards of arms to follow Bernoulli distributions with means $\theta^* = [.1, .2, .3, .4, .5, .6, .7, .8, .9]$. Furthermore, we assumed $m = 4$. We defined priors/posteriors to be Beta distributed with initial parameters $\alpha = \beta = 1$. The Beta distribution is a conjugate prior of the Bernoulli distribution. Its parameters can be interpreted as counts for events of success and failure. Hence the choice $\alpha = \beta = 1$ for each arm mimics a uniform prior. After an observation of the reward of arm l, an update consists of simply increasing either α or β of l, depending on whether it was a success or failure. The mean of the Beta distribution can therefore simply be computed as $\frac{\alpha}{\alpha+\beta}$.

The posterior mass put on certain events, in particular the confidence, was computed as described in Section 4.1. Figure 4.1 indicates the confidence for a given number of steps of our method compared to the the uniform allocation and Thompson sampling. We repeated each of the three methods with 50 different random seeds. Boxes indicate quartiles. We hope to convince the reader that TXTS increases in confidence much quicker than both alternatives.

As we don't have a closed form for the average measurement plan $\bar{\psi}_{n,l}$ at our disposal, we approximate it with the empirical sampling frequencies. Said frequencies simply indicate what fraction of the samples has been allocated to a certain arm.

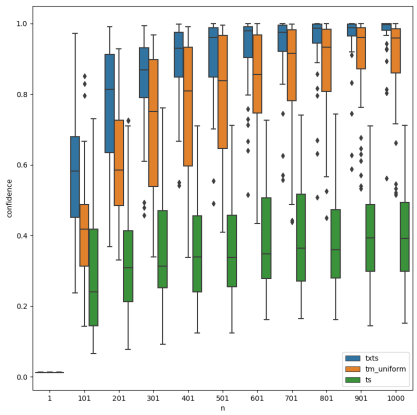

Figure 4.1: Confidence per steps for TXTS, uniform and Thompson sampling.

Figure 4.2 indicates the empirical sampling frequencies. The arms are ordered according to there true means with the worst being at the bottom and the best at the top. Adding up the subotimal, i.e. the 5 lower arms, and optimal, i.e. the 4 upper arms, we note that $\bar{\psi}_{S^*}$ can nicely be read to equal $\frac{1}{2}$ - just as we have argued theoretically. The value of $\frac{1}{2}$ is indicated by a horizontal bar.

Moreover, we compared the empirical sampling frequencies to the fixed optimal allocation from Section 3.2, the uniform allocation and the allocation produced by Thompson sampling. In order to illustrate progression over time, we plotted TXTS and Thompson sampling for different numbers of samples, i.e. timesteps. This can be seen in figure Figure 4.3. We hope to convey that TXTS steadily approaches the optimal allocation with increasing number of samples

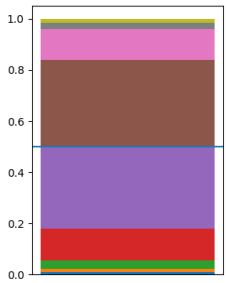

Figure 4.2: TXTS empirical allocation after 1000 steps.

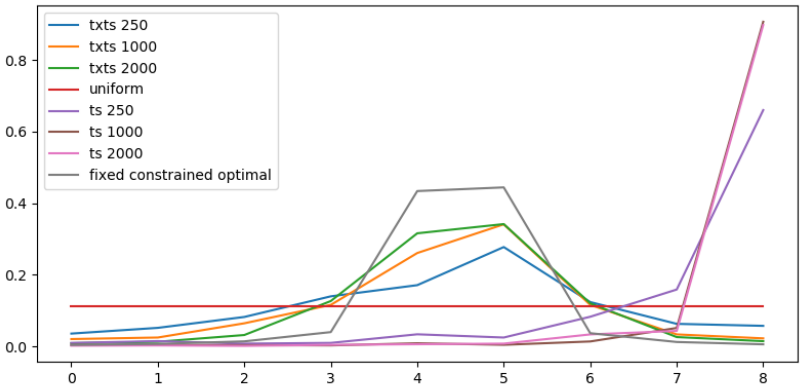

Figure 4.3: Comparison of allocations for different methods and numbers of samples.

As we know from Section 3.1, the optimal allocation gathers equal evidence for all pairs from $S^* \times S^{*c}$. We used the empirical sampling frequencies to approximate those coefficients. A qualitative comparison of the individual coefficients can be found in Figure 4.4. The individual $C_{j,i}$ were computed as described in Section 3.2. Seeing a discrepancy between the coefficients, we wondered whether the algorithm wasn't able to balance the evidence because of its lacking accuracy on θ^* and therefore also computed the coefficients with the estimated $\hat{\theta}$. The results were gathered by executing TXTS with 150 different random seeds, each amounting to 2000 samples. We interpret this result as a confirmation that Z.

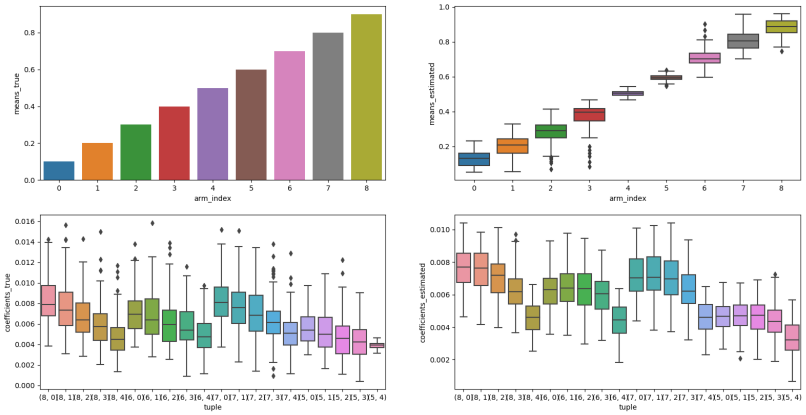

Figure 4.4: Comparison of estimated θ and coefficients $C_{j,i}$.

4.4 Further results and outlook

Looking at Proposition 3.9, we see that in order to show $\bar{\psi}_n \to \psi^{\frac{1}{2}*}$, we can show that once an arm has been oversampled on average, its likelihood of being sampled in the next step is very low. We expect the upper bound from Lemma 4.8 to be useful for this task as a similar approach has succeeded in Russo's case. In particular, the missing link in the chain is to show that if i has been oversampled, $\min_{j \in S^*} C_{j,i}(\bar{\psi}_j, \bar{\psi}_i) \geq \Gamma^*_{\frac{1}{2}}$. If the latter is satisfied, the probability $\psi_{n,i}$ is exponentially small in n and therefore the sum from Proposition 3.9 finite by the geometric series.

However, this crucial condition is not naturally satisfied. Investigating it more carefully, we realize that its satisfaction requires an excess of evidence over the optimal allocation. This can only happen in a given round if the optimal arms $j \in S^*$ have sampling probabilities greater equal to the optical allocation $\psi^{\frac{1}{2}*}$.

As we are not convinced that this holds true, we started looking into the scenario in which the minimizing j is currently undersampled. Our intuition is that if $j \in S^*$ is currently undersampled, it will soon be corrected for.

Lemma 4.9 proposes an analogous form of Proposition 3.9 for undersampling. Lemma 4.10 suggests that if an arm is undersampled, it will be given a lot of measurement effort very soon.

Lemma 4.9 *Given* $\sum_n (\frac{1}{2} - \psi_{n,l}) \mathbb{I}[\bar{\psi}_{n,l} \leq \psi^*_l - \delta] < \infty$ *we have that*

$$\limsup \bar{\psi}_{n,l} \geq \psi^*_j \text{ and } \liminf \bar{\psi}_{n,l} \geq \psi^*_j$$

Lemma 4.10 *If* $\alpha_{n,S^*} \to 1$ *and* $\bar{\psi}_{n,j} < \psi_j^{\frac{1}{2}*}$ *then there exists* $\delta > 0$ *with*

$$\psi_{n,\hat{j}} \geq \frac{1}{2} - m \exp(-n\delta) \tag{4.13}$$

4.5 Proofs

Proof (Lemma 4.2)

- Note that this statement is equal for top-1 and top-m. Hence we can employ Proposition 4.1 telling us that if $\sum_{n \in \mathbb{N}} \psi_{n,l} \to \infty$, it follows that

$$\Pi_n(\{\theta : \theta_l \in (\theta^*_l - \varepsilon, \theta^*_l + \varepsilon)\}) \to 1 \tag{4.14}$$

  Hence (4.14) holds for any $l \notin \mathcal{I}$.

- If $\mathcal{I} = \emptyset$, every arm is sampled infinitely often when the number of samples goes to infinity. This means that (4.14) holds for every arm. In other words, our estimate of each arm is arbitrarily concentrated

around its true value, i.e. $\Pi_n(\{\theta^*\}) \to 1$. Recalling our definitions of S^*, Θ_S and $\alpha_{n,S}$, we see that

$$\alpha_{n,S^*} = \Pi_n(\Theta_{S^*}) \geq \Pi_n(\{\theta^*\}) \to 1 \tag{4.15}$$

As $\alpha_{n,S}$ is bounded by $[0,1]$, we obtain the desired statement.

- Similar to θ_S, we define $\theta_{S,\varepsilon} = \{\theta \in \Theta | \min_{l_1 \in S} \theta_j \geq \max_{l_2 \notin S} \theta_i + \varepsilon\}$. In other words, $\theta_{S,\varepsilon}$ is the set of parameters under which S optimal with a distance of at least ε to the next-best arm. As a lower bound suffices for our statement, we tighten this condition to conveniently allow for usage of the previous points.

  We now pose two tightening conditions on parameters θ: First we restrict arms from $\mathcal{I}^c$ to be at most ε better than the *true* best parameter from $\mathcal{I}^c$. Second we restrict arms from $\mathcal{I}$, other than from S, to be worse than the true best parameter from $\mathcal{I}^c$. For the sake of convenience, let us define $\rho^* = \max_{S \not\subseteq \mathcal{I}} \min_{l_1 \in S} \theta^*_{l_1}$ on the true parameters. For $S \subset \mathcal{I}$, we have:

$$\theta_{S,\varepsilon} = \{\theta | (\min_{l_1 \in S} \theta_{l_1} \geq \max_{l_2 \in \mathcal{I} \setminus S} \theta_{l_1} + \varepsilon) \wedge (\min_{l_1 \in S} \theta_{l_1} \geq \max_{l_2 \notin \mathcal{I}} \theta_{l_2} + \varepsilon)\} \tag{4.16}$$

$$= \{\theta | (\forall l_2 \in \mathcal{I} \setminus S : \min_{l_1 \in S} \geq \theta_{l_2} + \varepsilon) \wedge (\min_{l_1 \in S} \theta_{l_1} \geq \max_{l_2 \notin \mathcal{I}} \theta_{l_2} + \varepsilon)\} \tag{4.17}$$

$$\supseteq \{\theta | (\min_{l_1 \in S} \theta_{l_1} \geq \rho^* + 2\varepsilon) \wedge (\forall l_2 \in \mathcal{I} \setminus S : \rho^* > \theta_{l_2}) \wedge (\min_{l_1 \in S} \theta_{l_1} \geq \max_{l_2 \notin \mathcal{I}} \theta_{l_2} + \varepsilon)\} \tag{4.18}$$

$$\supseteq \{\theta | (\min_{l_1 \in S} \theta_{l_1} \geq \rho^* + 2\varepsilon) \wedge (\forall l_2 \in \mathcal{I} \setminus S : \rho^* > \theta_{l_2}) \wedge (\rho^* + \varepsilon \geq \max_{l_2 \notin \mathcal{I}} \theta_{l_2})\} \tag{4.19}$$

$$= \underbrace{\{\theta | (\min_{l_1 \in S} \theta_{l_1} \geq \rho^* + 2\varepsilon) \wedge (\forall l_2 \in \mathcal{I} \setminus S : \rho^* > \theta_{l_2})\}}_{\text{A}} \setminus \underbrace{\{\theta | \max_{l_2 \notin \mathcal{I}} \theta_{l_2} > \rho^* + \varepsilon\}}_{\text{B}} \tag{4.20}$$

Where the last step follows from the fact that $\{x | p(x) \wedge q(x)\} = \{x | p(x)\} \setminus \{x | \neg q(x)\}$. By (4.14), we know that $\Pi_n(B) \to 0$. The second part of Proposition 4.1 tells us that over any open interval (θ'_l, θ''_l), we have $\inf_{n \in \mathbb{N}} \Pi_n(\{\theta | \theta_l \in (\theta'_l, \theta''_l)\ \forall l \in \mathcal{I}\}) > 0$. By defining (θ'_l, θ''_l) to be $(\rho + 2\varepsilon, \bar{\theta})$ for all arms from $S \subset \mathcal{I}$ and $(\underline{\theta}, \rho^* + \varepsilon)$ for all arms in $\mathcal{I} \setminus S$, we get $\inf_{n \in \mathbb{N}} \Pi_n(A) > 0$. Together, this gives us

$$\inf_{n \in \mathbb{N}} \alpha_{n,S} = \inf_{n \in \mathbb{N}} \Pi_n(\theta_{S,\varepsilon}) \geq \inf_{n \in \mathbb{N}} (\Pi_n(A) - \Pi_n(B)) > 0 \tag{4.21}$$

$$\liminf_{n \in \mathbb{N}} \alpha_{n,S} \geq \inf_{n \in \mathbb{N}} \alpha_{n,S} > 0 \tag{4.22}$$

□

Proof (Proposition 4.3) It will prove itself useful to first investigate the probability of a given arm l belonging to either S_1 or S_2, yet not both.

$$\begin{aligned}
&\Pr[l \in S_1 \wedge l \notin S_2] + \Pr[l \notin S_1 \wedge l \in S_2] && (4.23)\\
=&\sum_{S:l\in S} \Pr[S_1 = S] \sum_{S':l\notin S'} \Pr[S_2 = S'|S_1 = S] + \\
&\sum_{S':l\notin S'} \Pr[S_1 = S'] \sum_{S:l\in S} \Pr[S_2 = S|S_1 = S'] && (4.24)\\
=&\sum_{S:l\in S} \alpha_{n,S} \sum_{S':l\notin S'} \frac{\alpha_{n,S'}}{1-\alpha_{n,S}} + \sum_{S':l\notin S'} \alpha_{n,S'} \sum_{S:l\in S} \frac{\alpha_{n,S}}{1-\alpha_{n,S'}} && (4.25)\\
=&\sum_{S:l\in S} \frac{\alpha_{n,S}}{1-\alpha_{n,S}} \sum_{S':l\notin S'} \alpha_{n,S'} + \sum_{S':l\notin S'} \frac{\alpha_{n,S'}}{1-\alpha_{n,S'}} \sum_{S:l\in S} \alpha_{n,S} && (4.26)\\
=&\sum_{S:l\in S} \frac{\alpha_{n,S}}{1-\alpha_{n,S}} (1-\alpha_{n,l}) + \sum_{S':l\notin S'} \frac{\alpha_{n,S'}}{1-\alpha_{n,S'}} \alpha_{n,l} && (4.27)\\
=&(1-\alpha_{n,l}) \sum_{S:l\in S} \frac{\alpha_{n,S}}{1-\alpha_{n,S}} + \alpha_{n,l} \sum_{S':l\notin S'} \frac{\alpha_{n,S'}}{1-\alpha_{n,S'}} && (4.28)
\end{aligned}$$

This identity can now be leveraged for both lower and upper bound. But first, let's express the the measurement plan exactly.

$$\begin{aligned}
\psi_{n,l} =& \Pr[l \in S_1 \wedge l \notin S_2 \wedge I_n = l] + \Pr[l \notin S_1 \wedge l \in S_2 \wedge I_n = l] && (4.29)\\
=& \Pr[I_n = l|l \in S_1 \wedge l \notin S_2] \Pr[l \in S_1 \wedge l \notin S_2] + \\
& \Pr[I_n = l|l \notin S_1 \wedge l \in S_2] \Pr[l \notin S_1 \wedge l \in S_2] && (4.30)
\end{aligned}$$

Observe that both terms $\Pr[I_n = l|l \in S_1 \wedge l \notin S_2]$ and $\Pr[I_n = l|l \notin S_1 \wedge l \in S_2]$ correspond to a very similar situation: we know that l is part of the XOR, but we don't know how exactly the rest of the XOR looks like. To those quantities we can apply the aforementioned naïve bounds of $\frac{1}{2}$ and $\frac{1}{2m}$.

$$\begin{aligned}
\psi_{n,l} \leq& \frac{1}{2}(\Pr[l \in S_1 \wedge l \notin S_2] + \Pr[l \notin S_1 \wedge l \in S_2]) && (4.31)\\
=& \frac{1}{2}((1-\alpha_{n,l}) \sum_{S:l\in S} \frac{\alpha_{n,S}}{1-\alpha_{n,S}} + \alpha_{n,l} \sum_{S':l\notin S'} \frac{\alpha_{n,S'}}{1-\alpha_{n,S'}}) && (4.32)\\
\psi_{n,l} \geq& \frac{1}{2m}(\Pr[l \in S_1 \wedge l \notin S_2] + \Pr[l \notin S_1 \wedge l \in S_2]) && (4.33)\\
=& \frac{1}{2m}((1-\alpha_{n,l}) \sum_{S:l\in S} \frac{\alpha_{n,S}}{1-\alpha_{n,S}} + \alpha_{n,l} \sum_{S':l\notin S'} \frac{\alpha_{n,S'}}{1-\alpha_{n,S'}}) && (4.34)
\end{aligned}$$

□

Proof (Proposition 4.4) In order to show the first equality, we rely on the

idea that we only check if the first sampled set S_1 is equal to S.

$$\begin{aligned}
\psi_{n,S} &= \Pr[I_n \in S] && (4.35)\\
&= \Pr[S_1 = S \wedge I_n \in S_1] + \Pr[S_1 \neq S \wedge I_n \in S] && (4.36)\\
&= \Pr[I_n \in S_1 | S_1 = S] \Pr[S_1 = S] + \Pr[I_n \in S | S_1 \neq S] \Pr[S_1 \neq S] && (4.37)\\
&= \Pr[I_n \in S_1] \Pr[S_1 = S] + \Pr[I_n \in S | S_1 \neq S](1 - \Pr[S_1 = S]) && (4.38)\\
&= \frac{1}{2}\alpha_{n,S} + \Pr[I_n \in S | S_1 \neq S](1 - \alpha_{n,S}) && (4.39)
\end{aligned}$$

Note that $\Pr[I_n \in S_1] = \Pr[I_n \in S_2] = \frac{1}{2}$ as the XOR operation ensures that equally many elements from S_1 and S_2 in the XOR set. In combination with uniform sampling this yields aforementioned relation.

In order to show the second equality, we go a step further by checking if either of both sets equals S. For that purpose we first look into the probability that S is sampled as a second set and that I_n stems from S.

$$\begin{aligned}
&\Pr[S_2 = S \wedge I_n \in S_2] && (4.40)\\
&= \sum_{S' \neq S} \Pr[S_1 = S' \wedge S_2 = S \wedge I_n \in S_2] && (4.41)\\
&= \sum_{S' \neq S} \Pr[S_1 = S'] \Pr[S_2 = S | S_1 = S'] \Pr[I_n \in S_2 | S_1 = S' \wedge] && (4.42)\\
&= \sum_{S' \neq S} \alpha_{n,S'} \frac{\alpha_{n,S}}{1 - \alpha_{n,S'}} \frac{1}{2} = \frac{\alpha_{n,S}}{2} \sum_{S' \neq S} \frac{\alpha_{n,S'}}{1 - \alpha_{n,S'}} && (4.43)
\end{aligned}$$

$$\begin{aligned}
\psi_{n,S} &= \Pr[I_n \in S] && (4.44)\\
&= \Pr[S_1 = S \wedge I_n \in S_1] + \Pr[S_2 = S \wedge I_n \in S_2] + \Pr[S_1, S_2 \neq S \wedge I_n \in S] && (4.45)\\
&= \frac{\alpha_{n,S}}{2} + \frac{\alpha_{n,S}}{2} \sum_{S' \neq S} \frac{\alpha_{n,S'}}{1 - \alpha_{n,S'}} + \Pr[S_1, S_2 \neq S \wedge I_n \in S] && (4.46)
\end{aligned}$$

□

Proof (Lemma 4.5) Proposition 4.4 tells us that $\psi_{n,S} > \gamma \alpha_{n,S}$ for some constant $\gamma > 0$. Thanks to this we have $\sum_{n \in \mathbb{N}} \psi_{n,S} > \frac{1}{2} \sum_{n \in \mathbb{N}} \alpha_{n,S}$. For the sake of contradiction, assume that $\exists S'$ with $\sum_{n \in \mathbb{N}} \psi_{n,S'} < \infty$. According to Lemma 4.2's definition, we have $S \subset \mathcal{I}$. Hence we can apply its third clause and get $\liminf_{n \to \infty} \alpha_{n,S'} > 0$. It follows directly that $\sum_{n \in \mathbb{N}} \alpha_{n,S} \to \infty$. As a consequence $\sum_{n \in \mathbb{N}} \psi_{n,S}$ tends to infinity with growing n as well, which is a contradiction. □

Proof (Lemma 4.6) In Proposition 4.4's first form, $\Pr[I_n \in S | S_1 \neq S]$ is naturally bounded by 1. The desired statement follows immediately. □

Proof (Lemma 4.7) Proposition 4.3 tells us that

$$\psi_{n,l} \leq \frac{1}{2}((1-\alpha_{n,l}) \sum_{S: l \in S} \frac{\alpha_{n,S}}{1-\alpha_{n,S}} + \alpha_{n,l} \sum_{S': l \notin S'} \frac{\alpha_{n,S'}}{1-\alpha_{n,S'}})$$

Knowing that $\alpha_{n,S^*} \to 1$, we can bound $\psi_{n,i}$ in the following way:

$$\begin{aligned}
\psi_{n,i} \leq & \frac{1}{2}((1-\alpha_{n,i}) \frac{\sum_{S: i \in S} \alpha_{n,S}}{1-\alpha_{n,S^*}} + \alpha_{n,i} \frac{\sum_{S': i \notin S'} \alpha_{n,S'}}{1-\alpha_{n,S^*}}) && (4.47)\\
= & \frac{1}{2}((1-\alpha_{n,i}) \frac{\alpha_{n,i}}{1-\alpha_{n,S^*}} + \alpha_{n,i} \frac{1-\alpha_{n,i}}{1-\alpha_{n,S^*}}) && (4.48)\\
\leq & \frac{1}{2} \frac{2\alpha_{n,i}(1-\alpha_{n,i})}{\max_{S' \neq S^*} \alpha_{n,S'}} && (4.49)\\
\leq & \frac{\alpha_{n,i}}{\max_{S' \neq S^*} \alpha_{n,S'}} && (4.50)
\end{aligned}$$

for $i \notin S^*$ as well as

$$\psi_{n,j} \leq \frac{1-\alpha_{n,j}}{\max_{S' \neq S^*} \alpha_{n,S'}} \tag{4.51}$$

for $j \in S^*$.

Note that in (4.47) we used that for any $S \neq S^*$, $\alpha_{n,S} < \alpha_{n,S^*}$ thanks to our assumption $\alpha_{n,S^*} \to 1$. □

Proof (Lemma 4.8) We first want to find a lower bound for the denominator and then find an upper bound for the numerator.

Proposition 3.8 gives us that

$$\lim_{n \to \infty} \sup -\frac{1}{n} \log \Pi_n(\Theta^c_{S^*}) \leq \Gamma^*_{\frac{1}{2}}$$

Lemma A.1 tells us that there is a sequence $\varepsilon_n > 0, \varepsilon_n \to 0$ s.t.

$$\begin{aligned}
-\frac{1}{n} \log \Pi_n(\Theta^c_{S^*}) \leq & \Gamma^*_{\frac{1}{2}} + \varepsilon_n && (4.52)\\
\Pi_n(\Theta^c_{S^*}) \geq & \exp\{-n(\Gamma^*_{\frac{1}{2}} + \varepsilon_n)\} && (4.53)
\end{aligned}$$

We observe that $\Theta^c_{S^*} = \bigcup_{S' \neq S^*} \Theta_{S'}$. We use the union bound to notice that $\max_{S' \neq S^*} \alpha_{n,S'} \leq \Pi(\Theta^c_{S^*}) \leq \binom{k}{m} \max_{S' \neq S^*} \alpha_{n,S'}$. We have

$$\lim_{n \to \infty} \frac{1}{n} \log \frac{\alpha_{n,S}}{\binom{k}{m} \alpha_{n,S}} \to 0 \Rightarrow \alpha_{n,S} \doteq \binom{k}{m} \alpha_{n,S}$$

and hence by the Squeeze theorem $\max_{S' \neq S^*} \alpha_{n,S'} \doteq \Pi(\Theta^c_{S^*})$. Combining these two insights, we obtain our desired lower bound for the denominator:

$$\max_{S' \neq S^*} \alpha_{n,S'} \doteq \Pi_n(\Theta^c_{S^*}) \geq \exp\{-n(\Gamma^*_{\frac{1}{2}} + \varepsilon_n)\}$$

Let's investigate the numerator.

For $i \notin S^*$, we have $\Theta_i \subset \bar{\Theta}_i$ and hence:

$$\begin{aligned}
\alpha_{n,i} =& \Pi_n(\Theta_i) && (4.54)\\
\leq& \Pi_n(\bar{\Theta}_i) && (4.55)\\
\doteq& \exp\{-n \inf_{\theta \in \bar{\Theta}_i} D_{\bar{\psi}_n}(\theta^* || \theta)\} \text{ (Proposition 3.1)} && (4.56)\\
=& \exp\{-n \min_{j \in S^*} C_{j,i}(\bar{\psi}_j, \bar{\psi}_i)\} \text{ (Lemma 3.4)} && (4.57)
\end{aligned}$$

Simultaneously leveraging both bounds we obtain:

$$\begin{aligned}
\frac{\alpha_{n,i}}{\max_{S' \neq S^*} \alpha_{n,S'}} &\leq \frac{\exp\{-n \min_{j \in S^*} C_{j,i}(\bar{\psi}_j, \bar{\psi}_i)\}}{\exp\{-n(\Gamma^*_{\frac{1}{2}} + \varepsilon_n)\}} && (4.58)\\
&= \exp\{-n(\min_{j \in S^*} C_{j,i}(\bar{\psi}_j, \bar{\psi}_i) - \Gamma^*_{\frac{1}{2}} - \varepsilon_n)\} && (4.59)
\end{aligned}$$

□

Remark 4.11 *As C is strictly increasing in its second argument and by the assumptions of the lemma $\bar{\psi}_i > \psi^*_i + \delta$, we have that for any $j \in S^*$ $C_{j,i}(\psi^*_j, \bar{\psi}_i) = C_{j,i}(\psi^*_j, \psi^*_i) + \delta' = \Gamma^* + \delta'$ with $\delta' > 0$. It remains to show that a similar equality holds when minimizing over j and with $\bar{\psi}_j$ instead of ψ^* as first argument. One should be able to leverage $\bar{\psi}_{S^*} \to \frac{1}{2} = \psi^*_{S^*}$ along the way.*

Proof (Lemma 4.9) We first show the first statement and then use it for the second. For the sake of contradiction, assume that $\limsup \bar{\psi}_{n,l} < \psi^*_l$. This implies that there is a n_0 from which onward $\bar{psi}_{n,l} < \psi^*_l$.

$$\begin{aligned}
\sum_n (\frac{1}{2} - \psi_{n,l}) &= \sum_{n=1}^{n_0} (\frac{1}{2} - \psi_{n,l}) + \sum_{n > n_0} (\frac{1}{2} - \psi_{n,l}) && (4.60)\\
&= \sum_{n=1}^{n_0} (\frac{1}{2} - \psi_j) + \sum_{n > n_0} (\frac{1}{2} - \psi_{n,l}) \mathbb{I}[\bar{\psi}_{n,l} \leq \psi^*_l - \delta] && (4.61)\\
&= C && (4.62)
\end{aligned}$$

Where the last line follows from the assumption and the fact that the first sum is a finite amount of individually finite quantities. Hence we can write:

$$\begin{aligned}
\sum_n \psi_{n,l} &= -C + \sum_n \frac{1}{2} && (4.63)\\
\bar{\psi}_{n,l} &= \frac{-C}{n} + \frac{1}{2} \sum_n \frac{1}{n} && (4.64)\\
&= \frac{-C}{n} + \frac{1}{2} H_n && (4.65)
\end{aligned}$$

We know that H_n is asymptotically equivalent to $\log(n)$. Therefore we arrive at $\bar{\psi}_{n,l} \to \infty$, which is a contradiction. As a consequence we have $\limsup \bar{\psi}_{n,l} \geq \psi_j^*$.

The argument from Proposition 3.9 can be mirrored to show the second statement. □

Proof (Lemma 4.10) Thanks to $\alpha_{n,S^*} \to 1$: we can apply Lemma 4.7, and get for $j \in S^*$:

$$\psi_{n,j} \leq \frac{1-\alpha_{n,j}}{1-\alpha_{n,S^*}} \tag{4.66}$$

Additionally, according to Lemma 4.6, $\alpha_{n,S^*} \to 1$, allows us to assume that $\psi_{n,S^*} \to \frac{1}{2}$.

Assume for $\hat{j} \in S^*$ it holds that $\bar{\psi}_{\hat{j},n} < \psi_{\hat{j}}^{\frac{1}{2}*}$

$$\begin{aligned}
\psi_{n,\hat{j}} &= \psi_{n,S^*} - \sum_{j \in S^* \setminus \{\hat{j}\}} \psi_{n,j} && (4.67)\\
&\to \frac{1}{2} - \sum_{j \in S^* \setminus \{\hat{j}\}} \psi_{n,j} && (4.68)\\
&\geq \frac{1}{2} - \sum_{j \in S^* \setminus \{\hat{j}\}} \frac{1-\alpha_{n,j}}{1-\alpha_{n,S^*}} \text{ (Lemma 4.7)} && (4.69)\\
&= \frac{1}{2} - \sum_{j \in S^* \setminus \{\hat{j}\}} \frac{\Pi_n(\Theta^c_{m,j})}{\Pi_n(\Theta^c_{S^*})} && (4.70)\\
&= \frac{1}{2} - \sum_{j \in S^* \setminus \{\hat{j}\}} \frac{\exp\{-n \min_{i \notin S^*} C_{j,i}(\bar{\psi}_j, \bar{\psi}_i)\}}{\exp\{-n \min_{i \notin S^*} \min_{j \in S^*} C_{j,i}(\bar{\psi}_j, \bar{\psi}_i)\}} && (4.71)\\
&= \frac{1}{2} - \sum_{j \in S^* \setminus \{\hat{j}\}} \exp\{-n(\min_{i \notin S^*} C_{j,i}(\bar{\psi}_j, \bar{\psi}_i) - \min_{i \notin S^*} \min_{j \in S^*} C_{j,i}(\bar{\psi}_j, \bar{\psi}_i))\} && (4.72)\\
&= \frac{1}{2} - m \exp(-n\delta) && (4.73)
\end{aligned}$$

Where Equation (4.71) relies on a combination of Proposition 3.1 and Lemma 3.4, just as we've done in Lemma 4.8. The last step stems from the fact that a minimization over both indices will always yield a value lesser or equal to the minimization over only one of the indices. □

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 discrete 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 (3.6) $x \in \mathbb{R}$. 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 one-dimensional exponential family

We reiterate some of the facts mentioned by Russo and add some general properties.

- They have a scalar parameter θ.
- They come with the definition of function $T(x)$, $\nu(\theta)$, $h(x)$ and $A(\theta)$. T corresponds to the sufficient statistic and A to the log-partition function.
- The probability density function is defined by:

$$f_X(x|\theta) = h(x)\exp(\nu(\theta)T(x) - A(\theta))$$

- Their log-partition-function $A(\theta)$ is strictly convex and differentiable.
- They have conjugate priors.
- Their mean equals $A'(\theta) = \int T(y)p(y|\theta)d\nu(y)$. In the typical case of $T(y) = y$, this yields that $A'(\theta) = \mu(\theta)$.
- 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] D. Agarwal, B. Chen, and P. Elango. Explore/exploit schemes for web content optimization. In *2009 Ninth IEEE International Conference on Data Mining*, pages 1–10, Dec 2009.

[2] Christopher Baldassano and Naomi Leonard. Explore vs. exploit: Task allocation for multi-robot foraging. 11 2019.

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

[4] Vincent François-Lavet, Peter Henderson, Riashat Islam, Marc G. Bellemare, and Joelle Pineau. An introduction to deep reinforcement learning. *CoRR*, abs/1811.12560, 2018.

[5] 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.

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

[7] James McInerney, Benjamin Lacker, Samantha Hansen, Karl Higley, Hugues Bouchard, Alois Gruson, and Rishabh Mehrotra. Explore, exploit, and explain: Personalizing explainable recommendations with bandits. In *Proceedings of the 12th ACM Conference on Recommender Systems*, RecSys '18, pages 31–39, New York, NY, USA, 2018. ACM.

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

[9] Csaba Szepesvari. *Algorithms for Reinforcement Learning*. Morgan and Claypool Publishers, 2010.

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.