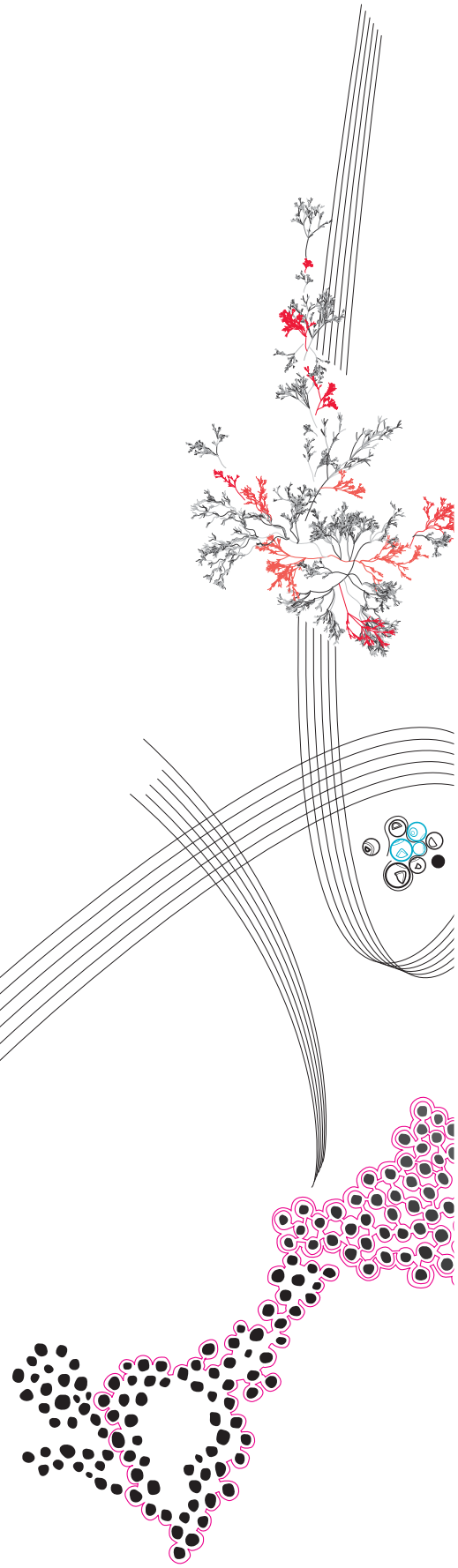

MSc Applied Mathematics
Final Project

On the sample complexity of finding rewards above the arithmetic mean

Tim Huitema

Supervisor: Wouter Koolen

September, 2024

Department of Computer Science
Faculty of Electrical Engineering,
Mathematics and Computer Science,
University of Twente

UNIVERSITY OF TWENTE.

Contents

Abstract

This thesis covers the sample complexity of a dynamic-threshold problem: the problem to find the set of arms that have a reward that is higher than the arithmetic mean of all rewards. This problem configuration differs from existing literature because of the dynamic behaviour of the threshold which is dependent on all arms. The assumption is made that rewards are sampled from an underlying Gaussian distribution. The results are acquired by using the Karush–Kuhn–Tucker conditions for a generic problem description. The sample complexity is compared to similar algorithms made using the Track-and-Stop strategy. We see that the sample complexity of our algorithm for comparable objectives is higher. For our algorithm the δ-PAC property is empirically confirmed. Unfortunately, a component of the algorithm remains dependent on iterative testing for optimality, this however only increases the computational run time and does not disprove the validity of the algorithm. In this thesis a method for accelerating this iterative testing is presented.

Keywords: Sample Complexity, Best Arm Identification, Pure exploration, Dynamic threshold

Chapter 1

Introduction

The key to clinical trials for cures is efficiently identifying the most effective treatments, dosages, or remedies. Minimizing the number of trials required to meet specific criteria is crucial, since prescribing less effective medicine could be detrimental. To model the problem of finding the optimal dosage in the early stage of a clinical trial, the paper [2] uses a multi-armed bandit approach. More applications of the bandit setting are readily available: [8] uses this application for smoothing the process of fine tuning hyper-parameters. The paper [13] proposes to use a bandit algorithm for dynamic pricing of products. It should be emphasised that there are infinite possibilities for using this bandit setting to mimic real-life situations, which makes the field important.

The multi-armed bandit problem is a setting in which a decision-maker is able to perform a certain action (or, in the terminology of the field, to pull an arm), and receive a certain reward. The rewards are assumed to be independent, and sampled from an underlying (unknown) distribution. By repeatedly pulling the arms, the decision-maker iteratively has a better understanding of the underlying distribution from which the rewards are sampled. To know which arm to sample at which time is referred to as the sampling rule. Most literature is about regret minimization. For regret minimization, the accumulated reward during the process is of importance. KL-UCB [6], UCB [1], Thompson sampling [16] are examples of algorithms for the sampling rule. However these are outside the scope of this project. This thesis will only cover pure exploration. We focus solely on minimizing the number of times we have to "pull" arms until we can answer a predetermined query. There are two branches in this field. The first is fixed budget, here the number of times we can pull arms is limited. This setting could be more inline with real-life situations where there could be an underlying cost for each time we choose a certain action. In this thesis, however, we concentrate on the fixed confidence approach, wherein a specific confidence level, denoted by δ, is used to bound the probability of failing to identify the correct answer within finite time. The goal is to minimize the number of pulls, also known as the sample complexity. For the analysis of the sample complexity we rely heavily on the groundbreaking work of Garivier and Kaufmann [7]. That paper dives into the characteristic time of pure exploration problems, which is an optimization problem. They use solutions to this minimization problem for their strategy, Track-and-Stop. Many different specialized strategies for the best arm identification problem have been proposed. However for these strategies there is a gap between the lower bound of the sample complexity and the actual sample complexity of the strategy.

Modifications have been made to the work of [7] in order to analyse a variety of queries.

One of the subjects that is closely related to the topic of this thesis is the identification of the set of arms for which the reward exceeds a predetermined stationary threshold. To illustrate, the objective is to identify arms that have a yearly return of more than 5%. However, in order to determine an appropriate threshold for identifying the optimal response to the query, it is necessary to have a certain degree of understanding regarding the rewards. Another approach, as proposed by [11], is to identify the arms that exhibit a range of ϵ around the maximum mean. In this context, it is also essential to have a clear understanding of the rewards. We believe that a dynamic threshold is a more suitable option for aligning with the goal of real-world scenarios.
This gives reason to try to work out the sample complexity for a similar problem. The problem in question is to find the set of arms that have a reward that is higher than the average of all rewards. We can express this as follows:

$$i^\star(\mu) = \left\{ i \in [m] : \mu_i \geq \frac{1}{m} \sum_{a=1}^{m} \mu_a \right\} \tag{1.1}$$

Where $\mu = \{\mu_1, \mu_2, \ldots, \mu_m\}$ are the means of the rewards for arms $i \in \{1, \ldots, m\}$. $i^\star(\mu)$ is the correct answer on rewards μ. We will use KKT conditions to obtain the necessary equations to model this and use the track-and-stop strategy to implement it.

This thesis is organized in the following manner. In Chapter 2 we give the necessary theoretical background to understand the bandit setting and the algorithm that is used. In Chapter 3 we give the main result of the thesis, the sample complexity of problem 1.1. In Chapter 4 we workout the sample complexity for comparable problems such that we can compare the results in Chapter 5. Afterwards we provide possible ideas for further research in Chapter 6.

Chapter 2

Theoretical background

This Chapter presents the theoretical background necessary for interpreting the main result of the thesis. It begins with a description of a bandit problem, then moves on to present the optimisation problem that is central to the result. It then describes the strategy used to implement the model and provides all the necessary information for its implementation. Finally, it offers an alternative to the strategy and explains how this could be implemented.

2.1 Problem description

The multi-armed bandit model is defined by m probability distributions $\nu_1, \ldots, \nu_m$ all with respective means $\mu = \{\mu_1, \mu_2, \ldots, \mu_m\}$. The scope of this thesis will only include probability distributions that are dependent on one parameter. Additionally, we assume that the rewards are sampled according to a normal distribution.

Following [7], a strategy is defined by

1. Sampling rule $(A_t)_t$
2. Stopping rule τ
3. Decision rule $\hat{i}_\tau(\mu)$

The sampling rule A_t is defined as the arm (or action) $a \in \{1, \ldots, m\}$ at time t we choose to pull. The stopping rule τ is the number of draws we need before we stop the decision process. For fixed confidence strategies we use τ_δ. The goal is to minimize the expected number of draws $E_\nu[\tau_\delta]$, also known as the sample complexity, while we minimize the probability that the decision rule is not the correct answer to the query. When we can guarantee $P(\hat{i}_\tau(\hat{\mu}) \neq i^\star(\mu)) \leq \delta$ and $P(\tau_\delta < \infty) = 1$, the strategy is called δ-PAC. $i^\star(\mu)$ is the correct answer, $\hat{i}_\tau(\hat{\mu})$ is the best answer after time τ_δ. The δ-PAC property is desirable because of the fact that the user has the control to fix the likelihood of getting incorrect answers while minimizing the samples needed.

2.2 Sample complexity

There are many problem-dependent lower bounds for the sample complexity. Most of them are based on the quest to find the best arm. The work [9] establishes a lower bound for general bandit problems that depends on the chosen distribution rather than on the specific problem itself. Before we can state the lower bound, we first need to define some definitions.

First we define the Kullback-Leibler divergence between two probability distributions P, Q

$$kl(P,Q) = \int_{-\infty}^{\infty} P(X) \log\left(\frac{P(X)}{Q(X)}\right) dx \tag{2.1}$$

We assume: $kl(P,Q) < \infty$.

For the first lemma we closely follow the reasoning as provided in the paper [7]. They make, as much bandit literature does, use of a change of distribution argument as has been provided by the groundbreaking work in paper [17].

Lemma 1 *Let τ be any almost surely finite stopping time with respect to the field of observations $\mathcal{F}_t$, for every event $\epsilon \in \mathcal{F}_\tau$*

$$\sum_{a\in[m]} E_\mu[N_a(\tau_\delta)] d(\mu_a, \lambda_a) \geq kl(P_\nu(\epsilon), P_{\nu'}(\epsilon))$$

$kl(P_\nu(\epsilon), P_{\nu'}(\epsilon))$ is the Kullback-Leibler divergence between $P_\nu(\epsilon)$ and $P_{\nu'}(\epsilon)$. $E_\mu[N_a(\tau_\delta)]$ is the expected number of samples for arm a when we stop at time τ_δ. $d(\mu_a, \lambda_a)$ is the Kullback-Leibler divergence between μ_a and λ_a. This lemma captures the relationship between the sample complexity and the Kullback-Leibler divergence.

Using this lemma, we can state the lower bound. For the proof we refer to Appendix: 10.1. For any bandit model μ and any δ-PAC strategy, the following inequality holds:

$$T^\star(\mu) kl(\delta, 1-\delta) \leq \mathbb{E}[\tau_\delta] \tag{2.2}$$

where $T^\star(\mu)$ is defined as

$$T^\star(\mu)^{-1} := \sup_{w\in\Sigma_m} \inf_{\lambda\in \text{Alt}(\mu)} \sum_{a=1}^{m} w_a d(\mu_a, \lambda_a), \tag{2.3}$$

where $d(\mu_a, \lambda_a)$ represents the Kullback-Leibler divergence between μ_a and λ_a. w_a which is the weight of arm a. Then $\sum_m = \{w \in \mathbb{R}^+ : w_1 + \cdots + w_m = 1\}$. Ultimately, $w_a^\star$ is the optimal proportion of samples allocated to arm a. The other unknown component is $Alt(\mu)$ which is:

$$Alt(\mu) := \{\lambda : i^\star(\lambda) \neq i^\star(\mu)\} \tag{2.4}$$

So our set Alt(μ) is a set of bandits, in our case named λ, for which the correct answer, $i^\star(\lambda)$, is not the same as for μ, $i^\star(\mu)$. $d(\mu_a, \lambda_a)$ is the Kullback-Leibler divergence between μ_a and λ_a.

As $\delta \to 0$ the lower bound goes to:

$$\mathbb{E}[\tau_\delta] \geq T^\star(\mu) \log\left(\frac{1}{\delta}\right) \tag{2.5}$$

2.3 Kullback–Leibler divergence

The Kullback-Leibler divergence, as defined in Equation 2.1, is used throughout this project. This divergence has a different form for the assumed probability distribution. In this thesis we make the assumption that the rewards are sampled according to an underlying normal distribution.We assume that the samples are taken from a normal distribution with mean μ and variance 1, i.e.,

$$Y_a \sim \mathcal{N}(\mu_a, 1)$$

where Y_a represents the random sample for mean a.This assumption leads to the following form of the divergence:

$$kl(\mu_i, \lambda_i) = \frac{(\mu_i - \lambda_i)^2}{2} \tag{2.6}$$

The derivation of this result can be found in the appendix: 10.6. It must be stressed that this is not the proper notation, this would be $\mathrm{kl}(\nu^{\mu}, \nu^{\lambda})$. However for (seemingly) aesthetic reasons this notation is abused in most literature. Most literature, for example papers [7] and [14] use a generalized divergence. This is because when their work is implemented the user of the algorithm can determine the assumed distribution. For the generic divergence, Bregman divergence is used.

2.4 Track-and-Stop strategy

Based on our knowledge the Track-and-Stop strategy, designed by [7] has been the only strategy that has been proven to have the sample complexity match the theoretical lower bound of the sample complexity. The Track-and-Stop strategy makes use of solutions acquired by solving the optimization problem listed in Equation 2.3. However this can lead to high computational costs for calculating the solutions needed for the strategy. As will be discussed in Section 2.5, there are ways to greatly reduce this. Prior to outlining the strategy, it is necessary to reiterate the fundamental objective of the Track-and-Stop strategy. This process yields the sampling rule, as detailed in the strategy description. Before we can list the Track-and-Stop strategy we first need to cover some necessary components of the strategy. Other components will be covered in separate Subsections.

2.4.1 Optimal weights $w^\star$

The Track-and-Stop strategy makes use of the solutions of Equation 2.3.

$$w^\star(\mu) \in \arg\max_{w \in \sum_m} \inf_{\lambda \in Alt(\mu)} \sum_{i=1}^{m} w_i d(\mu_i, \lambda_i) \tag{2.7}$$

To use this we have to prove continuity for the optimal weights. In other words, we need to have the assurance that as: $\hat{\mu} \to \mu$ we also need $w^\star(\hat{\mu}) \to w^\star(\mu)$. Continuity for best arm identification is proven in the paper [7]. An extension to all other single answer problems are made in [5]. In there it is proven that for single-answer problems the optimal weights are (upper hemi-) continuous. They also prove that optimal weights are convex. These two factors combined are enough for our purposes.

2.4.2 Observations $\hat{\mu}$

We do not immediately posses, the true value of the corresponding mean of the values for μ, we use preliminary estimates for this by using a rolling estimate on the true value of the mean.

$$\hat{\mu}_a(t) = \frac{1}{N_a(t)} \sum_{s=1}^{t} Y_s \mathbb{1}\{A_s = a\}$$

$N_a(t)$ is the number of draws of arm a at time t. Y_s is the observation of the reward that was pulled at time s. $\mathbb{1}\{A_s = a\}$ is equal to 1 if at time s arm a was pulled, 0 otherwise. By the law of large number we have that $P(\hat{\mu}(t) \to \mu) = 1$ as $N(t) \to \infty$.

2.4.3 Tracking rules

The Track-and-Stop strategy has two tracking rules[7].

C-Tracking

The cumulative tracking makes use of past solutions of the weights. The sampling rule is defined as

$$A_{t+1} \in \arg\max_{1 \le a \le m} \sum_{s \le t} w_a^\star(\hat{\mu}_a(s)) - N_a(t)$$

D-Tracking

D-tracking, or direct tracking. This sampling rule makes use of one more definition

$$u_t = \left\{ a : N_a(t) < \sqrt{t} - \frac{m}{2} \right\}$$

We can then define the sampling rule as follows:

$$A_{t+1} \in \begin{cases} \arg\min_a N_a(t) & \text{if } u_t \neq \emptyset \quad \textit{forced exploration} \\ \arg\max_a t w_a^\star(\hat{\mu}_a(t)) - N_a(t) & \text{if } u_t = \emptyset \quad \textit{direct tracking} \end{cases}$$

The first part is necessary since it forces the exploration of under sampled arms. This is also needed to make sure early inaccuracies of $\hat{\mu}$ do not lead to significant inaccuracies of the weights which subsequently could cause mistakes that do not improve over time. In Subsection stopping rules and thresholds we explore the question of when we have gathered enough information to stop the strategy.

2.4.4 Generalized Likelihood Ratio statistic

For the strategy to identify the correct answer, it must have enough information to exclude all other possible answers. To give a measure of this information we use the generalized log-likehood statistic. The generalized log-likelihood ratio statistic, as defined in [10] is defined as:

$$\mathcal{Z}(t) = \inf_{\lambda \in Alt(\mu)} \sum_{a=1}^{m} N_a(t) d(\hat{\mu}_a(t), \lambda_a)$$

$N_a(t)$ is the number of times arm a is pulled up until the time t. m is the number of arms, $\bar{\hat{\mu}}$ is the preliminary estimate of the average of all rewards. As we will prove below

in Equation 3.27 we show that the GLRT (generalized log-likelihood ratio) for our purpose is:

$$\mathcal{Z}(t) = \min_i \frac{N_i(t)m^2(\bar{\hat{\mu}}(t) - \hat{\mu}_i(t))^2}{2(m(m-2) + N_i(t)\sum_{i=1}^m \frac{1}{N_i(t)})} \tag{2.8}$$

2.4.5 Threshold stopping rule

Now that we know how to calculate the generalized log likelihood ratio statistic, we need something to compare it to. Exceeding this threshold would indicate that we have gathered enough information to exclude all other possible answers. In the paper [3] the following threshold function is listed:

$$\beta(t, \delta) = \log\left(\frac{1 + \log(t)}{\delta}\right) \tag{2.9}$$

This threshold is widely used as a threshold for δ-PAC problems. Formally, there does not exist a proof that this threshold is sufficient for all δ-PAC problems [4]. The justification for using this threshold comes from Proposition 12 from [7]. Here is stated that when the Global-Likelihood-Ratio stopping rule is used there exists an $\alpha > 1, R = R(\alpha, m)$ such that:

$$\beta(t, \delta) = \log(\frac{Rt^\alpha}{\delta})$$

when used as an threshold, together with the "Empirical-Best decision rule" ensures that the problem is δ-PAC.The global-Likelihood-Ratio stopping rule is the rule ensuring that we stop the strategy when the statistic 2.8 exceeds the threshold 2.9. The Empirical-best decision rule entails that we choose $\hat{i}_\tau(\hat{\mu}) = \left\{i \in [m] : \hat{\mu}_i(\tau_\delta) \geq \frac{1}{m}\sum_{i=1}^m \hat{\mu}_i(\tau_\delta)\right\}$

2.5 Gradient Ascent

Finding the optimal weights that solve the Equation 2.3 could be computational expensive. The best-arm-identification and our algorithm are examples of this. They require numerical solvers for each iteration to solve for the optimal weights. As a result, it may take so long for the optimal weights to be calculated that the algorithm becomes unusable for some applications. This raises the question of whether it would be feasible to approximate the weights or to refrain from calculating them at each iteration. The paper [12] dives into this question and tackles the optimization problem from a different angle. The adaptation works using analogies from the zero-sum game perspective. One player tries to play the best proportion for w, while the second player proposes the "hardest" alternative λ for the query in question. The following definitions is used

$$F(w, \mu) := \inf_{\lambda \in Alt(\mu)} \sum_i^m w_i d(\mu_i(t), \lambda_i)$$

Then gradient ascent is performed on this formula for the approximations of the new weights. The new weights are calculated as follows

$$\tilde{w}(t+1) = \arg\max_{w \in \sum_m} \eta_{t+1} \sum_{s=m}^t w \cdot clip_s(\nabla F(\tilde{w}(s), \hat{\mu}(s))) - kl(w, \pi)$$

Clipping the gradient is used to overcome some difficulties when the gradient is unbounded. η_t is the step size, π is a uniform distribution. γ_t is the exploration rate.

The last part to update the weights is

$$w'(t+1) = (1-\gamma_t)\tilde{w}(t+1) + \gamma_t \pi$$

$\tilde{w}$ is skewed towards a uniform distribution to force exploration.
For the purpose of our project this could be implemented as follows:

$$\nabla F(w,\mu) = -\min_i \frac{(\bar{\mu}-\mu_i)^2 m^3 (m-2)}{2w_i^2(m(m-2)+w_i A)^2}$$

Where $A = \sum_{i=1}^{m} \frac{1}{w_i}$. The proof for this is given in appendix: 10.4.

Chapter 3

KKT-conditions

In this Chapter we will outline the main result of this thesis. This will be done by using KKT-conditions for a general description of our problem, we will use the definitions made in the book [15]. We will first give a basic outline of what KKT-conditions entail. In the Subsection 3.1.1 we will workout the inner part of the Equation 2.3. The result for this is:

Theorem 1 *For every $w \in \Sigma_m$ and $\mu = \{\mu_1, \mu_2, \ldots, \mu_m\}$, we have*

$$\inf_{\lambda \in Alt(\mu)} \sum_{i=1}^{m} \frac{1}{2} w_i (\mu_i - \lambda_i)^2 = \min_i \frac{w_i m^2 (\bar{\mu} - \mu)^2}{2(m(m-2) + w_i A)}. \tag{3.1}$$

In Subsection 3.1.2 we will use this result to workout the outer part of Equation 2.3 and acquire the equations necessary to generate optimal weights. These Equations are given as follows:

Theorem 2 *The optimal $w^\star(\mu)$ is given by*

$$w^\star(\mu) \in \arg\max_{w \in \Sigma_m} \min_i \frac{w_i m^2 (\bar{\mu} - \mu)^2}{2(m(m-2) + w_i A)}, \tag{3.2}$$

where w_i and w_j are the solutions, computed as follows:

$$w_i = \sqrt{\frac{\sum_{j \in J} w_j^2}{|J| + m(m-2)}} \quad \forall i \in I$$
$$w_j = \frac{2\psi^\star m(m-2)}{m^2(\bar{\mu} - \mu_j)^2 - 2\psi^\star A} \quad \forall j \in J.$$

We define $I \subseteq [m]$ and $J = [m] \setminus I$, where $I \cap J = \emptyset$. A has the following definition: $A = \sum_{i=1}^{m} \frac{1}{w_i}$.We calculate $\psi^\star$ using the following transformation $\xi^\star = \psi^\star A$, We than calculate $\xi^\star$ by solving following equation:

$$F(\xi) = 0 = \sqrt{\frac{\sum_{j \in J} \left(\frac{2\xi m(m-2)}{m^2(\bar{\mu} - \mu_j)^2 - 2\xi} \right)^2}{\frac{|J| + m(m-2)}{|I|^2}}} - \frac{|I|^2 m(m-2)}{|J| + m(m-2)} \cdot \frac{1}{1 - \frac{m^2}{2\xi(|J| + m(m-2))} \sum_{j \in J} (\bar{\mu} - \mu_j)^2}$$

We can see that the equations for w_i, w_j, ξ are dependent on sets I and J and so are the KKT-conditions that we will derive. There exists one configuration of subsets I and J that solve all necessary KKT-conditions. We show in 3.1.4 how to find these.

3.0.1 KKT-conditions

Karush–Kuhn–Tucker conditions or KKT-conditions are necessary conditions for a solution to be optimal. The KKT-conditions can be extracted from a optimization problem. If we have a generic minimization problem:

$$\begin{aligned} \text{Minimize} \quad & f(x) \\ \text{subject to} \quad & g_i(x) \leq 0, \quad i = 1, \ldots, m \\ & h_j(x) = 0, \quad j = 1, \ldots, p \end{aligned}$$

where $f(x)$ is the objective function, $g_i(x)$ are the inequality constraint functions, and $h_j(x)$ are the equality constraint functions. We define $\lambda_i, \mu_j \in \mathbb{R}$

From this we can derive the Lagrangian function:

$$\mathcal{L}(x, \lambda, \mu) = f(x) + \sum_{i=1}^{m} \lambda_i g_i(x) + \sum_{j=1}^{p} \mu_j h_j(x)$$

where:

- λ_i are the Lagrange multipliers for the inequality constraints $g_i(x) \leq 0$,
- μ_j are the Lagrange multipliers for the equality constraints $h_j(x) = 0$.

If x^* is an optimal solution, the following conditions must be satisfied:

1. **Stationarity:**

$$\nabla f(x^*) + \sum_{i=1}^{m} \lambda_i \nabla g_i(x^*) + \sum_{j=1}^{p} \mu_j \nabla h_j(x^*) = 0$$

2. **Primal feasibility:**

$$g_i(x^*) \leq 0, \quad \forall i = 1, \ldots, m$$

$$h_j(x^*) = 0, \quad \forall j = 1, \ldots, p$$

3. **Dual feasibility:**

$$\lambda_i \geq 0, \quad \forall i = 1, \ldots, m$$

4. **Complementary slackness:**

$$\lambda_i g_i(x^*) = 0, \quad \forall i = 1, \ldots, m$$

For the KKT conditions to be necessary for optimality, we need to have that Slater's condition holds. For this condition to hold we must have that $\exists x : h_j(x) = 0$ and $g_i(x) < 0$. For the problem in 3.1.1 we can see that this condition is met for many different λ_i or λ_j , as long as $\lambda_k \neq \bar{\lambda}, \quad \forall k \in [m]$. For example, $\lambda_i = 1, \lambda_j = 0, \bar{\lambda} = 0.5 \quad \forall i \in I, j \in J$.For the problem in 3.1.2, Slater's condition is met for $w_k = \frac{1}{m} \forall k \in [m]$ and $\psi = 0$.

3.1 Finding the set of means above average

We define the bandit model $\mu = (\mu_1, ..., \mu_m)$. We assume that $\mu_i \neq \bar{\mu}$ for all $i \in \{1, 2, \ldots, m\}$, where $\bar{\mu} = \frac{1}{m}\sum_{i=1}^{m} \mu_i$ is the average of the means. Because this leads to theoretical problems. Since we do not seek unique arms that exceed the threshold, we do not have to worry about the possibility of multiple correct answers. In the paper [5], it is described that there are instances where multiple (distinct) sets of weights are optimal, which would imply that our solution would not be unique.

The query of interest, and the main query of this thesis is given as follows:

$$i^\star(\mu) = \left\{ i \in [m] : \mu_i \geq \frac{1}{m}\sum_{a=1}^{m} \mu_a \right\} \tag{3.3}$$

3.1.1 Infinum

To tackle the optimization problem 2.3 we first focus on the inner part of the equation. We therefore fix $w \in \sum_m$. We begin by addressing it from an optimization standpoint. We formulate KKT-conditions to get a better grip on the characteristics of the solution. By employing a generic formulation of our query, we can examine the behavior of λ and use this to derive $Alt(\mu)$. We use the following definitions

$$\bar{\lambda} = \frac{1}{m}\sum_{i}^{m} \lambda_i \tag{3.4}$$

Furthermore, we define $I \subseteq [m]$ and $J = [m] \setminus I$, where $I \cap J = \emptyset$. Let $X = \{X_i : i \in I\}$, $Y = \{Y_j : j \in J\}$, where $X_i, Y_j \in \mathbb{R}$ and $\nu \in \mathbb{R}$ be the Lagrangian multipliers. We make the assumption that rewards are sampled according to an underlying normal distribution. Hence the Kullback-Leibler divergence has the form 2.6. With that, our problem is:

$$\begin{aligned} &\inf_{\lambda} \sum_{i=1}^{m} \frac{1}{2} w_i(\mu_i - \lambda_i)^2 \\ \text{s.t.} \quad &\bar{\lambda} - \lambda_i \leq 0 \quad \forall i \in I, \\ &\lambda_j - \bar{\lambda} \leq 0 \quad \forall j \in J, \\ &\bar{\lambda} - \frac{1}{m}\sum_{i=1}^{m} \lambda_i = 0. \end{aligned}$$

The Lagrangian can then be expressed as follows:

$$\begin{aligned} \mathcal{L}(\lambda, \bar{\lambda}, X, Y, \nu) = \sum_{i=1}^{m} \frac{1}{2} w_i(\mu_i - \lambda_i)^2 &+ \sum_{i \in I} X_i(\bar{\lambda} - \lambda_i) \\ &+ \sum_{j \in J} Y_j(\lambda_j - \bar{\lambda}) \\ &+ \nu\left(\bar{\lambda} - \frac{1}{m}\sum_{i=1}^{m} \lambda_i\right) \end{aligned} \tag{3.5}$$

The KKT-conditions are:

Stationarity

$$-w_i(\mu_i - \lambda_i) - X_i - \frac{1}{m}\nu = 0 \quad \forall i \in I \tag{3.6}$$

$$-w_j(\mu_j - \lambda_j) + Y_j - \frac{1}{m}\nu = 0 \quad \forall j \in J \tag{3.7}$$

$$\sum_{i \in I} X_i - \sum_{j \in J} Y_j + \nu = 0 \tag{3.8}$$

Complementary slackness

$$X_i(\bar{\lambda} - \lambda_i) = 0 \quad \forall i \in I \tag{3.9}$$

$$Y_j(\lambda_j - \bar{\lambda}) = 0 \quad \forall j \in J \tag{3.10}$$

Dual feasibility

$$X_i \geq 0 \quad \forall i \in I \tag{3.11}$$

$$Y_j \geq 0 \quad \forall j \in J \tag{3.12}$$

Primal feasibility

$$\bar{\lambda} - \lambda_i \geq 0 \quad \forall i \in I \tag{3.13}$$

$$\lambda_j - \bar{\lambda} \geq 0 \quad \forall j \in J \tag{3.14}$$

$$\bar{\lambda} - \frac{1}{m}\sum_{i=1}^{m} \lambda_i = 0 \tag{3.15}$$

We begin by solving for stationarity. We take the sum of 3.6 and 3.7 to link all vectors that are dependent on I and J.

$$\sum_{i \in I} -w_i(\mu_i - \lambda_i) - X_i - \frac{1}{m}\nu = 0 = \sum_{j \in J} -w_j(\mu_j - \lambda_j) - Y_j - \frac{1}{m}\nu = 0$$

$$\sum_{i \in I} -w_i(\mu_i - \lambda_i) - \sum_{j \in J} w_j(\mu_j - \lambda_j) = 0$$

$$\sum_{i \in I} w_i(\lambda_i - \mu_i) = \sum_{j \in J} w_j(\mu_j - \lambda_j) \tag{3.16}$$

We now need to get an expression for λ since this is the only variable in this relation, we consider the following cases for solving complementary slackness Equations 3.9 and 3.10

Case 1 (a) We begin by considering the case where $\lambda_i = \bar{\lambda}$ and $X_i \neq 0 \ \forall i \in I$. We use this assumption in Equation 3.6.

$$-w_i(\mu_i - \bar{\lambda}) - X_i - \frac{1}{m}\nu = 0$$

$$X_i = -\frac{1}{m}\nu - w_i(\mu_i - \bar{\lambda}) \tag{3.17}$$

Case 1 (b). If we make the assumption $\lambda_j = \bar{\lambda}$ and $Y_j \neq 0 \ \forall j \in J$, we get the following:

$$-w_j(\mu_j - \bar{\lambda}) + Y_j - \frac{1}{m}\nu = 0$$

$$Y_j = w_j(\mu_j - \bar{\lambda}) + \frac{1}{m}\nu \tag{3.18}$$

Case 2 (a) We now consider the other case: $\lambda_i \neq \bar{\lambda}, X_i = 0 \ \forall i \in I$

$$-w_i(\mu_i - \lambda_i) - \frac{1}{m}\nu = 0$$
$$-w_i(\mu_i - \lambda_i) = \frac{1}{m}\nu$$

$$\lambda_i = \mu_i - \frac{\frac{1}{m}\nu}{w_i} \tag{3.19}$$

Case 2 (b) We now consider the other case: $\lambda_j \neq \bar{\lambda}, Y_j = 0 \ \forall j \in J$

$$-w_j(\mu_j - \lambda_j) - \frac{1}{m}\nu = 0$$
$$-w_j(\mu_j - \lambda_j) = \frac{1}{m}\nu$$

$$\lambda_j = \mu_j - \frac{\frac{1}{m}\nu}{w_j} \tag{3.20}$$

We now have an understanding of the structure of an optimal λ. We can use this for $Alt(\mu)$. When we constructed the KKT-conditions we fixed the set I, however when we want to find the $Alt(\mu)$ we need to search over many possible subsets I. If we would iterate over all arms and set one, i, equal to $\bar{\lambda}$ and set all other arms λ_j equal to $\mu_j - \frac{\frac{1}{m}\nu}{w_j}$, sets I and J would only exchange one arm. This would generate the subset of $Alt(\mu)$ that is different from the correct answer by one arm. We assume that λ that satisfies the infinum is in this subset. A visualization can be found in Appendix: 10.5. In this Figure we can see that one arm i that is originally in set J and now in set I.

We use the equations 3.16 and 3.20 in combination with the assumption that $|I| = 1$ and $|J| = m - 1$.

$$w_i(\bar{\lambda} - \mu_i) = \sum_{j \in [m] \setminus i} w_j(\mu_j - \lambda_j) \tag{3.21}$$

$$\lambda_j = -\frac{\frac{1}{m}\nu}{w_j} + \mu_j \tag{3.22}$$

$$w_j(\mu_j - \lambda_j) = \frac{1}{m}\nu$$
$$w_i(\bar{\lambda} - \mu_i) = \sum_{j\in[m]\setminus i} \frac{1}{m}\nu$$
$$= \frac{m-1}{m}\nu$$

From this we get the expression $\nu = \frac{m}{m-1} w_i(\bar{\lambda} - \mu_i)$. We now want to make an expression for $\bar{\lambda}$. We do this by using the Equations 3.15, 3.20 and $\lambda_i = \bar{\lambda}$.

$$\bar{\lambda} = \frac{1}{m}\sum_{i=1}^{m} \lambda_i$$
$$\bar{\lambda} = \frac{1}{m}(\bar{\lambda} - \sum_{j\in[m]\setminus i} [\frac{1}{m}\frac{1}{w_j}\nu + \mu_j])$$
$$\bar{\lambda} = -\frac{\nu}{m(m-1)} \sum_{j\in[m]\setminus i} \frac{1}{w_j} + \frac{1}{m-1} \sum_{j\in[m]\setminus i} \mu_j$$
$$\bar{\lambda} = -\frac{\nu}{m(m-1)} \sum_{j\in[m]\setminus i} \frac{1}{w_j} + \frac{1}{m-1}(m\bar{\mu} - \mu_i) \tag{3.23}$$

To now express $\bar{\lambda}$ and ν as only functions of w and μ. For the sake of clarity, we express ν and $\bar{\lambda}$ as:

$$\nu = a\bar{\lambda} - b$$
$$\bar{\lambda} = c\nu + d$$

with

$$a = \frac{m}{m-1} w_i$$
$$b = \frac{m}{m-1} w_i \mu_i$$
$$c = -\frac{1}{m(m-1)} \sum_{j\in[m]\setminus i} \frac{1}{w_j}$$
$$d = \frac{1}{m-1}(m\bar{\mu} - \mu_i)$$
$$\nu = a\bar{\lambda} - b$$
$$\nu = a(c\nu + d) - b$$
$$\nu = ac\nu + ad - b$$
$$\nu - ac\nu = ad - b$$
$$\nu = \frac{ad - b}{1 - ac}$$

$$\nu = \frac{m^2 w_i(\bar{\mu} - \mu_i)}{w_i \sum_{j\in[m]\setminus i} \frac{1}{w_j} + (m-1)^2} \tag{3.24}$$

We now have an expression for ν, which is always defined since the denominator is non-zero. We now express $\bar{\lambda}$

$$\bar{\lambda} = c\nu + d$$
$$\bar{\lambda} = c\frac{ad-b}{1-ac} + d$$
$$\bar{\lambda} = \frac{d-bc}{1-ac}$$

$$\bar{\lambda} = \frac{\mu_i(w_i \sum_{j\in[m]\setminus i} \frac{1}{w_j} + 1) + m^2\bar{\mu} - m(\bar{\mu} + \mu_i)}{w_i \sum_{j\in[m]\setminus i} \frac{1}{w_j} + (m-1)^2} \tag{3.25}$$

We now have every expression necessary to compute the inner part of the optimization problem 2.3

$$\inf_{\lambda\in \mathrm{Alt}(\mu)} \left(\sum_{i=1}^{m} w_a d(\mu_a, \lambda_a) \right) =$$
$$\min_i \inf_{\substack{\lambda_i=\bar{\lambda} \\ \lambda_j = \frac{\frac{1}{m}\nu}{w_j} + \mu}} \frac{w_i(\mu_i - \bar{\lambda})^2}{2} + \sum_{j\in[m]\setminus i} \frac{w_j(\frac{1}{m}\nu)^2}{w_j^2} =$$
$$\min_i w_i \frac{(\mu_i - \bar{\lambda})^2}{2} + \frac{(\frac{1}{m}\nu)^2}{2} \sum_{j\in[m]\setminus i} \frac{1}{w_j}$$

we will address both parts separately.

$$w_i \frac{(\mu_i - \bar{\lambda})^2}{2} =$$
$$\frac{1}{2} w_i (\mu_i - \frac{\mu_i(w_i \sum_{j\in[m]\setminus i} \frac{1}{w_j} + 1) + m^2\bar{\mu} - m(\bar{\mu} + \mu_i)}{w_i \sum_{j\in[m]\setminus i} \frac{1}{w_j} + (m-1)^2)})^2 =$$
$$\frac{m^2(m-1)^2(\mu_i - \bar{\mu})^2}{2(w_i \sum_{j\in[m]\setminus i} \frac{1}{w_j} + (m-1)^2)^2}$$

The last part is

$$\frac{(\frac{1}{m}\nu)^2}{2} \sum_{j\in[m]\setminus i} \frac{1}{w_j} =$$
$$\frac{m^2 w_i^2 \sum_{j\in[m]\setminus i} \frac{1}{w_j} (\bar{\mu} - \mu)^2}{2(w_i \sum_{j\in[m]\setminus i} \frac{1}{w_j} + (m-1)^2)^2}$$

If we now add the parts together we arrive at the wanted expression, we name this expression c_i

$$\frac{m^2((m-1)^2 w_i + w_i^2 \sum_{j\in[m]\setminus i} \frac{1}{w_j})(\bar{\mu} - \mu_i)^2}{2((m-1)^2 + w_i \sum_{j\in[m]\setminus i} \frac{1}{w_j})^2}$$

Finally, we can reduce this to:

$$c_i = \frac{w_i m^2(\bar{\mu} - \mu_i)^2}{2((m-1)^2 + w_i \sum_{j\in[m]\setminus i} \frac{1}{w_j})} \tag{3.26}$$

so to conclude

$$\inf_{\lambda \in Alt(\mu)} \sum_{i=1}^{m} \frac{1}{2} w_i (\mu_i - \lambda_i)^2 = \min_i c_i$$

3.1.2 Argmax

We now focus on the outer part of equation 2.3. We simplify 3.26 by rewriting $\sum_{j\in[m]\setminus i} \frac{1}{w_j}$ such that this term depends on all weights. We also do this to simplify the derivative.

$$c_i = \frac{w_i m^2(\bar{\mu} - \mu_i)^2}{2((m-1)^2 + w_i \sum_{j\in[m]\setminus i} \frac{1}{w_j})} = \tag{3.27}$$

$$\frac{w_i m^2(\bar{\mu} - \mu_i)^2}{2(m(m-2) + w_i A)} \tag{3.28}$$

$$A = \sum_{i=1}^{m} \frac{1}{w_i} \tag{3.29}$$

For the KKT-conditions we need to workout gradients of c_i, these are than given as follows:

$$\frac{\partial c_i}{\partial w_i} = \frac{(m-2)m^3(\bar{\mu} - \mu_i)^2}{2(m(m-2) + w_i A)^2}$$

$$\frac{\partial c_i}{\partial A} = -\frac{m^2 w_i^2(\bar{\mu} - \mu_i)^2}{2(m(m-2) + w_i A)^2}$$

We need to make one more transformation. Using equation 3.27, equation 2.3 is now equal to:

$$T^\star(\mu)^{-1} = \max_{w \in \sum_m} \min_i c_i$$

We cannot simply use KKT conditions if this optimization problem has both a maximization and a minimization part. We transform this by adding the constraint: $\psi \leq c_i$ and maximizing this ψ. Since we maximize we the Lagrangian will have $-\psi$ instead of ψ. We define $\nu, X \in \mathbb{R}$ and $\gamma \in \mathbb{R}^m$ To to find the necessary weights, we use the following optimization problem:

$$\begin{aligned}
\max_{\psi, w, A} \quad & \psi \\
\text{subject to:} \quad & w_i \geq 0, \quad \forall i, \\
& 1 - \sum_{i=1}^{m} w_i = 0, \\
& \psi \leq c_i, \quad \forall i, \\
& A = \sum_{i=1}^{m} \frac{1}{w_i}.
\end{aligned}$$

The Lagrangian is as follows:

$$\mathcal{L}(\psi, A, w_i, X, \nu, \gamma_i) = -\psi - \nu\left(1 - \sum_i w_i\right) + X\left(A - \sum_i \frac{1}{w_i}\right) + \sum_i \gamma_i(\psi - c_i)$$

Stationarity:

$$-1 + \sum_{i=1}^{m} \gamma_i = 0 \tag{3.30}$$

$$\nu + X\frac{1}{w_i^2} - \gamma_i c_i'(w_i) = 0 \quad \forall i \in [m] \tag{3.31}$$

$$X - \sum_{i=1}^{m} \gamma_i c_i'^{(A)} = 0 \tag{3.32}$$

Complementary Slackness:

$$\gamma_i(\psi - c_i) = 0 \quad \forall i \in [m] \tag{3.33}$$

Primal Feasibility:

$$A = \sum_{i=1}^{m} \frac{1}{w_i} \tag{3.34}$$

$$\sum_{i=1}^{m} w_i = 1 \tag{3.35}$$

Dual Feasibility:

$$w_i \geq 0 \quad \forall i \in [m] \tag{3.36}$$

$$\gamma_i \geq 0 \quad \forall i \in [m] \tag{3.37}$$

By working out the case where equation 3.33 is solved by setting all c_i equal to ψ, which is covered in Section 10.2. We found that the equation for the optimal weights 3.49 , cannot hold for every μ. Therefore, we concluded that there must exist a set of indices for which $c_i \neq \psi$. We therefore define $\gamma_i = 0$ for $i \in I$ and set $\gamma_j \geq 0$ for $j \in J$. We define $I \subseteq [m]$ and $J = [m] \setminus I$, where $I \cap J = \emptyset$. We begin by finding an expression for w_i with the help of equation 3.31.

$$\nu + X\frac{1}{w_i^2} = 0$$
$$X\frac{1}{w_i^2} = -\nu$$
$$-\frac{X}{\nu} = w_i^2$$
$$w_i = \sqrt{-\frac{X}{\nu}} \quad \forall i \in I$$

We now find an expression for w_j using equation 3.33.

$$\psi - c_j = 0$$
$$\psi = \frac{w_j m^2(\bar{\mu} - \mu_j)^2}{2(m(m-2) + w_j A - 1)}$$
$$w_j m^2(\bar{\mu} - \mu_j)^2 = 2\psi(m(m-2)) + 2\psi w_j A - 2\psi$$
$$w_j m^2(\bar{\mu} - \mu_j)^2 - 2\psi w_j A = 2\psi(m(m-2))$$

$$w_j = \frac{2\psi m(m-2)}{m^2(\bar{\mu} - \mu_j)^2 - 2\psi A} \quad \forall j \in J \tag{3.38}$$

Using the expressions for w_i, w_j and the definition of A 3.34, we do the following:

$$\sum_{i=1}^{m} \frac{1}{w_i} = \sum_{i \in I} \frac{1}{w_i} + \sum_{j \in J} \frac{1}{w_j}$$
$$= |I| \sqrt{-\frac{\nu}{X}} + \sum_{j \in J} \frac{m^2(\bar{\mu} - \mu_j)^2 - 2\psi A}{2\psi m(m-2)}$$
$$= |I| \sqrt{-\frac{\nu}{X}} + \frac{1}{2\psi m(m-2)} \left(m^2 \sum_{j \in J} (\bar{\mu} - \mu_j)^2 - 2\psi A |J| \right)$$
$$= |I| \sqrt{-\frac{\nu}{X}} + \frac{m^2}{2\psi m(m-2)} \sum_{j \in J} (\bar{\mu} - \mu_j)^2 - \frac{A|J|}{m(m-2)}$$
$$\sum_{i=1}^{m} \frac{1}{w_i} + \frac{|J|}{m(m-2)} \sum_{i=1}^{m} \frac{1}{w_i} = |I| \sqrt{-\frac{\nu}{X}} + \frac{m^2}{2\psi m(m-2)} \sum_{j \in J} (\bar{\mu} - \mu_j)^2$$

$$\sum_{i=1}^{m} \frac{1}{w_i} = \frac{|I| m(m-2)}{|J| + m(m-2)} \sqrt{-\frac{\nu}{X}} + \frac{m^2}{2\psi(|J| + m(m-2))} \sum_{j \in J} (\bar{\mu} - \mu_j)^2 \tag{3.39}$$

Using equation 3.31, we can workout γ_j and use this definition to solve for X in equation 3.32.

$$\nu + X \frac{1}{w_j^2} - \gamma_j c_j'(w) = 0$$
$$\gamma_j = \frac{\nu + X \frac{1}{w_j^2}}{c_j'(w)}$$

$$\gamma_j = \left(\nu + X \frac{1}{w_j^2} \right) \frac{2(m(m-2) + w_j A)^2}{(m-2) m^3 (\bar{\mu} - \mu_j)^2} \tag{3.40}$$

Now we insert this expression for γ_j into Equation 3.32

$$
\begin{aligned}
&X - \sum_{i=1}^{m} \gamma_i c_i'^{(A)} = 0 \\
&X - \sum_{j \in J} \gamma_j c_j'^{(A)} = 0 \\
&X - \sum_{j \in J} \left(\frac{\nu + X \frac{1}{w_j^2}}{c_j'(w)} \right) c_j'^{(A)} = 0 \\
&X - \sum_{j \in J} \left(\left(\nu + X \frac{1}{w_j^2} \right) \frac{2(m(m-2) + w_j A)^2}{(m-2)m^3(\bar{\mu} - \mu_j)^2} \right) \left(-m^2 w_j^2 \frac{(\bar{\mu} - \mu_j)^2}{2(m(m-2) + w_j A)^2} \right) = 0 \\
&X + \sum_{j \in J} \left(\nu + X \frac{1}{w_j^2} \right) \frac{w_j^2}{m(m-2)} = 0 \\
&X + \frac{1}{m(m-2)} \left(\nu \sum_{j \in J} w_j^2 + X|J| \right) = 0 \\
&X \left(1 + \frac{|J|}{m(m-2)} \right) = - \frac{\nu \sum_{j \in J} w_j^2}{m(m-2)}
\end{aligned}
$$

$$
X = - \frac{\nu \sum_{j \in J} w_j^2}{m(m-2) + |J|} \tag{3.41}
$$

We have now derived an expression for X that involves $\sum_{j \in J} w_j^2$ instead of w_j. This is advantageous, as the expression for X now depends solely on constants and ν.

We now solve for ν. We do this by manipulating the expression we have for ψ via the expression for w_j. Our starting point is the equation 3.30

$$
\sum_{j \in J} \gamma_j = \sum_{j \in J} \left(\nu + X \frac{1}{w_j^2} \right) \frac{2(m(m-2) + w_j A)^2}{(m-2)m^3(\bar{\mu} - \mu_j)^2} = 1
$$

Using the definition of ψ

$$
\begin{aligned}
&\psi = \frac{w_j m^2 (\bar{\mu} - \mu_j)^2}{2(m(m-2) + w_j A - 1)} \\
&m(m-2) + w_j A - 1 = \frac{w_j m^2 (\bar{\mu} - \mu_j)^2}{2\psi} \\
&(m(m-2) + w_j A - 1)^2 = \frac{w_j^2 m^4 (\bar{\mu} - \mu_j)^4}{4\psi^2} \\
&2(m(m-2) + w_j A - 1)^2 = \frac{w_j^2 m^4 (\bar{\mu} - \mu_j)^4}{2\psi^2} \\
&\frac{2(m(m-2) + w_j A - 1)^2}{m^3 (\bar{\mu} - \mu_j)^2} = \frac{w_j^2 m (\bar{\mu} - \mu_j)^2}{2\psi^2}
\end{aligned}
$$

Filling in this definition and using the relation for X and ν

$$\sum_{j\in J}\left(\nu - \frac{\nu\sum_{j\in J} w_j^2}{m(m-2)+|J|}\frac{1}{w_j^2}\right)\frac{w_j^2 m(\bar{\mu}-\mu_j)^2}{2(m-2)\psi^2} = 1$$

$$\frac{\nu m}{2(m-2)\psi^2}\left(\sum_{j\in J} w_j^2 - \frac{\sum_j w_j^2}{m(m-2)+|J|}\right)(\bar{\mu}-\mu_j)^2 = 1$$

$$\nu = \frac{2(m-2)\psi^2}{m\left(\sum_{j\in J} w_j^2(\bar{\mu}-\mu_j)^2 - \frac{\sum_{j\in J} w_j^2}{|J|+m(m-2)}\sum_{j\in J}(\bar{\mu}-\mu_j)^2\right)} \tag{3.42}$$

If we fill this into the equations for X, 3.41 and use those results for equations 3.40, 3.39 and 3.1.2 we get the following expressions:

$$X = -\frac{2(m-2)\psi^2\sum_{j\in J} w_j^2}{m(|J|+m(m-2))}\frac{2(m-2)\psi^2}{m\left(\sum_{j\in J} w_j^2(\bar{\mu}-\mu_j)^2 - \frac{\sum_{j\in J} w_j^2}{|J|+m(m-2)}\sum_{j\in J}(\bar{\mu}-\mu_j)^2\right)} \tag{3.43}$$

$$\gamma_j = \frac{(\bar{\mu}-\mu_j)^2\left(w_j^2 - \frac{\sum_{j\in J} w_j^2}{|J|+m(m-2)}\right)}{\sum_{j\in J} w_j^2(\bar{\mu}-\mu_j)^2 - \frac{\sum_{j\in J} w_j^2\sum_{j\in J}(\bar{\mu}-\mu_j)^2}{m(m-2)+|J|}} \quad \forall j \in J \tag{3.44}$$

$$A = \frac{|I|^2 m(m-2)}{|J|+m(m-2)}\frac{1}{1-\sum_{j\in J} w_j} + \frac{m^2}{2\psi(|J|+m(m-2))}\sum_{j\in J}(\bar{\mu}-\mu_j)^2 \tag{3.45}$$

$$w_i = \sqrt{\frac{\sum_{j\in J} w_j^2}{|J|+m(m-2)}} \quad \forall i \in I \tag{3.46}$$

The final expression that we need to use is equation 3.35. We use the definition of w_i.

$$|I|\sqrt{\frac{\sum_{j\in J} w_j^2}{|J|+m(m-2)}} = 1 - \sum_{j\in J} w_j$$

$$\sqrt{\frac{\sum_{j\in J} w_j^2}{|J|+m(m-2)}} = \frac{1-\sum_{j\in J} w_j}{|I|}$$

$$\sum_{j\in J} w_j^2 = \frac{|J|+m(m-2)}{|I|^2}(1-\sum_{j\in J} w_j)^2$$

We cannot directly use this relation since $w_{j\in J}$ depends on ψ and on $\sum_j w_j^2$ through its dependence on A. To tackle this problem we make use of a variable change. We define:

$\xi = \psi A$. We will alter all necessary relations with this variable change.

$$\sum_{j\in J} w_j^2 = \frac{|J| + m(m-2)}{|I|^2}\left(1 - \sum_{j\in J} w_j\right)^2$$

$$\sum_{j\in J}\left(\frac{2\psi m(m-2)}{m^2(\bar{\mu}-\mu_j)^2 - 2\psi A}\right)^2 = \frac{|J| + m(m-2)}{|I|^2}\left(1 - \sum_{j\in J}\left(\frac{2\psi m(m-2)}{m^2(\bar{\mu}-\mu_j)^2 - 2\psi A}\right)\right)^2$$

$$\frac{1}{A^2}\sum_{j\in J}\left(\frac{2\psi A m(m-2)}{m^2(\bar{\mu}-\mu_j)^2 - 2\psi A}\right)^2 = \frac{|J| + m(m-2)}{|I|^2}\left(1 - \frac{1}{A}\sum_{j\in J}\left(\frac{2\psi A m(m-2)}{m^2(\bar{\mu}-\mu_j)^2 - 2\psi A}\right)\right)^2$$

$$\frac{\frac{1}{A^2}\sum_{j\in J}\left(\frac{2\psi A m(m-2)}{m^2(\bar{\mu}-\mu_j)^2 - 2\psi A}\right)^2}{\frac{|J| + m(m-2)}{|I|^2}} = \left(1 - \frac{1}{A}\sum_{j\in J}\left(\frac{2\psi A m(m-2)}{m^2(\bar{\mu}-\mu_j)^2 - 2\psi A}\right)\right)^2$$

$$\frac{\sum_{j\in J}\left(\frac{2\psi A m(m-2)}{m^2(\bar{\mu}-\mu_j)^2 - 2\psi A}\right)^2}{\frac{|J| + m(m-2)}{|I|^2}} = \left(A - \sum_{j\in J}\left(\frac{2\psi A m(m-2)}{m^2(\bar{\mu}-\mu_j)^2 - 2\psi A}\right)\right)^2$$

$$A = \sqrt{\frac{\sum_{j\in J}\left(\frac{2\xi m(m-2)}{m^2(\bar{\mu}-\mu_j)^2 - 2\xi}\right)^2}{\frac{|J| + m(m-2)}{|I|^2}}} + \sum_{j\in J}\frac{2\xi m(m-2)}{m^2(\bar{\mu}-\mu_j)^2 - 2\xi} \tag{3.47}$$

We also need to convert the other relations that we have left that uses ψ or A.

$$A = \frac{|I|^2 m(m-2)}{|J| + m(m-2)}\frac{1}{1 - \sum_{j\in J}\frac{2\psi m(m-2)}{m^2(\bar{\mu}-\mu_j)^2 - 2\psi A}} + \frac{m^2}{2\psi(|J| + m(m-2))}\sum_{j\in J}(\bar{\mu}-\mu_j)^2$$

$$A = \frac{|I|^2 m(m-2)}{|J| + m(m-2)}\frac{1}{1 - \frac{1}{A}\sum_{j\in J}\frac{2\psi A m(m-2)}{m^2(\bar{\mu}-\mu_j)^2 - 2\psi A}} + A\frac{m^2}{2\psi A(|J| + m(m-2))}\sum_{j\in J}(\bar{\mu}-\mu_j)^2$$

$$A = \sum_{j\in J}\frac{2\psi A m(m-2)}{m^2(\bar{\mu}-\mu_j)^2 - 2\psi A} + \frac{|I|^2 m(m-2)}{|J| + m(m-2)}\frac{1}{1 - \frac{m^2}{2\psi A(|J|+m(m-2))}\sum_{j\in J}(\bar{\mu}-\mu_j)^2}$$

$$A = \sum_{j\in J}\frac{2\xi m(m-2)}{m^2(\bar{\mu}-\mu_j)^2 - 2\xi} + \frac{|I|^2 m(m-2)}{|J| + m(m-2)}\frac{1}{1 - \frac{m^2}{2\xi(|J|+m(m-2))}\sum_{j\in J}(\bar{\mu}-\mu_j)^2} \tag{3.48}$$

3.1.3 case: $|J| = m$

As can be seen from the equations 3.46 and 3.38, we explicitly choose to have two types of weights, one for weights in set I, the other in set J. If we deal with a certain set μ for which the set $J = \{1 \dots m\}$ is needed to fulfill the KKT-conditions, the equations we need, have another form. Namely we do not need to have 3.46. If we do this, equations 3.38 and 3.48 simplify and we can calculate ψ analytically. We then have a closed form equation for the weights, namely:

$$w_i = \frac{\frac{1}{m^2(\bar{\mu}-\mu_i)^2 - \frac{m}{m-1}\sum_{i=1}^m(\bar{\mu}-\mu_i)^2}}{\frac{1}{\sum_{i=1}^m m^2(\bar{\mu}-\mu_i)^2 - \frac{m}{m-1}\sum_{i=1}^m(\bar{\mu}-\mu_i)^2}} \tag{3.49}$$

The proof for this can be found in Appendix 10.2. We can see from this equation that this could lead to a violation of equation 3.36.

3.1.4 Implementation

As can be seen from equations 3.47 and 3.48 we do not have an analytical expression we can solve to get the necessary values for ψ to yield our optimal weights. To overcome this problem we find the root numerically by using a dichotomic search on the following function:

$$F(\xi) = \sqrt{\frac{\sum_{j\in J}\left(\frac{2\xi m(m-2)}{m^2(\bar{\mu}-\mu_j)^2-2\xi}\right)^2}{\frac{|J|+m(m-2)}{|I|^2}}} - \frac{|I|^2 m(m-2)}{|J|+m(m-2)} \cdot \frac{1}{1-\frac{m^2}{2\xi(|J|+m(m-2))}\sum_{j\in J}(\bar{\mu}-\mu_j)^2}$$

When we acquire ξ we use this in term to get the values for ψ and A from equation 3.47 and the original variable change $\xi = A\psi$. We need these values to compute the optimal weights using equations 3.46 and 3.38. The last part needed to compute the optimal weights is J.

Finding correct set J

Finding the correct sets I and J is of upmost importance. There exists only one pair of sets for which all KKT-conditions are met. We can go through all possible subsets and check for each set if all KKT conditions are met, however the number of subsets scales exponentially $(2^m - 1)$ for an m-armed bandit problem. (-1 because we exclude J to be empty). We did not find a method to analytically determine sets I and J. We rely on the following heuristic:

Heuristic: Finding J

1. **Compute the mean:**

$$\bar{\mu} = \frac{1}{m}\sum_{i=1}^{m}\mu_i$$

2. **Define the distance:**

$$\Delta_i = (\mu_i - \bar{\mu})^2$$

3. **Sort by distance:** Sort the indices of μ_i's based on their Δ_i values in ascending order:

$$\Delta_1 \leq \Delta_2 \leq \cdots \leq \Delta_m$$

4. **Initialization:** Initialize the set J with the arm with the smallest distance Δ:

$$J = \{\mu_1\}$$

5. **Verification:** Check whether the KKT conditions are satisfied. If not, iteratively add the next smallest μ_i (based on Δ_i) to J, and repeat the verification until the conditions are met.

From Figure 3.1 we can see the performance of using this heuristic. As can be seen, the number of subsets that needed to be checked until we find the correct KKT grows less than linearly with the number of arms m for the heuristic. The heuristic is built onto two pillars. First the following conjecture:

Conjecture 1 *Consider the set of weights $\mu = \{\mu_1, \mu_2, \ldots, \mu_m\}$, and let the overall mean be defined as*

$$\bar{\mu} = \frac{1}{m} \sum_{i=1}^{m} \mu_i.$$

If we order the arms according to their distance to the mean $\bar{\mu}$, the arm with the smallest distance corresponding to μ_1 is always in J

Secondly, we use the fact that the set J is an interval.

Theorem 3 *Let $\mu_1, \ldots, \mu_m$ be a set of means of rewards, and let μ be ordered according to their distance from the mean $\overline{\mu}$, such that μ_1 corresponds to mean with the smallest distance and μ_k corresponds to the mean with the k-th smallest distance. Then there exists a set $J \subseteq \{1, \ldots, m\}$, which is an interval, meaning:*

- *If $\mu_i \in J$ and $\mu_k \in J$, then for every $j \in \{1, \ldots, m\}$ such that $|\bar{\mu} - \mu_j| \leq |\bar{\mu} - \mu_k|$, we have $\mu_j \in J$.*

In other words, if both μ_i and μ_k are in J, then all μ_j that have a distance closer to the mean than μ_k are also in set J.

The proof for this theorem can be found in the appendix: 10.3.

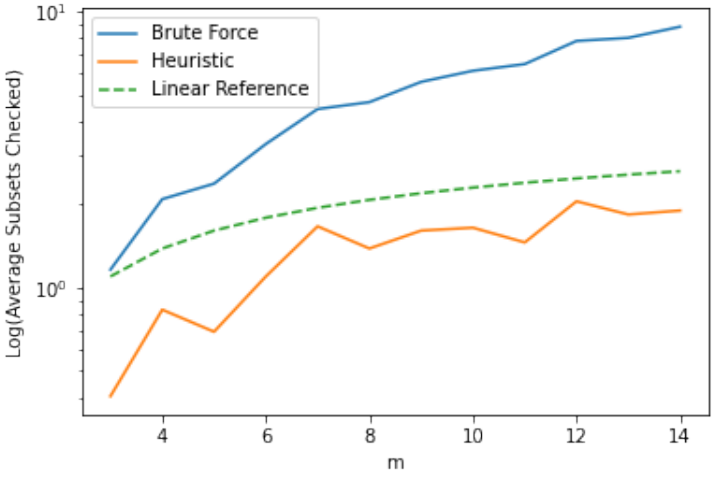

FIGURE 3.1: The plot shows the average of 10 experiments, where each experiment involves generating a random subset μ with increasing length m

Bounds

To use the dichotomic search, we need have bounds for the possible values of ξ. We know that both A and ψ are strictly positive. We can now focus on the equation for A to find the bounds:

$$A = \sum_{j \in J} \frac{2\xi m(m-2)}{m^2(\bar{\mu} - \mu_j)^2 - 2\xi} + \frac{|I|^2 m(m-2)}{|J| + m(m-2)} \frac{1}{1 - \frac{m^2}{2\xi(|J| + m(m-2))} \sum_{j \in J} (\bar{\mu} - \mu_j)^2}$$

we see that $\xi < \frac{m^2(\bar{\mu}-\mu_j)^2}{2} \forall j \in J$. Otherwise the denominator of the first term is negative. From the second term we see that

$$1 > \frac{m^2 \sum_j (\bar{\mu} - \mu_j)^2}{2\xi(|J| + m(m-2))}$$

$$\xi > \frac{m^2 \sum_j (\bar{\mu} - \mu_j)^2}{2(|J| + m(m-2))}$$

$$\frac{m^2 \sum_j (\bar{\mu} - \mu_j)^2}{2(|J| + m(m-2))} < \xi < \frac{m^2(\bar{\mu} - \mu_j)^2}{2} \quad \forall j \in J$$

3.1.5 Overview

To conclude, we developed an algorithm that determines the weights with respect to a certain configuration I and J. To acquire the optimal weights, we need to find the correct set I and J. We provided a heuristic to accelerate this process.

Strategy overview

1. **Sampling rule:**

$$A_{t+1} \in \begin{cases} \arg\min_a N_a(t) & \text{if } u_t \neq \emptyset \quad \textit{forced exploration} \\ \arg\max_a t w_a^\star(\hat{\mu}_a(t)) - N_a(t) & \text{if } u_t = \emptyset \quad \textit{direct tracking} \end{cases}$$

2. **Stopping rule:**

$$\tau_\delta = \inf \left\{ t \in \mathbb{N} : \min_i \frac{N_i(t) m^2 (\bar{\hat{\mu}}(t) - \hat{\mu}_i(t))^2}{2(m(m-2) + N_i(t) \sum_{i=1}^m \frac{1}{N_i(t)})} > \log\left(\frac{1 + \log(t)}{\delta}\right) \right\}$$

3. **Recommendation:**

$$\hat{i}_{\tau_\delta}(\hat{\mu}) = \left\{ i \in [m] : \hat{\mu}_i(\tau_\delta) \geq \frac{1}{m} \sum_{i=1}^m \hat{\mu}_i(\tau_\delta) \right\}$$

Chapter 4

Results for existing problems

We can compare the performance of our model with other implementations of the Track-and-Stop strategy. However we need to carefully construct a sequence of μ for which the objective of identifying the set of means above the average is the same as for the other queries for which we want to compare. We first work these related problems out using KKT-conditions.

4.0.1 Sample complexity of thresholding bandit

The paper [14] mentions multiple possible queries for pure exploration. One of them closely resembles the problem of this thesis. However with the significant difference of a stationary threshold which is not, necessarily, dependent on means. The correct answer for the problem of finding the arms above a threshold $\alpha \in \mathbb{R}$ is:

$$i^\star(\mu) = \{i \in [m] : \mu_i \geq \alpha\}$$

the alternative set of μ is defined as

$$Alt(\mu) = \bigcup_j \{\lambda : (\mu_j - \alpha)(\lambda_j - \alpha) < 0\}$$

This notation ensures that when some arm in μ exceeds the threshold, this particular arm in λ does not and thus does not belong to this set, and vice versa. When we evaluate the inner part of the optimization problem we get the following result

$$\inf_{\lambda \in \text{Alt}(\mu)} \left(\sum_{i=1}^{m} w_i d(\mu_i, \lambda_i) \right) =$$
$$\inf_{\lambda:(\mu_i-\alpha)(\lambda_i-\alpha)<0} \min_i \left(\sum_{a=i}^{m} w_i \frac{(\mu_i - \lambda_i)^2}{2} \right) =$$
$$\min_i \left(w_i \frac{(\mu_i - \alpha)^2}{2} \right)$$

we can calculate the entire optimization problem as follows

$$\arg\max_{w \in \sum_k} \min_i \frac{1}{2} w_i (\mu_i - \alpha)^2$$

To solve this we rewrite the problem such that we are able to solve this with KKT-conditions.

$$\begin{aligned}
&\max_{\psi,w} \psi \\
&\sum_{i=1}^{m} w_i = 1 \\
&\quad w_i \geq 0 \quad \forall i \in [m] \\
\psi - \frac{1}{2} w_i (\mu_i - \alpha)^2 &\leq 0
\end{aligned}$$

We define $\nu \in \mathbb{R}$ and $x \in \mathbb{R}^m$. The Lagrangian for this problem is:

$$\mathcal{L}(\psi, w_i, \nu, x_i) = -\psi + v \left(\sum_{i=1}^{m} w_i - 1 \right) + \sum_{i=1}^{m} x_i \left(\psi - \frac{1}{2} w_i (\mu_i - \alpha)^2 \right)$$

The KKT conditions are:

1. Stationarity:

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \psi} &= -1 + \sum_{i=1}^{m} x_i = 0 \\
\frac{\partial \mathcal{L}}{\partial w_i} &= v - \frac{1}{2} x_i (\mu_i - \alpha)^2 = 0 \quad \forall i \in [m]
\end{aligned}$$

2. Primal feasibility:

$$\begin{aligned}
&\sum_{i=1}^{m} w_i = 1 \\
&\quad w_i \geq 0 \quad \forall i \in [m]
\end{aligned}$$

3. Dual feasibility:

$$x_i \geq 0 \quad \forall i \in [m]$$

4. Complementary slackness:

$$x_i \left(\psi - \frac{1}{2} w_i (\mu_i - \alpha)^2 \right) = 0 \quad \forall i \in [m]$$

From the complementary slackness condition:

$$x_i \left(\psi - \frac{1}{2} w_i (\mu_i - \alpha)^2 \right) = 0$$

We assume that $x_i \neq 0$ because the corresponding constraint is active. This constraint ensures that $\frac{1}{2} w_i (\mu_i - \alpha)^2$ is minimized, which directly influences the optimal solution.

Since this minimization is crucial to determining the optimal point, the constraint must remain active. We also see that: $\psi - \frac{1}{2}w_i(\mu_i - \alpha)^2 \leq 0$. Therefore we have:

$$\psi = \frac{1}{2}w_i(\mu_i - \alpha)^2$$

Solving for w_i:

$$w_i = \frac{2\psi}{(\mu_i - \alpha)^2}$$

Summing over all i and using the constraint $\sum_{i=1}^{m} w_i = 1$:

$$\sum_{i=1}^{m} w_i = \sum_{i=1}^{m} \frac{2\psi}{(\mu_i - \alpha)^2} = 1$$

Solving for ψ:

$$2\psi \sum_{i=1}^{m} \frac{1}{(\mu_i - \alpha)^2} = 1 \Rightarrow \psi = \frac{1}{2\sum_{i=1}^{m} \frac{1}{(\mu_i - \alpha)^2}}$$

Substituting ψ back into the expression for w_i:

$$w_i = \frac{2\left(\frac{1}{2\sum_{i=1}^{m} \frac{1}{(\mu_i - \alpha)^2}}\right)}{(\mu_i - \alpha)^2} = \frac{\frac{1}{((\mu_i - \alpha)^2)}}{\sum_{i=1}^{m} \frac{1}{(\mu_i - \alpha)^2}}$$

This is a closed form equation for the computation of the optimal weights. This is advantageous because, in contrast to for example Best Arm Identification, we do not have to use numerical solvers to compute equations and is thus computationally cheaper. In theory, this problem could have the same answer as our algorithm. When the threshold is set close to the initial average, and the means don't vary significantly, the average remains stable, ensuring that the set of means above and below the average doesn't change.

Interestingly enough there is another (possibly more) query with the exact same formula for the optimal weights, namely: How many arms are above a threshold? A part of the derivation can be found in the appendix: 10.7.

4.0.2 Sample proportions for best arm identification

Garivier and Kaufmann's work, [7], has had a significant impact on the field of pure exploration. Their work is on the complexity of finding the best arm, within a single parameter bandit problems. The query in question is defined as:

$$i^\star(\mu) = \{i : \mu_i > \mu_j \quad \forall j \in [m] \setminus i\}$$

The alternative of μ is written by the authors to be equal to

$$Alt(\mu) = \bigcup_{a \neq 1} \{\lambda \in S : \lambda_a > \lambda_1\}$$

Where $\lambda_1 > \max_{b \neq 1} \lambda_b$. μ is defined such that $\mu_1 > \mu_2 \geq \cdots \geq \mu_m$. This definition implies that we focus on bandits where there is a unique best arm.

The inner part of the optimization problem is then worked at in the following manner:

$$\begin{aligned}
&\inf_{\lambda \in \text{Alt}(\mu)} \left(\sum_{a=1}^{m} w_j d(\mu_a, \lambda_a) \right) \\
&= \min_{a \neq 1} \inf_{\lambda \in S : \lambda_a > \lambda_1} \left(\sum_{a=1}^{m} w_a d(\mu_a, \lambda_a) \right) \\
&= \min_{a \neq 1} \inf_{\lambda \in S : \lambda_a > \lambda_1} w_a d(\mu_a, \lambda_a) + w_1 d(\mu_1, \lambda_1)
\end{aligned}$$

The minimization of $w_a d(\mu_a, \lambda_a) + w_1 d(\mu_1, \lambda_1)$ is done analytically and the corresponding value for λ is then:

$$\lambda_a = \lambda_1 = \frac{w_1}{w_1 + w_a} \mu_1 + \frac{w_a}{w_1 + w_a} \mu_a$$

The whole optimization problem is then rewritten as

$$\arg\max_{w \in \sum_m} w_1 \min_{a \neq 1} g_a \left(\frac{w_a}{w_1} \right)$$

$$g_a(x) = d(\mu_1, m_a(x)) + x d(\mu_a, m_a(x))$$

$$m_a(x) = \frac{\mu_1 + x \mu_a}{1 + x}$$

After further manipulation we end up with the following relation to find the optimal weights:

$$w_a^\star(\mu) = \frac{x_a(y^\star)}{\sum_a x_a(y^\star)}$$

Where $x_a(y) = g_a^{-1}(y)$ and $y^\star$ is the unique solution to the equation:

$$\sum_{a=2}^{m} \frac{d\left(\mu_1, m_a(x_a(y))\right)}{d\left(\mu_a, m_a(x_a(y))\right)} = 1$$

This problem potentially closely resembles the problem of finding the set of means above the mean of the rewards, only in the case when one mean has a dramatically high reward such that it is the only mean above the average. Then we are essentially searching for the arm with the biggest reward.

Now that we have the equations for the optimal weights for both algorithms (Best Arm Identification and arms above threshold). We will compare the performance in the next section.

Chapter 5

Numerical experiments

In this Chapter we implement the results for existing problems from Chapter 4 and compare the performance to our algorithm. We also test influence of the confidence δ on the expected stopping time. Also the influence of the variance $s^2 = \frac{1}{m}\sum_{i=1}^{m}(\mu_i - \bar{\mu})^2$ of the numbers inside set μ. In the numerical experiments, our algorithm is referred to as *FindingMeano*. The algorithm for identifying arms above a given threshold is denoted as *AboveThreshold*, while the algorithm for identifying the best arm is simply referred to as *Best Arm Identification*.

5.0.1 Influence of δ

In Figure 5.1 we investigate the relationship between δ and the expected stopping time of our algorithm 3.1.5. From this Figure we can see that the average values for the sample

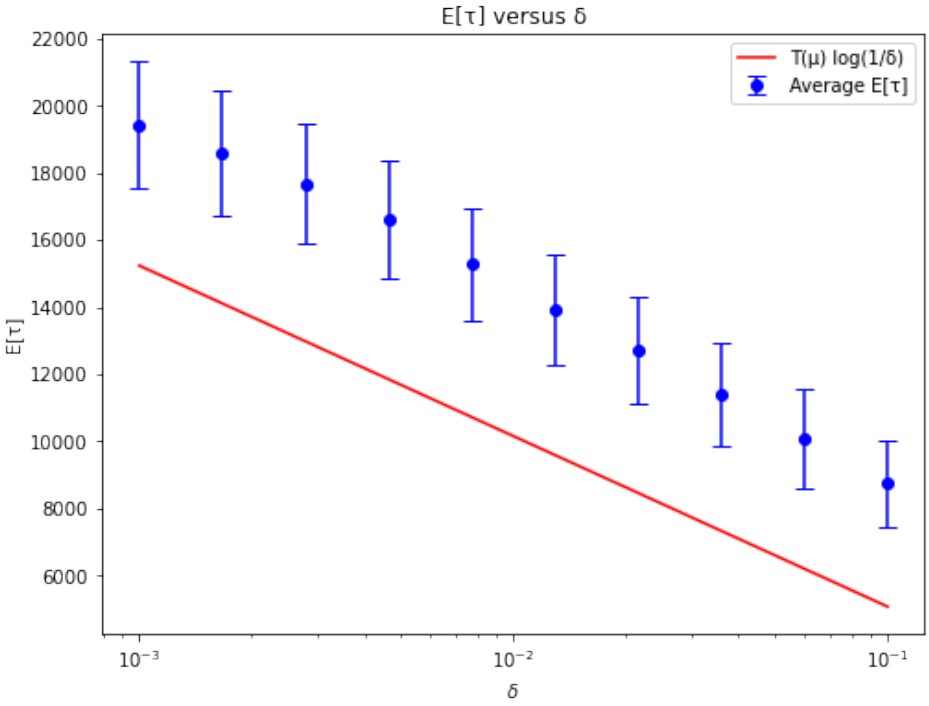

FIGURE 5.1: The Figure depicts the relationship between the sample complexity and δ. $\mu = [0.29, 0.12, 0.07]$, $T^\star(\mu) = 2205$, N = 100 repetitions for each δ

complexity are lower for higher values of δ. We additionally plotted a 95% confidence interval around the average sample complexity.

5.0.2 Influence of σ

To investigate the relationship between the sample complexity and the variance of the mean of the rewards the following Figure is made: 5.2 In this Figure, all μ_i for $i \in \{1, 2, 3, 4\}$

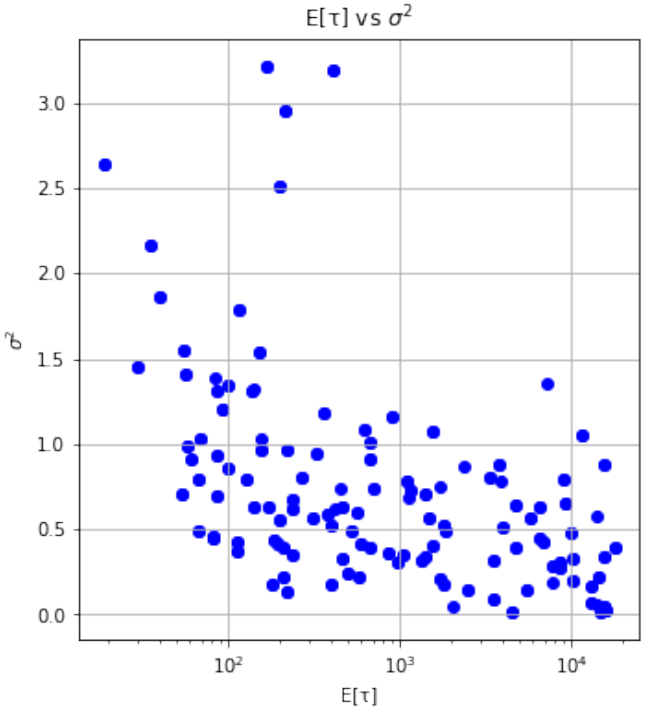

FIGURE 5.2: The Figure depicts the relationship between the sample complexity and the variance of μ. The set of values μ is generated from the distribution $\mathcal{N}(0, 1)$.

are drawn from a normal distribution with a mean of zero and a variance of one, denoted as $\mathcal{N}(0, 1)$. From this, we measure the variance of the generated set of means of rewards using the following formula: $s^2 = \frac{1}{m-1} \sum_{i=1}^{m} (\mu_i - \bar{\mu})^2$ with $\bar{\mu} = \frac{1}{m} \sum_{i=1}^{m} \mu_i$. We keep track of the relation between the sample complexity and this variance. We can see that for lower values of this measured variance we see more instances of (relative) high values of the sample complexity. However these observations do not immediately lead to significant conclusions.

5.0.3 Evolution of weights

We plot the evolution of the weights over time in Figure 5.3. In the Figure we can see the evolution of the change of weights. The grey lines indicate that at that time there is a shift of the set J where that specific arm is involved. So or the set J now includes or excludes that specific arm. We can see that at these times there is a significant change in value of that specific weight. After a certain time, where there is no change in the set J anymore, we see that all the weights are relatively stable.

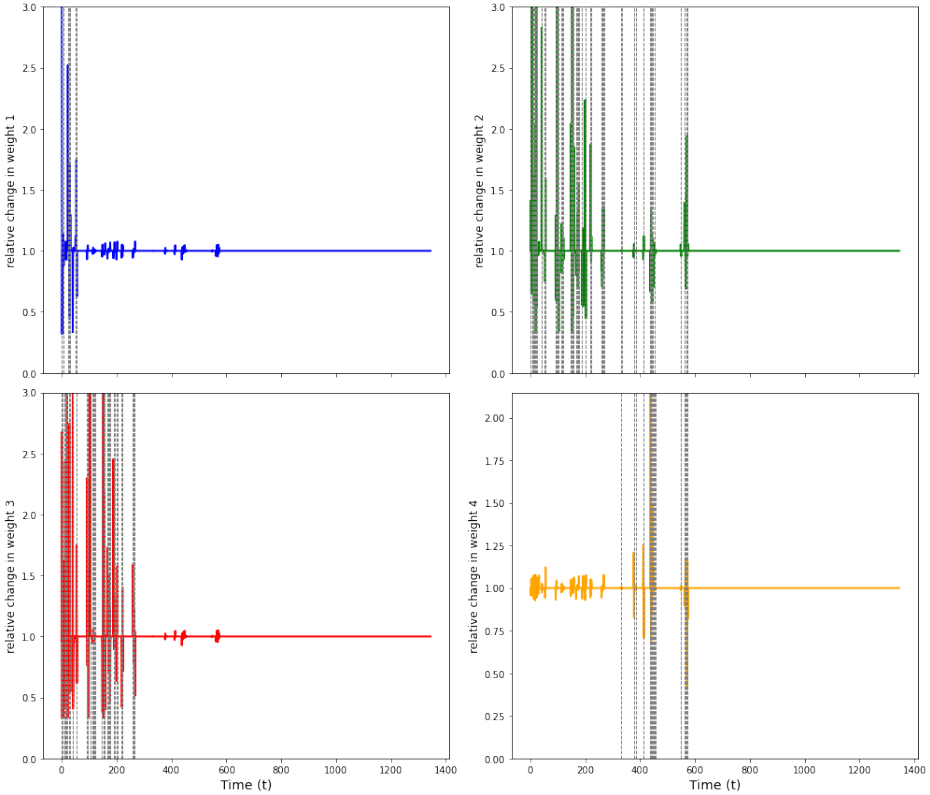

FIGURE 5.3: The plot depicts the evolution of the weights over time. $\mu = [0.1, 0.25, 0.3, 0.4]$, $\delta = 0.01$, grey lines indicate switch of set J, $T^\star(\mu) = 28800$

5.0.4 Comparisons

A comparison can be made between our algorithm, which we call FindingMeano and other algorithms that have been developed for Track-and-Stop. This can be achieved by selecting an appropriate μ and ensuring that the resulting answer is identical for both. Firstly, we compare our algorithm with one designed for threshold bandits. The threshold is then set to an average of the mean rewards. Secondly, a set μ is selected, whereby the highest value is so extreme such that it is the sole value above the mean. This renders the identification of all arms with an above-average reward identical to the selection of the arm with the highest reward.

From Figure 5.4 we can clearly see that the algorithm for threshold bandit outperforms our algorithm for the configurations mentioned above

From Figure 5.5 we can see that best arm identification outperforms our algorithm when μ is chosen such that the answer for both our algorithm and Best Arm Identification are equivalent.

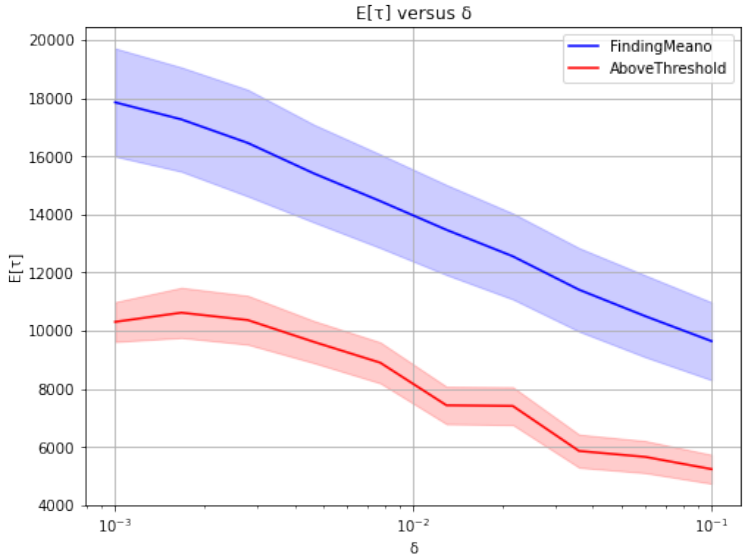

FIGURE 5.4: Plot shows the relationship between δ and the sample complexity for both models: AboveThreshold and FindingMeano. The values of μ are generated as follows: $\mu = [0.29, 0.25, 0.07]$, $T^{\star}_{FindingMeano}(\mu) = 2067$, $T^{\star}_{AboveThreshold}(\mu) = 1297$

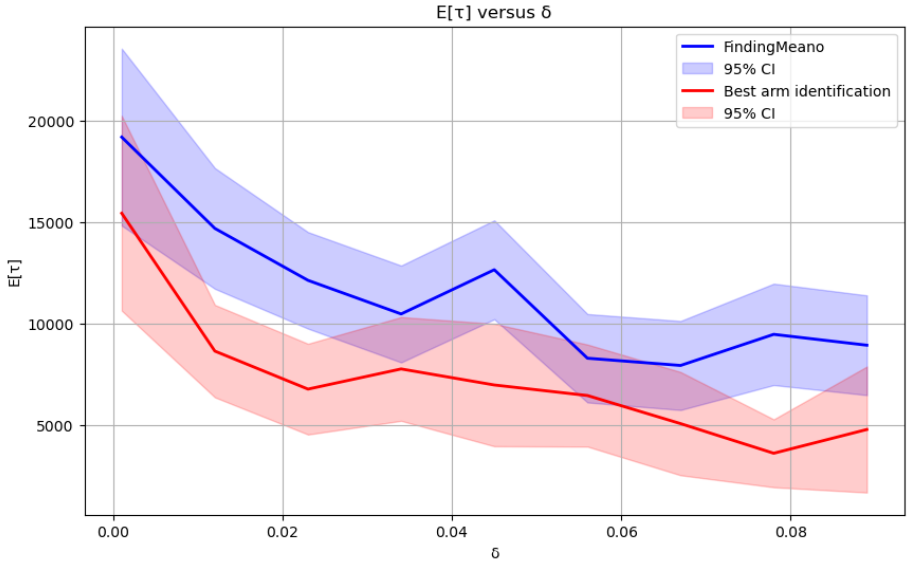

FIGURE 5.5: Plot shows the relationship between δ and the sample complexity for both models: Best Arm Identification and FindingMeano ($T^{\star}(\mu) = 2205$). The values of μ are generated as follows: $\mu = [0.29, 0.12, 0.07]$.

Chapter 6

For further research

6.0.1 Extension of the objective

For further research it would be interesting to work out a similar problem as our query:

$$i^{\star}(\mu) = \{i \in [m] : \mu_i \geq \bar{\mu} + k\sigma\}$$

Where $k \in \mathbb{R}$, $\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\mu_i - \bar{\mu})^2}$ The derivation for this problem is challenging. We refer to our efforts in Appendix: 10.8. As mentioned in Appendix, we end up with a term we do not yet know how to analytically work out. We believe that this extension of the objective is more in line with real applications. It is very similar to the objective of finding the K-best arms that is worked out in paper: [18] but with the adjustment that we do not have to have a reasonable value for K when we use it.

Chapter 7

Discussion

From Figure 3.1 we can see that the number of subsets we have to check, until we find the correct subset J, scales less than linear with the number of arms. However iteratively having to verify if all the KKT-conditions are met is still costly. The computational cost of the algorithm is inordinate because of this. Especially for larger values of m. So to find a proof such that we can immediately find the correct subset J is of importance.

In Figure 5.3 we see the relationship between the relative change in weights over time. We can immediately see that the significant changes in weights occur during the moments when the change in set J occurs (the one involving the arms in question). This begs the question whether it is necessary to recalculate the weights at every iteration, instead of only performing this procedure when there is a change in the set J.

The extent to which we can use our expression is limited. As can be seen from the equation 3.38, that is used to find the weights for the means in set J. We can see that when we would have an arm with a reward that is precisely equal to the average of all rewards, the weight is negative, which is infeasible. This limitation is purely theoretical since the probability for a continuous function to be precisely equal to this is zero.

The numerical experiments were acquired using D-tracking. C-tracking was used to assess the whether the implementation was correct and if results where considerably different. It is acknowledged that D-tracking may not converge, as previously observed in the literature [5]. However, no evidence was found to suggest that this phenomenon occurs in the context of our algorithm.

The results that we find in Figure 5.1 are as expected. We can visually see a general downward trend for the sample complexity when we increase δ. This relationship seems obvious since we can stop earlier when we relax the probability that we do not give the correct answer. Additionally, we indicate with the red line: $T(\mu)\log(\frac{1}{\delta})$. As $\delta \to 0$ we know that this must be lower than the sample complexity if the algorithm is δ-PAC. From the Figure we can see that this holds. If we look closely, we can see that for lower values of δ the sample complexity is gradually converging.

From the Figure 5.2 we can see the relationship between the variance of the set of random numbers μ and the sample complexity. The result is as expected. We could say that it is generally more difficult to distinguish between numbers that are close together. But this is not the whole picture. A smaller variance does not immediately mean that the differences

from the mean are also closer, which would make it harder to identify whether it is above or below the average..

In Figure 5.4 we have compared the performance for our model with the threshold bandit model. It should be stressed that the comparison is not an entirely accurate representation of reality, as it is based on the assumption that the final mean of all rewards has already been determined. Furthermore, the threshold is not dependent on the preliminary estimates of μ. This means that the weights for the thresholding bandit problem do not have the problem that the preliminary estimate of the mean could be significantly different from the actual mean. This could also explain the difference in the performance. Not only of the average stopping time but also on the variance. We see that for the thresholding bandit model, the variance is significantly lower. The equation for the optimal weights used to model the threshold bandit does not incorporate the fact that the threshold is dynamic. Therefore simply making the threshold equal to the current estimation of the mean of all the rewards would not be accurate.

In Figure 5.5 we have compared the performance of our model with the Best arm identification for chosen μ such that the objectives align. We can see from the Figure that the Best Arm Identification algorithm outperforms our model throughout the simulation. For lower values of δ the 95% interval overlaps more heavily, indicating that it might be harder to differentiate between the performances for both the algorithms. We believe that the performance difference could be entirely explained by the simple fact that Best Arm Identification is more equipped to perform the chosen objective.

We believe that the benefits by calculating the weights using the approach covered in [12] do not outweigh the loss of precision. The benefit of using this method is that we reduce the complexity of the inner part of the optimization problem. However the inner part of the optimization is the minimization over all arms, which is linear in the number of arms. We also do not solve the problem relating to finding the correct set of J.

Considerable effort has been invested in attempting to prove a method to analytically find the set J. However, all such attempts have ultimately failed. The comparison between the values of sets J and I has not yielded any conclusive results. Furthermore, a more detailed examination of the KKT conditions did not lead to the discovery of any new KKT conditions. We did find properties of the set J that have been empirically tested to be true, however the converse is not true. An example of this is:

$$\sum_{j \in J} (\bar{\mu} - \mu_j)^2 \geq \sum_{i \in I} (\bar{\mu} - \mu_i)^2$$

However there could be multiple sets for which this is the case.

As has been covered in Section 3.1.4, the conjecture is that the arm with the smallest distance. During all the necessary numerical experiments, as well as all other computations we did not find a singular case for which the arm with the smallest distance to the mean was not in set J. However, without a formal proof we cannot rule the possibility out.

Chapter 8

Conclusion

In this thesis we covered the sample complexity of the dynamic-threshold problem of finding all arms with a reward that is higher than the arithmetic mean of all rewards. For this analyse we made the assumption that the rewards are sampled according to a normal distribution that is governed by one parameter (the mean). We found that the sample complexity (and the KKT-conditions) are dependent on certain subsets I and J. We did not find an analytical method to find these. However we found an accelerated method to find the appropriate configurations of set I and J. We compared the sample complexity with other algorithms that have the same answer. We found that our algorithm has a higher stopping time than the Best Arm Identification algorithm and Thresholding bandit algorithm for instances of μ for which the correct answer was identical. All algorithms are implemented using the Track-and-Stop strategy. We also verified the δ-PAC property for our algorithm, FindingMeano.

Chapter 9

Acknowledgement

I would like to express my sincere gratitude to my supervisor Wouter Koolen for his support and essential feedback for this thesis. Additionally, I would like to thank Dr. Rianne de

Heide and Dr. Antonios Antoniadis for being part of the assessment committee.

Chapter 10

Appendix

10.1 Lower bound

Proof. For this proof we closely follow the reasoning as provided in the paper [9]. Using the log-likelihood ratio L_t and using $(Y_{a,s})$ which is the sequence of outcomes from arm a at time s. The log-likelihood ratio is defined as

$$L_t = \sum_{a=1}^{m} \sum_{s=1}^{t} \mathbb{1}\{A_s = a\} \log\left(\frac{f_t(Y_{a,s})}{f'_t(Y_{a,s})}\right)$$

$$= \sum_{a=1}^{m} \sum_{s=1}^{N_a(t)} log\left(\frac{f_t(Y_{a,s})}{f'_t(Y_{a,s})}\right)$$

If we than use the the fact that:

$$E_\nu\left[\log\left(\frac{f_t(Y_{a,s})}{f'_t(Y_{a,s})}\right)\right] = \mathrm{kl}(\mu_a, \lambda_a)$$

we than have

$$E_\nu[L_\sigma] = \sum_{a=1}^{m} E_\nu[N_a(\tau)] kl(\mu_a, \lambda_a) \geq kl(P_\mu(\epsilon), P_\lambda(\epsilon))$$

By 1, ϵ could be any event. If we choose the appropriate events

$$P_\mu(\epsilon) = P_\mu(\hat{I}_{\tau_\delta} \neq a^\star) = \delta \qquad P_\lambda(\epsilon) = P_\lambda(\hat{I}_{\tau_\delta} \neq a^\star) = 1 - \delta$$

$$E_\mu[\tau_\delta] \sum_{a=1}^{m} \frac{E_\mu[N_a(\tau_\delta)]}{E_\mu[\tau_\delta]} kl(\mu_a, \lambda_a) \geq kl(\delta, 1-\delta)$$

we know the following:

$$E_\mu[\tau_\delta] = \sum_{a=1}^{m} E_\mu[N_a[\tau_\delta]]$$

So by dividing them, as done in the previous equation, we essentially have a parameter that is the allocation of all arms. This is the parameter w_a, better known as the weight of the arm a. We use the notation $\sum_m = \{w \in \mathbb{R}^+ : w_1 + \cdots + w_m = 1\}$ to express these weights. At last, [9] makes two observations:

1. For each arm there exists a lower bound on the expected number of draws
2. this must hold for all δ-PAC problems

The first distinction gives rises to the idea to seek a set of alternatives $\lambda \in Alt(\mu)$ to minimize over, this such that we do not have to find separately the alternatives for each arm. Secondly the problem must hold for all δ-PAC strategies and therefore the weight of all arms is replaced by taking the supremum. If we combine these two observations we have the lower bound on the number of draws needed.

$$kl(\delta, 1-\delta) \leq E[\tau_\delta] \sup_{w \in \sum_m} \inf_{\lambda \in Alt(\mu)} \sum_{a=1}^{m} w_a d(\mu_a, \lambda_a)$$

10.2 Case: $|J| = m$

We start by reiterating the KKT-condition, for which we follow a different trajectory. This is complementary slackness:

$$\gamma_i(\psi - c_i) = 0$$

If $|\mathbf{J}| = \mathbf{m}$:

$$\psi = c_i \quad \forall i \in [m]$$

$$\psi = \frac{w_i m^2 (\bar{\mu} - \mu)^2}{2(m(m-2) + w_i A)}$$

$$w_i = \frac{2\psi m(m-2)}{m^2(\bar{\mu} - \mu_j)^2 - 2\psi A}$$

We first express A

$$A = \sum_{i=1}^{m} \frac{1}{w_i}$$

$$= \sum_{i=1}^{m} \frac{m^2(\bar{\mu} - \mu_j)^2 - 2\psi A}{2\psi m(m-2)}$$

$$= \sum_{i=1}^{m} \frac{m^2(\bar{\mu} - \mu_j)^2}{2\psi m(m-2)} - \frac{2\psi A}{2\psi m(m-2)}$$

We can reconfigure all the terms, and we end up with

$$A = \frac{m}{m-1} \frac{1}{2\psi} \sum_{i=1}^{m} (\bar{\mu} - \mu_i)^2$$

If we fill this into our original expression we get:

$$w_i = \frac{2\psi m(m-2)}{m^2(\bar{\mu} - \mu_i)^2 - \frac{m}{m-1} \sum_{i=1}^{m} (\bar{\mu} - \mu_i)^2} \tag{10.1}$$

If we then use the expression $\sum_i w_i = 1$, we find the appropriate ψ to be

$$\psi = \frac{\frac{1}{2m(m-2)}}{\frac{1}{m^2(\bar{\mu}-\mu_i)^2 - \frac{m}{m-1}\sum_{i=1}^{m}(\bar{\mu}-\mu_i)^2}}$$

We can use this expression for ψ in our original expression for the weights to get the result as has been established.

10.3 Reasoning heuristic

We explain the proof for the fact that the set J is an interval by a proof by contradiction. We assume that at least the arm that has lowest distance to the mean is in J. We try to prove that adding from lowest to highest will eventually result in the correct set J, so without every there being gaps between consecutive arms (in terms of distance to mean).

Proof by Contradiction: We sort all arms according to their respective squared distance to the mean. Without loss of generality we label them $\mu_1 < \mu_2 < \cdots < \mu_m$. We are given a set J where arm i,j and k are included in set J. We have $\mu_i < \mu_j < \mu_k$. First we assume the contrary: $\mu_i, \mu_k \in J, \mu_j \notin J$.

We know that $\forall \mu_i \in J, c_i = \psi$

$$c_i = c_k \quad , c_j \neq c_i$$

Given the definition of c_i:

$$c_i = \frac{w_i m^2(\bar{\mu}-\mu_i)^2}{2(m(m-2)+w_i A)},$$

$$c_j = \frac{w_j m^2(\bar{\mu}-\mu_j)^2}{2(m(m-2)+w_j A)},$$

$$c_k = \frac{w_k m^2(\bar{\mu}-\mu_k)^2}{2(m(m-2)+w_k A)},$$

Expressing c_i in terms of another arm
Given that $c_i = c_k$ and $c_j \neq c_i$, let's explore these relationships.
We introduce $\epsilon < 1$ such that:

$$(\bar{\mu}-\mu_i)^2 = \epsilon(\bar{\mu}-\mu_j)^2$$

Then we have

$$\frac{w_i m^2(\bar{\mu}-\mu_i)^2}{2m(m-2)+w_i A}$$

$$= \frac{w_i m^2(\bar{\mu}-\mu_j)^2\epsilon}{2m(m-2)+w_i A}$$

so we have

$$\frac{w_i m^2(\bar{\mu}-\mu_2)^2\epsilon}{2m(m-2)+w_i A} \neq \frac{w_j m^2(\bar{\mu}-\mu_j)^2}{2(m(m-2)+w_j A)}$$

We can rearrange this accordingly:

$$\frac{w_i}{w_j}\frac{m(m-2)+w_jA}{m(m-2)+w_iA} \neq \frac{(\bar{\mu}-\mu_j)^2}{(\bar{\mu}-\mu_j)^2\epsilon}$$

We know that $\epsilon < 1$, combining the two makes:

$$\frac{w_i(m(m-2)+w_jA)}{w_j(m(m-2)+w_iA)} < 1$$

We assume $m > 2$ so from this we know that $w_i < w_j$.

We can repeat this process again, but then transform:

$$(\bar{\mu}-\mu_k)^2 = \epsilon(\bar{\mu}-\mu_j)^2$$

for $\epsilon > 1$. If we repeat the steps we get the conclusion $w_j < w_k$. If we analyze the dependency of c_i (same dependency for j and k) on w:

$$c_i = \frac{w_i m^2(\bar{\mu}-\mu_i)^2}{2(m(m-2)+w_iA)}$$

$$= \frac{m^2(\bar{\mu}-\mu_i)^2}{2(\frac{m(m-2)}{w_i}+A}$$

We can see that an increase of w_i, as for an increase in $(\bar{\mu}-\mu_i)^2$ both lead to a higher c_i. And $w_k > w_i \quad (\bar{\mu}-\mu_k)^2 > (\bar{\mu}-\mu_i)^2$, which in term leads to $c_k > c_i$. This contradicts the assumption that $c_i = c_k$. **Conclusion:** This contradiction implies that our initial assumption must be false. Therefore, if arm i is in J and arm j is not, then arm k cannot be in J. □

10.4 Gradient Ascent

$$F(w,\mu) = \inf_{\lambda \in Alt(\mu)} \sum_{i=1}^{m} w_i d(\mu_i, \lambda_i)$$

$$= \min_i \frac{w_i m^2(\bar{\mu}-\mu_i)^2}{2(m(m-2)+w_iA)}$$

This function is directly dependent on w_i but also indirectly because of $A = \sum_{i=1}^{m} \frac{1}{w_i}$. To compute the gradient with respect to w_i we use the product rule.

$$\frac{\delta F}{\delta w_i} = \frac{\delta F}{\delta w}\frac{\delta A}{\delta W}$$

$$= -\frac{1}{w_i^2}\frac{(\bar{\mu}-\mu_i)^2 m^3(m-2)}{2(m(m-2)+w_iA)^2}$$

so

$$\nabla F(w,\mu) = \min_i \frac{1}{w_i^2}\frac{(\bar{\mu}-\mu_i)^2 m^3(m-2)}{2(m(m-2)+w_iA)^2}$$

10.5 Alt

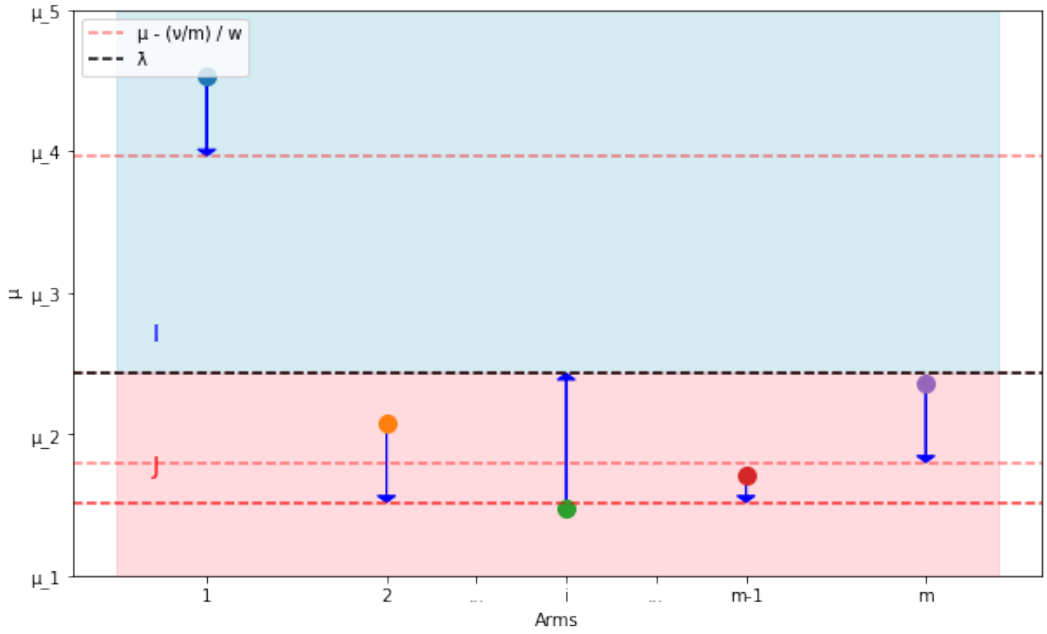

Figure depicting the effect of $\lambda_i = \bar{\lambda}$ and $\lambda_j = \mu_j - \frac{\frac{1}{m}\nu}{w_j}$.

10.6 Kullback-Leibler divergence

Let $P = \mathcal{N}(\mu, \sigma_1^2)$ and $Q = \mathcal{N}(\lambda, \sigma_2^2)$ be two normal distributions with probability density functions:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma_1^2}\right)$$

$$q(x) = \frac{1}{\sqrt{2\pi\sigma_2^2}} \exp\left(-\frac{(x-\lambda)^2}{2\sigma_2^2}\right)$$

The Kullback-Leibler (KL) divergence from P to Q is defined as:

$$D_{KL}(P||Q) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)}\, dx$$

We first simplify the logarithmic part of the integral.

$$\log \frac{p(x)}{q(x)} = \log \left[\frac{\frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma_1^2}\right)}{\frac{1}{\sqrt{2\pi\sigma_2^2}} \exp\left(-\frac{(x-\lambda)^2}{2\sigma_2^2}\right)} \right]$$

Simplifying the logarithm:

$$\log \frac{p(x)}{q(x)} = \log \frac{\sigma_2}{\sigma_1} + \frac{(x-\mu)^2}{2\sigma_1^2} - \frac{(x-\lambda)^2}{2\sigma_2^2}$$

Thus, the KL divergence becomes:

$$D_{KL}(P||Q) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma_1^2}\right) \left[\log \frac{\sigma_2}{\sigma_1} + \frac{(x-\mu)^2}{2\sigma_1^2} - \frac{(x-\lambda)^2}{2\sigma_2^2}\right] dx$$

This can be split into three separate integrals:

First part

$$\underbrace{\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma_1^2}\right)}_{\text{Integral from } -\infty \text{ to } \infty \text{ of } p(x)=1} \underbrace{\left[\log \frac{\sigma_2}{\sigma_1}\right]}_{\text{Constant}}$$

$$= \log \frac{\sigma_2}{\sigma_1}$$

Second part

$$\underbrace{\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma_1^2}\right) \frac{(x-\mu)^2}{2\sigma_1^2} dx}_{\text{Expectation of } \frac{(x-\mu)^2}{2\sigma_1^2}}$$

This integral represents the expectation of $\frac{(x-\mu)^2}{2\sigma_1^2}$):

$$\mathbb{E}\left[\frac{(x-\mu)^2}{2\sigma_1^2}\right] = \underbrace{\frac{1}{2\sigma_1^2}}_{\text{Constant}} \underbrace{\mathbb{E}_P\left[(x-\mu)^2\right]}_{\sigma_1^2 \text{ (variance of } P(x))}$$

Since $\mathbb{E}_P\left[(x-\mu)^2\right] = \sigma_1^2$ (the variance of the distribution $P(x)$):

$$\mathbb{E}_P\left[\frac{(x-\mu)^2}{2\sigma_1^2}\right] = \underbrace{\frac{1}{2\sigma_1^2} \cdot \sigma_1^2}_{\frac{1}{2}} = \frac{1}{2}$$

Thus, the second integral simplifies to $\frac{1}{2}$.

Third part

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma_1^2}\right) \frac{(x-\lambda)^2}{2\sigma_2^2} dx$$

Expanding $(x-\lambda)^2$ as:

$$(x-\lambda)^2 = (x-\mu+\mu-\lambda)^2 = (x-\mu)^2 + 2(x-\mu)(\mu-\lambda) + (\mu-\lambda)^2$$

The integral becomes:

$$\frac{1}{2\sigma_2^2}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi\sigma_1^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma_1^2}\right)\left[(x-\mu)^2+2(x-\mu)(\mu-\lambda)+(\mu-\lambda)^2\right]dx$$

First part

$$\frac{1}{2\sigma_2^2}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi\sigma_1^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma_1^2}\right)(x-\mu)^2\,dx=\frac{\sigma_1^2}{2\sigma_2^2}$$

As again $\mathbb{E}_P\left[(x-\mu)^2\right]=\sigma_1^2$

Second part

$$\frac{1}{2\sigma_2^2}\cdot 2(\mu-\lambda)\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi\sigma_1^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma_1^2}\right)(x-\mu)\,dx=\mathbb{E}[X-\mu]=\mathbb{E}[X]-\mu=0$$

Third part

$$\frac{1}{2\sigma_2^2}(\mu-\lambda)^2\underbrace{\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi\sigma_1^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma_1^2}\right)dx}_{\text{Integral of } p(x)=1}=\frac{(\mu-\lambda)^2}{2\sigma_2^2}$$

Combining these, the third integral evaluates to:

$$\frac{\sigma_1^2}{2\sigma_2^2}+\frac{(\mu-\lambda)^2}{2\sigma_2^2}$$

Final expression Substituting back into the KL divergence expression:

$$D_{KL}(P||Q)=\log\frac{\sigma_2}{\sigma_1}+\frac{\sigma_1^2+(\mu-\lambda)^2}{2\sigma_2^2}-\frac{1}{2}$$

If we use $\sigma_1=\sigma_2=1$ we have the result as we use it:

$$D_{KL}(P||Q)=\frac{(\mu-\lambda)^2}{2}$$

10.7 Sample complexity: how many arms are above a threshold

To answer the query: How many arms are above a threshold? We use the following definition for the correct answer:

$$i^\star(\mu)=\sum_{i=1}^{m}\mathbb{1}\{\mu_i\geq\alpha\}$$

We assume a normal distribution for the Kullback-Leibler divergence.

$$\inf_{\lambda\in\mathrm{Alt}(\mu)}\left(\sum_{a=1}^{m}w_a\frac{(\mu_a-\lambda_a)^2}{2}\right)$$

To elaborate on the alternative of μ we explore both the case for which the correct answer for λ lies above and below the correct answer of μ. If we set the correct number of arms in μ, above the threshold α to be equal to π, we seek: $i^\star(\lambda) = \pi + 1$ To increase the correct answer, we pick arm k such that $\mu_k < \alpha$ and we set $\lambda_k = \alpha$.

$$\inf_{\lambda \in \text{Alt}(\mu)} \left(\sum_{a=1}^{m} w_a \frac{(\mu_a - \lambda_a)^2}{2} \right) =$$

$$\min_{k:\mu_k<\alpha, \lambda_k=\alpha} \left(\sum_{a=1}^{K} w_a \frac{(\mu_a - \lambda_a)^2}{2} \right)$$

For all $\lambda_i \neq \lambda_k$ choose $\lambda_i = \mu_i$. With this relaxation we have:

$$\min_{k:\mu_k<\alpha} w_k \frac{(\mu_k - \alpha)^2}{2}$$

To calculate $i^\star(\lambda) = \pi - 1$ we pick vector k such that $\mu_k \geq \alpha$ and we set $\lambda_k < \alpha$

$$\inf_{\lambda \in \text{Alt}(\mu)} \left(\sum_{a=1}^{m} w_a \frac{(\mu_a - \lambda_a)^2}{2} \right) =$$

$$\min_{k:\mu_k\geq\alpha, \lambda_k<\alpha} \left(\sum_{a=1}^{m} w_a \frac{(\mu_a - \lambda_a)^2}{2} \right)$$

we set $\forall i \neq k : \lambda_i = \mu_i$

$$\min_{k:\mu_k\geq\alpha} w_k \frac{(\mu_k - \alpha)^2}{2}$$

We now now the correct form for when $\mu \geq \alpha$ and $\mu \leq \alpha$. We want to include those possibilities both therefore we therefore have:

$$\inf_{\lambda \in \text{Alt}(\mu)} \left(\sum_{a=1}^{m} w_a \frac{(\mu_a - \lambda_a)^2}{2} \right)$$
$$= \min_k w_k \frac{(\mu_k - \alpha)^2}{2}$$

From here we use the exact same reasoning as for problem 4.0.1 to acquire the equation for the optimal weights.

10.8 For further research

We begin by defining sets $I \in [m]$ $J \in [m] \setminus I$, we now use KKT-conditions on a generic problem description of the problem.

KKT Conditions:

$$\inf_{\lambda} \sum_{i=1}^{m} w_i \frac{(\mu_i - \lambda_i)^2}{2}$$

subject to

$$\begin{aligned}
&\lambda_i \geq \bar{\lambda} + k\sigma && \text{for all } i \in I,\\
&\lambda_j < \bar{\lambda} + k\sigma && \text{for all } j \in J,\\
&\bar{\lambda} = \frac{1}{m}\sum_{i=1}^{m} \lambda_i,\\
&\sigma = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(\bar{\lambda} - \lambda_i)^2}.
\end{aligned}$$

Lagrangian:

$$\begin{aligned}
\mathcal{L}(\lambda_i, \sigma, \bar{\lambda}, X_i, Y_j, z, \nu) = &\sum_{i=1}^{m} w_i \frac{(\mu_i - \lambda_i)^2}{2} + \sum_{i \in I} X_i(\bar{\lambda} + k\sigma - \lambda_i) + \sum_{j \in J} Y_j(\lambda_j - \bar{\lambda} - k\sigma)\\
&+ \nu\left(\bar{\lambda} - \frac{1}{m}\sum_{i=1}^{m}\lambda_i\right)\\
&+ z\left(\sigma - \sqrt{\frac{1}{m}\sum_{i=1}^{m}(\bar{\lambda} - \lambda_i)^2}\right)
\end{aligned}$$

Stationarity

$$-w_i(\mu_i - \lambda_i) - X_i - \frac{1}{m}\nu + \frac{1}{m}\frac{\lambda_i - \bar{\lambda}}{\sigma}z = 0 \quad \forall i \in I \tag{10.2}$$

$$-w_j(\mu_j - \lambda_j) + Y_j - \frac{1}{m}\nu + \frac{1}{m}\frac{\lambda_j - \bar{\lambda}}{\sigma}z = 0 \quad \forall j \in J \tag{10.3}$$

$$\sum_{i \in I} X_i - \sum_{j \in J} Y_j + \nu - \frac{1}{m}\frac{\lambda_i - \bar{\lambda}}{\sigma}z = 0 \tag{10.4}$$

$$k\sum_{i \in I} X_i - k\sum_{j \in J} Y_j + z = 0 \tag{10.5}$$

If we multiply equation 10.4 with k and insert it into equation 10.5 we can get an expression for ν.

$$\nu = z\left(\frac{\lambda_i - \bar{\lambda}}{m\sigma} + \frac{1}{k}\right)$$

Using this newly acquired equation into equations 10.2 and 10.3 gives the following relations:

$$X_i = -w_i(\mu_i - \lambda_i) + \frac{z}{m}\frac{\lambda_i - \bar{\lambda}}{m\sigma}(1 - \frac{1}{m}) - \frac{z}{km} \tag{10.6}$$

$$Y_j = w_j(\mu_j - \lambda_j) - \frac{z}{m}\frac{\lambda_j - \bar{\lambda}}{m\sigma}(1 - \frac{1}{m}) + \frac{z}{km} \tag{10.7}$$

We now simplify equation 10.5. From equations 10.7 and 10.6 we can see that all the terms have the same sign. We can make use of the fact that:

$$\sum_{i=1}^{m}(\lambda_i - \bar{\lambda}) =$$
$$\sum_{i=1}^{m}\lambda_i - \sum_{i=1}^{m}\bar{\lambda} =$$
$$\sum_{i=1}^{m}\lambda_i - \sum_{i=1}^{m}\frac{1}{m}\sum_{i}^{m}\lambda_i =$$
$$\sum_{i=1}^{m}\lambda_i - \sum_{i=1}^{m}\lambda_i = 0$$

We also see that, if we take the sum of X_i and Y_j we simplify the constant terms:

$$-\frac{|I|z}{m} - \frac{|J|z}{m} + z =$$
$$-z\frac{|I| + |J|}{m} + z = 0$$

Then equation 10.5 becomes:

$$-k\sum_{i=1} X_i - k\sum_{j} Y_j + z = 0$$
$$-k\sum_{i=1}^{m} w_i(\mu_i - \lambda_i) = 0$$
$$\sum_{i=1}^{m} w_i(\mu_i - \lambda_i) = 0 \tag{10.8}$$

This is a remarkable result, because it is the same result as we had for the problem 1.1, namely in equation 3.16. We again use the reasoning that was used to generate $\text{Alt}(\mu)$ for the problem of finding the arms with a reward that is higher than the arithmetic mean. We assume that the minimizer λ ($\inf_{\lambda \in Alt(\mu)}$) is in the subspace for an answer that is different from the correct answer by only arm. We again use complementary slackness:

$$X_i(\bar{\lambda} + k\sigma - \lambda_i) = 0 \tag{10.9}$$

$$Y_j(\lambda_j - \bar{\lambda} - k\sigma) = 0 \tag{10.10}$$

We set $Y_j = 0$ and set $\lambda_i - \bar{\lambda} - k\sigma = 0$. We than have the following equations:

$$\lambda_j = \frac{k\mu_j m^2 \sigma w_j + k(m-1)\bar{\lambda}z + m\sigma z}{k(m^2\sigma w_j + (m-1)z)} \tag{10.11}$$

$$\lambda_i = \bar{\lambda} + k\sigma \tag{10.12}$$

Using these two equations directly is not enough. We still need to workout $\bar{\lambda}$, σ and z. However this cannot be simply worked out. If we look at the equations that are given from the condition primal feasibility:

$$\bar{\lambda} = \frac{1}{m}\sum_{i=1}^{m}\lambda_i \tag{10.13}$$

$$\sigma = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(\bar{\lambda} - \lambda_i)^2} \tag{10.14}$$

We see for both equations that we have to use the summation of λ_i and λ_j. Taking the sum of λ_j complicates finding expressions we can analytically solve. We have not found a way to do this. What we can do, is express z using the following relations:

$$\sum_{i=1}^{m} w_i(\mu_i - \lambda_i) = 0$$

$$Y_j = w_j(\mu_j - \lambda_j) - \frac{z}{m}\frac{\lambda_j - \bar{\lambda}}{m\sigma}(1 - \frac{1}{m}) + \frac{z}{km}$$

If we manipulate Y_j such that we solve for $w_j(\mu_j - \lambda_j)$, take into account we set Y_j equal to zero we get and use the definition $\lambda_i = \bar{\lambda} + k\sigma$:

$$\begin{aligned}
&\sum_{i}^{m} w_i(\mu_i - \lambda_i) = 0 \\
&= \sum_{j} w_j(\mu_j - \lambda_j) + w_i(\mu_i - \lambda_i) \\
&= w_i(\mu_i - \lambda_i) + \sum_{j\in[m]\setminus i}^{m} \frac{z}{m}\frac{\lambda_j - \bar{\lambda}}{\sigma}(1 - \frac{1}{m}) - \frac{z(m-1)}{km}
\end{aligned}$$

We can manipulate the summation of λ_j and use our definition of λ_i such that

$$\begin{aligned}
&\sum_{j\in[m]\setminus i} (\lambda_j - \bar{\lambda}) \\
&= \sum_{i}^{m} \lambda_i - \lambda_i - (m-1)\bar{\lambda} \\
&= m\bar{\lambda} - \bar{\lambda} - k\sigma - (m-1)\bar{\lambda} \\
&= -k\sigma
\end{aligned}$$

Using this we can obtain an expression for z, in terms of $\bar{\lambda}$ and σ.

$$z = \frac{km^2 w_i(\mu_i - \bar{\lambda} - k\sigma)}{(m-1)(k^2 - m)} \tag{10.15}$$

Although the expression z does not simplify 10.11 we can see some restrictions. $k^2 \neq m$ is not allowed for example.

Bibliography

[1] Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. J. Mach. Learn. Res., 3(null):397–422, mar 2003.

[2] Maryam Aziz, Emilie Kaufmann, and Marie Karelle Riviere. On multi-armed bandit designs for dose-finding clinical trials. Journal of Machine Learning Research, 22:1–38, 2021. arXiv:1903.07082.

[3] Antoine Barrier. Contributions to a theory of pure exploration in sequential statistics. URL: https://theses.hal.science/tel-04192097.

[4] Antoine Barrier. Contributions to a Theory of Pure Exploration in Sequential Statistics To cite this version : HAL Id : tel-04192097 Discipline : Mathématiques Contributions à une théorie de l ' exploration pure en statistique séquentielle. 2023.

[5] Rémy Degenne and Wouter M. Koolen. Pure exploration with multiple correct answers. Advances in Neural Information Processing Systems, 32:1–31, 2019.

[6] Aurélien Garivier and Olivier Cappé. The KL-UCB algorithm for bounded stochastic bandits and beyond. Journal of Machine Learning Research, 19:359–376, 2011. arXiv:1102.2490.

[7] Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. Journal of Machine Learning Research, 49:998–1027, 2016.

[8] Kevin Jamieson and Ameet Talwalkar. Non-stochastic best arm identification and hyperparameter optimization. Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, AISTATS 2016, pages 240–248, 2016. arXiv:1502.07943.

[9] Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. Journal of Machine Learning Research, 17:1–42, 2016.

[10] Emilie Kaufmann and Wouter Koolen. Mixture martingales revisited with applications to sequential tests and confidence intervals. 11 2018. URL: http://arxiv.org/abs/1811.11419.

[11] Blake Mason, Lalit Jain, Ardhendu Tripathy, and Robert Nowak. Finding all e-good arms in stochastic bandits. Advances in Neural Information Processing Systems, 2020-Decem, 2020.

[12] Pierre Ménard. Gradient Ascent for Active Exploration in Bandit Problems. pages 1–21, 2019. URL: http://arxiv.org/abs/1905.08165, arXiv:1905.08165.

[13] Kanishka Misra, Eric Schwartz, and Jacob Abernethy. Dynamic online pricing with incomplete information using multiarmed bandit experiments. Marketing Science, 38, 03 2019. doi:10.1287/mksc.2018.1129.

[14] Chao Qin and Wei You. Dual-directed algorithm design for efficient pure exploration. 10 2023. URL: http://arxiv.org/abs/2310.19319.

[15] Stephen Boyd Lieven Vandenberghe. Convex Optimization. 2013.

[16] WILLIAM R THOMPSON. ON THE LIKELIHOOD THAT ONE UNKNOWN PROBABILITY EXCEEDS ANOTHER IN VIEW OF THE EVIDENCE OF TWO SAMPLES. Biometrika, 25(3-4):285–294, 12 1933. arXiv:https://academic.oup.com/biomet/article-pdf/25/3-4/285/513725/25-3-4-285.pdf, doi:10.1093/biomet/25.3-4.285.