



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 Thompson Sampling	4
2.4 Notation	4
2.5 Best Arm Identification	5
2.5.1 Model	5
2.5.2 Optimal allocation	6
2.5.3 A Constrained Optimal Allocation	7
2.5.4 Top-Two Thompson Sampling algorithm	8
2.5.5 Alternative approaches	9
2.6 Top- m Arm Identification	9
2.6.1 Current approaches: LUCB	9
3 Characterizing the Constrained Optimal Top-m Allocation	11
3.1 Statements	12
3.2 A Concrete Optimal Allocation	15
3.3 Proofs	16
4 Algorithm	21
4.1 Analysis/Properties	21
4.2 Empirical behaviour	21
4.3 Proofs	21
5 Conclusion	23
A Appendix	25

CONTENTS

A.1 Computing $C_{j,i}$ for Bernoulli means	25
A.2 Facts about the exponential family	26
Bibliography	27

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 constraints 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 bandits 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. It is typically assumed that the outcome of the selection of an individual arm 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 gathered knowledge.

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, where quality is usually assessed by high means. Hence it is part of the task to make estimated best and true best equal as well as obtaining a high *confidence* on this statement.

The bandits model can be used for all processes involving sequential decision making. Yet, an important simplification is the assumption of the distributions being fixed over time. This assumption might be more or less representative of the true, to-be-modelled processes.

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 Thompson Sampling

Thompson sampling is a sampling strategy often employed for choosing arms in the bandits scenario. By default, it particularly lends itself to the cumulative reward setting. The technique is based on having and updating prior distributions over all arms. In contrast to greedily selecting the arm that empirically maximizes the metric of desire, Thompson sampling suggests to randomly draw beliefs on arms from the prior distributions. 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. Intuitively, this is very appealing for the cumulative reward scenario as it creates a natural balance between exploration and exploitation.

2.4 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:

$$\Theta := [0, 1]^k \quad (2.1)$$

$$\Theta_{1,l} := \{\theta \in \Theta \mid l = \arg \max_{l' \in [k]} \theta_{l'}\} \quad (2.2)$$

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

$$= \{\theta \in \Theta \mid \min_{j_1 \in S} \theta_{j_1} > \max_{j_2 \notin S} \theta_{j_2}\} \quad (2.4)$$

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

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:

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

$$= \Pr[S \text{ is top-}m \mid D_n] \quad (2.7)$$

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

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

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

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 $\sum_{l \in [k]} \psi_l = 1$.

The n -th sampled arm is denoted as l_n . For adaptive strategies, the allocation can change after every sample. After n sample we refer to the allocation as ψ_n and therefore after n samples we have $\psi_{n,l} = \Pr[l_n = l]$ for arm l . We also define the allocation property of a set by the sum of its arms: $\psi_S = \sum_{l \in S} \psi_l$. In the case of adaptive strategies, we might as well be interested in average allocation up to a certain sample n . We write:

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

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

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 also known as '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 fix amount of samples or minimize amount of samples for a fixed confidence.

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 $\Pr[\text{output of algorithm is } \varepsilon\text{-correct}] \geq 1 - \delta$, 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 e.g. aerodynamic properties of the cars. Elaborate simulations can come with significant resource demands, such as compute power and direct and indirect costs of losing time. Coming to the same conclusion, consisting of the superiority of one of the designs, with fewer simulations can be crucial.
- 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 best possible bound. Intuitively it is not possible to match the performance of that optimal allocation without the knowledge of the underlying truth. Hence the desired statement is to say that a concrete 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 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 most the arms in a way that convinces 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:

$$\Pi_n(\Theta_{l^*}) = 1 - \Pi_n(\Theta_{l^*}^c) = 1 - \exp\{-n\Gamma^*\} \quad (2.11)$$

$$\Gamma^* = \max_{\psi} \min_{\theta \in \Theta_{l^*}^c} \sum_{l \in [k]} \psi_l d(\theta_l^* || \theta_l) \quad (2.12)$$

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 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

$$C_i(\psi_{l^*}, \psi_i) := \min_{x \in \mathbb{R}} \psi_{l^*} d(\theta_{l^*}^* || x) + \psi_i d(\theta_i^* || x) \quad (2.13)$$

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

$$\forall i_1, i_2 \neq l^* : C_{i_1}(\psi_{l^*}, \psi_{i_1}) = C_{i_2}(\psi_{l^*}, \psi_{i_2}) \quad (2.14)$$

Moreover, he shows that the optimal allocation is unique.

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:

$$\Gamma_\beta^* = \max_{\psi: \psi_{l^*} = \beta} \min_{\theta \in \Theta_{l^*}^c} \sum_{l \in [k]} \psi_l d(\theta_l^* || \theta_l) \quad (2.15)$$

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 1 is to repeat drawing with Thompson sampling from the prior until two different candidates are obtained. 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.

Algorithm 1 Given a prior Π_n at step n

```

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

```

The attentive reader might wonder how exactly prior/posteriors are updated. This is addressed in the empirical section 4.2.

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 similiary to those mentioned in 2.5.1. Natural explications for desiring the identification m instead of 1 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 algorithhm can be seen in 2. 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 2 Given a prior Π_n at 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 $\mu < \varepsilon$

2. BACKGROUND

Confidence estimation 1

Confidence estimation 2

Confidence estimation 3

Chapter 3

Characterizing 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 $\Pi_n(\Theta_{m,S^*})$. 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. Proofs of the statements are provided in Section 3.3. Moreover, we will show an example of a concrete optimal allocation in Section 3.2.

Overall we hope to convince the reader that the chracterization 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 that is the hardest to distinguish from it, we will compare the a pair of an optimal and a suboptimal arm are hardest to distinguish from another. Note that again, distinction revolves around two factors: the frequency with which an arm has been sampled, i.e. quantity, as well as the proximity of its true mean, i.e. quality.

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

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

3.1 Statements

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

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

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

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

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

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

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

Lemma 3.2

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

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

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

$$\Pi_n(\Theta_{m,S^*}^c) \doteq \max_{i \notin S^*} \Pi_n(\tilde{\Theta}_i) \doteq 1 - \min_{j \in S^*} \Pi_n(\Theta_j) \quad (3.3)$$

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

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

In particular, $C_{j,i}$ can be thought of as evidence that j is distinct from i . Quite naturally, we want this evidence 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 ,

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

Intuitively, this lemma tells us that the weighted sum of KL divergences can be reduced to looking at only two arms for parameters in $\bar{\Theta}_i$. 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 is from the true mean of i , captured by the KL divergence between both arms and an x inbetween, 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:

$$\Pi_n(\Theta_{S^*}^c) \doteq \max_{i \notin S^*} \Pi_n(\hat{\Theta}_i) \quad (\text{Lemma 3.3}) \quad (3.6)$$

$$\doteq \max_{i \notin S^*} \exp\left\{-n \inf_{\theta \in \bar{\Theta}_i} \sum_{l \in [k]} \bar{\psi}_{n,l} d(\theta_l^* || \theta_l)\right\} \quad (\text{Proposition 3.1}) \quad (3.7)$$

$$= \max_{i \notin S^*} \exp\left\{-n \min_{j \in S^*} C_{j,i}(\psi_j, \psi_i)\right\} \quad (\text{Lemma 3.4}) \quad (3.8)$$

$$= \exp\left\{-n \min_{i \notin S^*} \min_{j \in S^*} C_{j,i}(\psi_j, \psi_i)\right\} \quad (3.9)$$

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 (3.9) as small as possible. We have:

$$\psi^* = \arg \min_{\psi} \exp\left\{-n \min_{i \notin S^*} \min_{j \in S^*} C_{j,i}(\psi_j, \psi_i)\right\} \quad (3.10)$$

$$= \arg \max_{\psi} \min_{i \notin S^*} \min_{j \in S^*} C_{j,i}(\psi_j, \psi_i) \quad (3.11)$$

For the sake of convenience we define the optimal exponent:

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

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:

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

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

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 3.7 we will show that the optimal measurement plan will fulfill (3.15). Before doing so, we set the stage by enunciating 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

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

Proposition 3.7 *The solution to the optimization problem 3.13 is the allocation $\psi^{\frac{1}{2}*}$, which is unique and satisfies*

$$\forall j_1, j_2 \in S^*, \forall i_1, i_2 \notin S^* : C_{j_1, i_1}(\psi_{j_1}^*, \psi_{i_1}^*) = C_{j_2, i_2}(\psi_{j_2}^*, \psi_{i_2}^*) \quad (3.15)$$

If $\psi_n = \psi^*$ for all n , then

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

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

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

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 ψ^* 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 respectively. Thanks to this assumption, the KL divergences can be minimized analytically with great convenience, we refer to the appendix, Appendix A.1 for details. Figure 3.1 indicates the probability mass put onto each arm for each scenario. The results were produced with 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 we refer to the simulation code in our repository ¹.

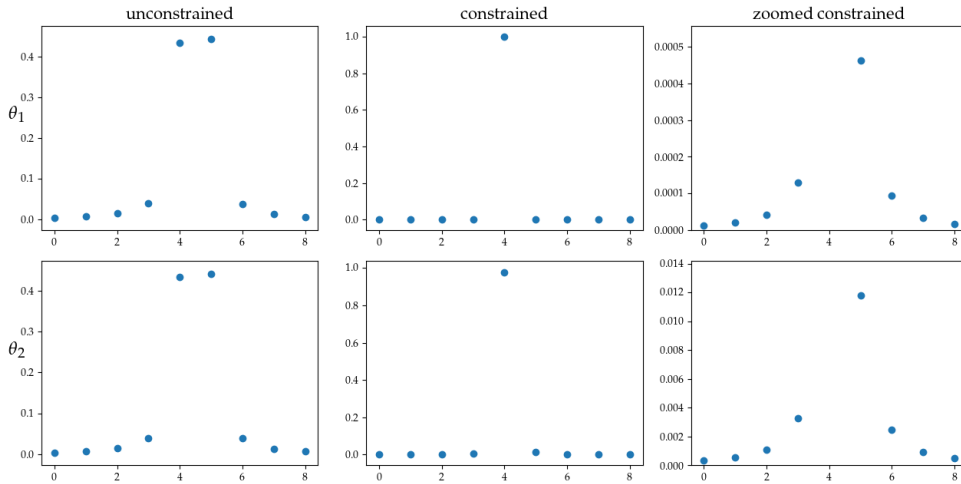


Figure 3.1: Unconstrained and constrained optimal allocation for θ_1 and θ_2 , top-4

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

¹https://github.com/kkleindev/tts/compute_optimal_allocation.py

3.3 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 Θ_j .

$$\Theta_{m,S^*}^c = \{\theta \in \Theta \mid \min_{j \in S^*} \theta_j > \max_{i \notin S^*} \theta_i\}^c \quad (3.16)$$

$$= \{\theta \in \Theta \mid \max_{i \notin S^*} \theta_i \geq \min_{j \in S^*} \theta_j\} \quad (3.17)$$

$$= \{\theta \in \Theta \mid \exists i \notin S^* : \theta_i \geq \min_{j \in S^*} \theta_j\} \quad (3.18)$$

$$= \{\theta \in \Theta \mid \exists i \notin S^* : \text{top-}m(\theta, S^* \cup \{i\}) \neq S^*\} \quad (3.19)$$

$$= \bigcup_{i \notin S^*} \{\theta \in \Theta \mid \text{top-}m(\theta, S^* \cup \{i\}) \neq S^*\} \quad (3.20)$$

$$= \bigcup_{i \notin S^*} \bar{\Theta}_i \quad (3.21)$$

$$\Theta_{m,S^*} = \{\theta \in \Theta \mid \min_{j \in S^*} \theta_j > \max_{j \in S^*} \theta_j\} \quad (3.22)$$

$$= \{\theta \in \Theta \mid \bigwedge_{j \in S^*} \theta_j > \max_{j \in S^*} \theta_j\} \quad (3.23)$$

$$= \bigcap_{j \in S^*} \{\theta \in \Theta \mid \theta_j > \max_{j \in S^*} \theta_j\} \quad (3.24)$$

$$= \bigcap_{j \in S^*} \{\theta \in \Theta \mid j \in \text{top-}m(\theta, [k])\} \quad (3.25)$$

$$= \bigcap_{j \in S^*} \Theta_j \quad (3.26)$$

$$\Theta_{m,S^*}^c = \Theta - \Theta_{m,S^*} \quad (3.27)$$

$$\Theta_{m,S^*}^c = \Theta - \bigcap_{j \in S^*} \Theta_j \quad (3.28)$$

□

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 i s and each single one leads to a set, whose density is bounded by the maximal density of all such sets. This gives us:

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

We use the definition of \doteq to check whether lower bound and upper bound are equal.

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

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

$$\Pi_n(\Theta_{m,S^*}^c) \doteq \max_{i \notin S^*} \Pi_n(\tilde{\Theta}_i) \quad (3.31)$$

For $j \in S^*$, we follow a very similar 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.

$$1 - \min_{j \in S^*} \Pi(\Theta_j) \leq \Pi(\Theta_{m,S^*}^c) \leq 1 - \min_{j \in S^*} (\Pi(\Theta_j))^m \quad (3.32)$$

$$(3.33)$$

Again, we will check whether upper and lower bound equal in the \doteq sense.

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

Observe that $1 \geq \min_{j \in S^*} \Pi(\Theta_j) \geq \alpha_{n,S^*} \rightarrow 1$. 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

$$\Pi_n(\Theta_{m,S^*}^c) \doteq 1 - \min_{j \in S^*} \Pi_n(\Theta_j) \quad (3.35)$$

□

Proof (Lemma 3.4)

$$\min_{\theta \in \tilde{\Theta}_i} D_\psi(\theta^* || \theta) = \min_{\theta \in \tilde{\Theta}_i} \sum_{j=1}^k \psi_j d(\theta_j^* || \theta_j) \quad (3.36)$$

$$= \min_{\theta \in \tilde{\Theta}_i} \sum_{j \in S^*} \psi_j d(\theta_j^* || \theta_j) + \psi_i d(\theta_i^* || \theta_i) + \sum_{j \notin S^* \cup \{i\}} \psi_j d(\theta_j^* || \theta_j) \quad (3.37)$$

$$= \min_{\theta \in \tilde{\Theta}_i} \sum_{j \in S^*} \psi_j d(\theta_j^* || \theta_j) + \psi_i d(\theta_i^* || \theta_i) \quad (3.38)$$

$$= \min_{j \in S^*} \min_{\theta \in \tilde{\Theta}_i} \psi_j d(\theta_j^* || \theta_j) + \psi_i d(\theta_i^* || \theta_i) \quad (3.39)$$

$$= \min_{j \in S^*} \min_{\theta \in \tilde{\Theta}_i} \psi_j d(\theta_j^* || \theta_j) + \psi_i d(\theta_i^* || \theta_j) \quad (3.40)$$

$$= \min_{j \in S^*} \min_{x \in \mathbb{R}} \psi_j d(\theta_j^* || x) + \psi_i d(\theta_i^* || x) \quad (3.41)$$

$$= \min_{j \in S^*} C_{j,i}(\psi_j, \psi_i) \quad (3.42)$$

(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 their true value - as the KL divergence is minimized for those 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 $\text{top-}m(\theta, S^* \cup \{i\}) \neq S^*$. 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. \square

Proof (Lemma 3.5) For the sake of clarity, let us define $f(x, (\psi_j, \psi_i)) = \psi_j d(\theta_j^* || x) + \psi_i d(\theta_i^* || x)$. 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. \square

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 (3.41) for a given j gives us:

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

$$= \psi_j((\theta_j^* - x)A'(\theta_j^*) - A(\theta_j^*) + A(x)) + \psi_i((\theta_i^* - x)A'(\theta_i^*) - A(\theta_i^*) + A(x)) \tag{3.44}$$

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

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

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

\square

Proof (Proposition 3.7) We prove in the following order: (i) (3.15) must hold for an optimal allocation, (ii) an optimal allocation is unique, (iii) what the convergence rate of the optimal allocation is and (iv) that none can do better.

- (i) Suppose that ψ^* is optimal but does not satisfy (3.15). Hence for some $i_1, i_2 \notin S^*, j_1, j_2 \in S^*$ with $(j_1, i_1) \neq (j_2, i_2)$:

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

This implies:

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

Consider the the measurement plan ψ^ε with

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

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

$$C_{j_1, i_1}(\psi_{j_1}^\varepsilon, \psi_{i_1}^\varepsilon) > C_{j_2, i_2}(\psi_{j_2}^\varepsilon, \psi_{i_2}^\varepsilon) \quad (3.46)$$

while shifting enough measurement to all other arms such that

$$\min_{i \notin S^*} \min_{j \in S^*} C_{j, i}(\psi_j^\varepsilon, \psi_i^\varepsilon) > \min_{i \notin S^*} \min_{j \in S^*} C_{j, i}(\psi_j^*, \psi_i^*) \quad (3.47)$$

Hence ψ^ε obtains more evidence on the worst possible pair than ψ^* and thereby achieves a better convergence rate (3.14). In other words, ψ^* 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. $\psi_l^1 \neq \psi_l^2$. 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

$$C_{j, i_x}(\psi_j^2, \psi_{i_x}^2 + \varepsilon) = C_{j, i_x}(\psi_j^1, \psi_{i_x}^2 + \varepsilon) \quad (3.48)$$

$$= C_{j, i_x}(\psi_j^1, \psi_{i_x}^1) \quad (3.49)$$

$$= C^* \quad (3.50)$$

$$= C_{j, i_x}(\psi_j^2, \psi_{i_x}^2) \quad (3.51)$$

Which is a contradiction as ε is positive C_{j, i_x} monotonically increasing.

- More than 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 $\psi_{i_y}^1 \geq \psi_{i_y}^2$ or $\psi_{i_y}^1 < \psi_{i_y}^2$ 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 $\psi_{j_x}^1 < \psi_{j_x}^2$ has to hold, otherwise both i_x and j_x were allocated more weight in ψ^1 than in ψ^2 . By the monotonicity of C_{j_x, i_x} , this would imply greater evidence for ψ^1 than for ψ^2 , a contradiction to the assumption that both are optimal.

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

$$\begin{aligned} - \psi_{i_x}^1 &= \psi_{i_x}^2 + \varepsilon \\ - \psi_{j_x}^1 &= \psi_{j_x}^2 - \varepsilon' \\ - \psi_{i_y}^1 &= \psi_{i_y}^2 - \varepsilon'' \end{aligned}$$

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

$$C_{j_x, i_y}(\psi_{j_x}^2 - \varepsilon', \psi_{i_y}^2 - \varepsilon'') = C_{j_x, i_y}(\psi_{j_x}^1, \psi_{i_y}^1) \quad (3.52)$$

$$= C^* \quad (3.53)$$

$$C_{j_x, i_y}(\psi_{j_x}^2, \psi_{i_y}^2) \quad (3.54)$$

which, again, is a contradiction by the monotonicity of the KL divergence. \square

l is either inside or outside of S^* . As both cases are analogous, let's assume $l = j_x \in S^*$. Combining the knowledge of (3.15) and $C_{j,i}$ being strictly increasing in the first argument leads to $C_{j_x, i}(\psi_{j_x}^1, \psi_i^1) > C_{j_x, i}(\psi_{j_x}^2, \psi_i^2)$ for all $i \notin S^*$. Hence ψ^2 is not optimal, which is a contradiction.

(iii)

(iv)

Algorithm

4.1 Analysis/Properties

4.2 Empirical behaviour

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

4.3 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 discrete distributions:

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

where Y corresponds to 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:

$$d(\theta_l||x) = \theta_l \log\left(\frac{\theta_l}{x}\right) + (1 - \theta_l) \log\left(\frac{1 - \theta_l}{1 - x}\right) \quad (\text{A.1})$$

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 .

$$\frac{d(d(\theta_l||x))}{dx} = -\frac{\theta_l}{x} + \frac{(1 - \theta_l)}{1 - x} \quad (\text{A.2})$$

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.

$$\frac{df(x)}{dx} = \psi_j \left(\frac{\theta_j}{x} + \frac{(1-\theta_j)}{1-x} \right) + \psi_i \left(\frac{\theta_i}{x} + \frac{(1-\theta_i)}{1-x} \right) \quad (\text{A.3})$$

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

$$\frac{df(x)}{dx} = 0 \Rightarrow (1-x_0)(\psi_j \theta_j + \psi_i \theta_i) = x_0(\psi_j(1-\theta_j) + \psi_i(1-\theta_i)) \quad (\text{A.5})$$

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

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

$$\Rightarrow x_0 = \frac{\psi_j \theta_j + \psi_i \theta_i}{\psi_j + \psi_i} \quad (\text{A.8})$$

Hence we have a very intuitive analytical solution for x : it is an average of the means of j and i , 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:

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

- The KL divergence satisfies:

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

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

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.