Ministry of Science and Higher Education of Russia
Moscow Institute of Physics and Technology (National Research University)
Landau Phystech-School of Physics & Research
Department of Fundamental Interactions and Cosmology

Fisher Information-driven Optimal Experiment Design

A thesis submitted for the degree of
Bachelor of Science

Author
Vladimir Palmin

Thesis Supervisor
Alexander Nozik

Moscow
2021

Abstract

The Fisher information is a powerful technique. It can be used to solve such problems as experiment designing. Since the costs of experiments are increasing, this problem is becoming more important. So this thesis provides the solution to optimise a time strategy of an experiment that measures a spectrum. This optimisation results in the sufficient reduction of error of a target parameter, e.g., a particle mass. This reduction might be explained in terms of correlation reducing that also occurs. Because reducing the correlation with the parameter represented systematic errors means that the influence of systematic errors also decreases. This thesis also provides the Bayesian view on the optimisation. It was shown that prior information on a nuisance parameter might reduce a target parameter error via correlation reduction. Thus, this method may provide a great opportunity for further experiments.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Alexander Nozik, for his invaluable supervision, continuous support, and patience during my study. Additionally, I would like to thank Mikhail Zelenyi and Roland Grinis for their help and advice. My appreciation also goes out to my family and friends for their encouragement and support all through my studies.

Contents

1 Introduction

The development of physics is slowing down by more expensive experiments, both in time and money. Such an extensive way of development leads to that there are fewer and fewer ways to measure everything in all directions. Therefore, the design of experiments should be more elaborate. On the other hand, statistics forbids varying data to optimize results, or in other words, one can not simply select the data that are the most preferable to fit a model. Thus, a new proven methodology is required. This methodology should help redefine the way experiments are designed.

The background to this work is the phenomenon discovered in the Troitsk nu-mass experiment [1]. It consists in the fact that when the neutrino mass is varied in a model, there is a sharp decrease in the systematic error at a certain point (Fig. 1). So it is possible to reduce systematic errors only by varying the model, one might even say that this improvement comes out of nowhere. Although such a phenomenon is quite interesting, as described above, a data-independent methodology is needed to explain it.

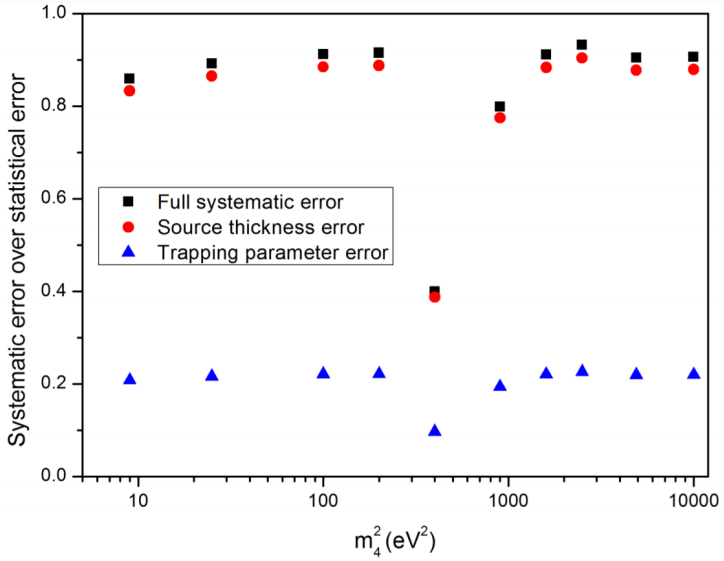


Figure 1: The systematic error of the fraction of the heavy neutrino in the electron neutrino, U^2, in Feldman and Cousins approach from different sources divided by statistical error for different masses. And m_4 is the mass of the heavy neutrino eigenstate.

The approach used to obtain Fig. 1 is described in [2]. The article and the approach are devoted to the problem of constructing confidence intervals and their variations. Feldman and Cousins approach solves some problems with the traditional classical intervals, such as non-physical intervals or intervals with empty sets, and gives more correct and stable results.

The mathematics used to solve such problems is well known - the Fisher information. Nevertheless, its application in physics, including the development of special techniques, is not widespread. The Fisher information can be used both for the design of experiment equipment and for a strategy of measurements. A strategy means that, for example, how much time is used to measure a part of the spectrum. In this work, this particular

example is considered. Therefore, requirements for an experiment can already be given. An experiment to which the method described in this thesis will be applied should measure a spectrum and make it possible to measure one part of the spectrum longer than another one.

The first chapter of this thesis presents the mathematical formulation of the problem and the necessary derivations, as well as the introduction to the concept of the Fisher information. This is followed by a description of the development of a numerical solution to the problem using the tritium spectrum as an example, and its application to the spectrum of axions in the second chapter. Finally, in the third chapter, the modified method is demonstrated in the context of Bayesian statistics, i.e. using prior information on systematic uncertainties and a target parameter, is demonstrated.

2 Fisher Information

This chapter is dedicated to the derivation of the key formula on which the method is based.

2.1 Basic Concepts

2.1.1 Likelihood

The Fisher information is an advanced technique, so basic terms and concepts will be discussed for clarification. We want to formulate an experiment mathematically, so we first use theorems from theoretical physics that describe the phenomenon measured in the experiment. Next, we also want to consider uncertainties into consideration. That results in a statistical model. And a statistical model is typically defined by a function $f(x_i|\theta)$ that represents how a parameter θ is functionally related to possible outcomes of a random variable x_i.

To illustrate, consider the most famous example − the coin toss model. An experimenter tosses a coin and records the outcome, heads or tails. There are only two possible outcomes and they can be denoted by zeros and ones. The function that describes such a model is called the Bernoulli distribution

$$f(x_i|\theta) = \theta^{x_i}(1-\theta)^{1-x_i} \tag{1}$$

Where $x_i \in \{0, 1\}$ and $\theta \in (0, 1)$, since θ represents a probability that $x_i = 1$. Probability is discussed in detail in Chapter Four. Thus, if a coin is fair, then θ is equal to 0.5. If θ is known, fixing it in the functional relationship f formally gives a function $p_\theta(x_i) = f(x_i|\theta)$, which in mathematics is called a probability density function. In physics, however, it is referred to a probability distribution. The ambiguity is that a probability distribution is the other name of a cumulative distribution function in mathematics. The difference is that a cumulative distribution function is the area under the probability density function from minus infinity to x for a scalar continuous distribution. The thesis has application in physics, so we use the 'physics' term. Moreover, $p_\theta(x_i) = P(X = x_i|\theta)$ can be considered as a data generative device.

In general, experiments consist of n trials that yield a possible set of outcomes. These n random variables are typically assumed to be independent and identically distributed (i.i.d.). Identically distributed means that each of these n random variables is determined by one and the same θ, while independence implies that the joint distribution of all these n random variables is simultaneously given by a product, i.e.

$$L(\theta|x) = f(x|\theta) = f(x_1|\theta) \times ... \times f(x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta) \tag{2}$$

Where $f(x|\theta)$ yields the joint distribution and $L(\theta|x)$ is referred to a likelihood func-

tion. One could think that we just relabeled $f(x|\theta)$ as the likelihood function. However, the reality is quite different. For the distribution f, θ is fixed and focused on an ever-changing x. For the likelihood function, on the other hand, a sample point x is a constant and a parameter θ is varying over the entire range of possible parameter values.

The likelihood function plays a fundamental role in statistics. It consists of one of the most powerful techniques, the maximum likelihood estimator (MLE), which is discussed in detail in Chapter Four. The MLE provides estimates that are unbiased, consistent, and efficient. Unbiased means that the bias between the expected value of this estimate and the true value of the parameter is zero. A consistent estimator implies that the resulting sequence of estimates converges in probability to the true value with an increase in the number of experiments. And efficient estimates have the lowest variance among unbiased estimators. For more details and information on likelihoods, see the article [3].

To calculate the MLE, it is easier to work with the log-likelihood than with the likelihood, since the product is harder to calculate than the sum. For example, we proceed in the calculations for the Bernoulli distribution:

$$\log L(\theta|x) = \log \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} = \log\theta \sum_{i=1}^{n} x_i + \log(1-\theta)\sum_{i=1}^{n}(1-x_i) \tag{3}$$

We want to find the maximum hence we should calculate the first derivative and set it equal to zero:

$$\frac{\partial}{\partial\theta}\log L(\theta|x) = \frac{\sum\limits_{i=1}^{n} x_i}{\theta} - \frac{n - \sum\limits_{i=1}^{n} x_i}{1-\theta} = 0 \tag{4}$$

As a result, the MLE equals to $\theta_{MLE} = \frac{1}{n}\sum\limits_{i=1}^{n} x_i$. To be sure that it is the maximum and not the minimum, one could calculate the second derivative. If this is negative at the optimal point, θ_{MLE} is the maximum. Indeed,

$$\frac{\partial^2}{\partial\theta^2}\log L(\theta|x) = -\frac{\sum\limits_{i=1}^{n} x_i}{(\frac{1}{n}\sum\limits_{i=1}^{n} x_i)^2} - \frac{n - \sum\limits_{i=1}^{n} x_i}{(1-\frac{1}{n}\sum\limits_{i=1}^{n} x_i)^2} = -\frac{n^2}{\sum\limits_{i=1}^{n} x_i} - \frac{n^2}{n - \sum\limits_{i=1}^{n} x_i} < 0 \tag{5}$$

2.1.2 Covariance Matrix

As described above, the task is to estimate parameters. So we need some object that describes the parameters as part of a whole. Such an object is called a covariance, a measure of the joint variability of two random variables. And a covariance matrix is a square matrix that gives the covariance between each pair of elements:

$$\Sigma_{ij} = \mathrm{cov}(X_i, X_j) = E\left(X_i - E(X_i)\right) E\left(X_j - E(X_j)\right) \tag{6}$$

Thus, the entries on the diagonal of the Σ are the variances of each element and off-diagonal elements are Pearson correlation coefficients

$$\sigma_{X_i}^2 = \text{var}(X_i) = \text{cov}(X_i, X_i), \quad \text{corr}(X_i, X_j) = \frac{\text{cov}(X_i, X_j)}{\sigma_{X_i}\sigma_{X_j}} \tag{7}$$

As a result, a covariance matrix can be seen as a generalization of scalar-valued variance to higher dimensions. For example, it is used in the multivariate normal distribution in the same way as the variance in the one-dimensional normal distribution.

2.2 Definition

The Fisher information is the core of experimental design, since it characterizes the amount of information contained in the data X about an unknown parameter θ. To understand this, one could think of likelihoods of some data samples. If the likelihoods were fairly flat on average, it would be difficult to estimate the desired parameter. Since it would not be clear which one to choose, they would all be almost the same. If, on the other hand, the likelihoods had a strong peak, it would be much easier to choose because the best estimate would be located in a small area around the peak. The example of one likelihood of the Bernoulli distribution for two different parameters θ is shown in Figure 2. It is seen that the peak is narrower for the right figure than for the other one. And if one draws the derivatives of several likelihoods, then their variation in the vicinity of the true value will be larger for the larger θ.

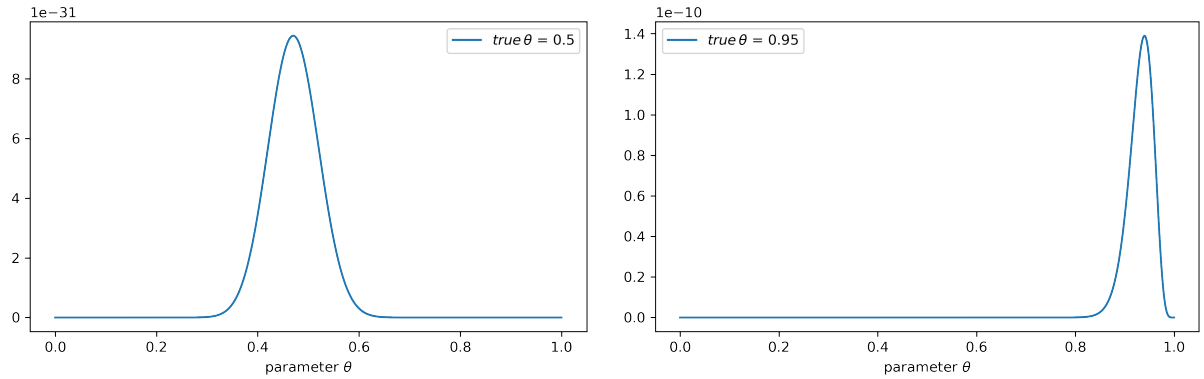


Figure 2: The likelihoods of the Bernoulli distribution for the true values $\theta = 0.5$ and $\theta = 0.95$.

This reasoning leads to several different but identical ways of defying the Fisher information. One of them is to use the expected value of the partial derivative squared with respect to θ of the natural logarithm of the likelihood function:

$$\mathcal{I}(\theta) = E\left[\left(\frac{\partial}{\partial\theta} log L(\theta|X)\right)^2 \middle| \theta\right] \tag{8}$$

The derivative squared because, under certain regularity conditions, which hold for most well-behaved distributions, the expected value evaluated at the true parameter value

θ is 0:

$$\int \frac{\partial log P(x|\theta)}{\partial \theta} P(x|\theta) dx = \int \frac{\frac{\partial}{\partial \theta} P(x|\theta)}{P(x|\theta)} P(x|\theta) dx = \frac{\partial}{\partial \theta} \int P(x|\theta) dx = \frac{\partial}{\partial \theta} 1 = 0 \qquad (9)$$

And $P(\theta|x)$ is the probability distribution. Equation (8) can be generalized to the case when θ is a vector of parameters then the Fisher information takes the form of a matrix called the Fisher Information Matrix (FIM)

$$[\mathcal{I}(\theta)]_{ij} = E\left[\left(\frac{\partial}{\partial \theta_i} log L(X;\theta)\right)\left(\frac{\partial}{\partial \theta_j} log L(X;\theta)\right)\middle|\theta\right] \qquad (10)$$

2.3 Examples and Common Strategies

To understand a concept better, one could consider examples of using this concept. Some of the most interesting use of the Fisher information is below.

2.3.1 Examples

The Fisher information is a powerful tool, so it is used in various fields. And designing an experiment involves two things: a strategy and a setup. In this work, the Fisher information is used to optimize a strategy, but it can also be used to optimize a setup. For example, it was used to prove that interferometers with more compact instantaneous beam patterns are more sensitive since they extract more spatial information from each detected photon in the work [4].

This technique is as well used in machine learning because accurate parameter estimates are often not a priority since they have no direct physical meaning. Instead, one would like to minimize the uncertainty in the model predictions for several quantities of interest. So, this approach is used to optimize the experimental design for source localization in an uncertain ocean environment in the paper [5].

Deep convolutional neural networks (CNNs) are a promising area. However, training a generalizable CNN requires a large amount of training data, which is difficult, expensive, and time-consuming to obtain in medical settings. So, the Fisher information has been used to overcome such problems in the study of the brain [6]. The paper is the first example of Fisher information-based active learning [7] applied to CNN models, since the significantly large parameter space of the CNN models leads to very large Fisher information matrices that are difficult to form and manipulate. As a result, an approximation of the Fisher information based on an implicit re-parameterization of the model helped outperform competing methods in improving the performance of the model after labeling a very small portion of the target data set in transfer learning [8] scenarios.

The Fisher information can also be used to evaluate the accuracy of estimates. For instance, in the article [9], application to parameter estimation in single-molecule microscopy

is considered. Fisher information has been used to determine the limit of accuracy with which parameters can be estimated. It serves as a benchmark against which the standard deviation of a particular estimator can be evaluated, indicating how much room for improvement might exist. It relates to the Cramér-Rao lower bound described below.

This concept is also used in finance too. In the paper [10], the Fisher information is used to demonstrate how constraints based on knowledge of system data can be used to construct probability laws. The Fisher information is used to formulate the equilibrium and then determine the dynamic laws of the economic system. In this way, the dynamic laws arise from an analysis of the flow of information about the investment value and the investment opportunities that produce these flows engender. Or the Fisher information might be used to derive the probability distribution function for prices in financial markets, as shown in [11].

However, there is another way to optimize a design. It is Bayesian optimization, which was used, for example, in optimising the active muon shield for the SHiP experiment at CERN [12]. But the Fisher information has a stronger mathematical basis. Moreover, Bayesian optimization solves the global optimization of black-box and expensive to evaluate functions such as deep neural networks. So it is not the best choice for other problems. Moreover, Bayesian and Fisher information-driven strategy optimization may be two sides of the same coin.

2.3.2 Common Strategies

There are four popular types of optimal designs:

- A-optimal design minimises the trace of the inverse FIM, which is equivalent to minimizing the total parameter variance.
- D-optimal design minimises the determinant of the inverse FIM, which is equivalent to the volume of the joint confidence interval of the parameters.
- E-optimal design minimises the maximum eigenvalue of the FIM, thereby minimising the uncertainties in the worst-case direction in the parameter space.
- V-optimal design minimises the variance of the model predictions under user-defined operating conditions of interest specified by the matrix W.

The D-optimal design is the most popular one. In addition, there is a geometrical interpretation as illustrated for the two-parameter case in Fig. 3. A more detailed description can be found in [13] and [14]. However, there is usually only one parameter of interest in a physical experiment. Therefore, none of these types was used in this work. A more reasonable way is to directly minimise the target parameter error using the FIM.

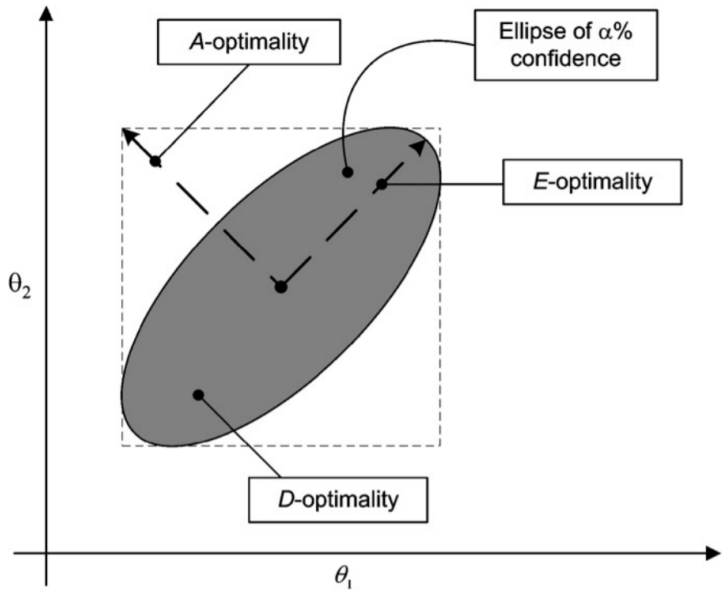


Figure 3: Geometrical interpretation of the standard criteria for the experiment design. The grey area represents the confidence region of the parameters (typically, 90% or 95%).

2.4 Derivation of the Main Equation

The definition (8) can be calculated, using the definition of the expected value, as

$$\begin{aligned}
\mathcal{I}(\theta)_{ij} = \int_X L \frac{\partial}{\partial\theta_i} logL \frac{\partial}{\partial\theta_j} logL \, d\mu = \int_X \sum_k \frac{\partial logP_k}{\partial\theta_i} \sum_k \frac{\partial logP_k}{\partial\theta_j} \prod_k P_k \, d\mu_k = \\
= \sum_k \int \left(\int \frac{\partial logP_k}{\partial\theta_i} \frac{\partial logP_k}{\partial\theta_j} P_k \, d\mu_k \right) \prod_{n\neq k} P_n \, d\mu_n + \\
+ \sum_k \int \left(\int \frac{\partial logP_k}{\partial\theta_i} P_k \, d\mu_k \right) \left(\sum_{n\neq k} \frac{\partial logP_k}{\partial\theta_j} \right) \prod_{n\neq k} P_n \, d\mu_n + \\
+ \sum_k \int \left(\int \frac{\partial logP_k}{\partial\theta_j} P_k \, d\mu_k \right) \left(\sum_{n\neq k} \frac{\partial logP_k}{\partial\theta_i} \right) \prod_{n\neq k} P_n \, d\mu_n = \\
= \sum_k \int \left(\int \frac{\partial logP_k}{\partial\theta_i} \frac{\partial logP_k}{\partial\theta_j} P_k \, d\mu_k \right) \prod_{n\neq k} P_n \, d\mu_n = \\
\sum_k \left(\int \frac{\partial logP_k}{\partial\theta_i} \frac{\partial logP_k}{\partial\theta_j} P_k \, d\mu_k \right) = \sum_k I_k
\end{aligned} \tag{11}$$

The second last equality follows from the shown in the equation (9).

Recall that we are solving the spectrum problem where each point has its own measurement time. The next step in formulating a mathematical problem statement is to take into account the properties of the spectrum in the problem under consideration. The spectrum is determined by the count rate $\mu = \mu(energy, \theta)$ and the measurement time T. Since the spectrum is the number of events in the bin, the distribution of values in the bin is Poisson. However, since the number of events is considered to be quite large,

one could write the distribution of values in the k-th bin as Gaussian with Poisson error $\sqrt{N} = \sqrt{\mu_k T_k}$

$$P_k = \frac{1}{\sqrt{2\pi\mu_k T_k}} e^{\frac{-(y_k - \mu_k T_k)^2}{2\mu_k T_k}} \tag{12}$$

Where $y_k(E_k, T_k)$ is the expected value in the k-th bin and E_k, T_k are the energy and the measurement time of the k-th bin.

One could put this into equation (11)

$$\frac{\partial log P_k}{\partial \theta_i} = \frac{(y_k - \mu_k T_k) T_k}{y_k} \frac{\partial \mu_k}{\partial \theta_i} \tag{13}$$

Define

$$\frac{(y_k - \mu_k T_k)}{y_k} = x \tag{14}$$

So, the equation could be rewritten as

$$I_k = \frac{1}{\sqrt{2\pi y_k}} T_k^2 \frac{\partial \mu_k}{\partial \theta_i} \frac{\partial \mu_k}{\partial \theta_j} \int_{-\infty}^{\infty} x^2 (-y_k) e^{-\frac{y_k}{2} x^2} \, dx \tag{15}$$

The last integral is known and equals $\sqrt{\frac{2\pi}{y_k}}$ and $\frac{T_k^2}{y_k} = \frac{T_k^2}{\mu_k T_k}$. Thus, the final formula would be

$$[\mathcal{I}(\theta)]_{ij} = \sum_k T_k \frac{\partial \mu_k}{\partial \theta_i} \frac{\partial \mu_k}{\partial \theta_j} \frac{1}{\mu_k} \tag{16}$$

From this formula one could understand that the strategy means $T(E)$, or in other words, how long which part of the spectrum should be measured.

The last step is the Cramér-Rao bound. It expresses a lower bound on the variance of unbiased estimators of a fixed, though unknown parameter. Thus, the Cramér-Rao bound determines the lowest error that can be achieved among all unbiased methods. This case is called an efficient estimator and is formulated as follows

$$\Sigma = \mathcal{I}^{-1} \tag{17}$$

Where Σ is the covariance matrix. Thus, the problem is to minimise the inverse of the Fisher Information Matrix. In addition, the minimization should depend on the target parameter. It may not have a theoretical solution, but a numerical solution will work. And the next chapter is devoted to the development of such a solution.

3 Application to Spectra

This chapter is devoted to the development and application of a numerical solution to the problem formulated in the previous one.

3.1 Beta Spectrum of Tritium

3.1.1 Beta Decay Theory

Beta decay is a type of radioactive decay which consists in the fact that a nucleus spontaneously emits leptons, electron or positron and electron neutrino or antineutrino. After that, the nucleus turns into a nucleus with the same mass number, but with an atomic number one greater or less by one.

Beta decay was believed to break the energy law due to its continuous energy spectra. Wolfgang Pauli postulated the existence of another particle, called neutrino, to conserve energy and momentum. Later, Enrico Fermi developed a theory of beta decay to include the neutrino, presumed to be massless as well as chargeless. Fermi developed an important relationship, linked the initial and final states and referred to as Fermi's Golden Rule:

$$w_{if} = \frac{2\pi}{\hbar}|M_{if}|^2\rho_f \tag{18}$$

Where w_{if} is the transition probability, M_{if} is the matrix element for the interaction and ρ_f is the density of final state. This leads to an expression for the kinetic energy spectrum N of emitted betas as follows:

$$N(KE_e) = C\sqrt{KE_e^2 + KE_e\, m_e c^2}\,(Q - KE_e)^2\,(KE_e + m_e c^2)\,F(Z', KE_e) \tag{19}$$

Where $F(Z', KEe)$ is called the Fermi function. It takes into account the nuclear coulomb interaction, which shifts this distribution towards lower energies due to the coulomb attraction between the daughter nucleus and the emitted electron. Q represents the transition energy yield and as such is the upper bound for the kinetic energy of an electron KEe, or in other words, the maximum energy.

The Fermi function can be approximated by:

$$F(Z', KE_e) \approx \frac{2\pi\eta}{1 - e^{-2\pi\eta}},\ \eta = \pm Z'\alpha E_e/pc \tag{20}$$

Where $\alpha \approx \frac{1}{137}$ is the fine-structure constant, $E_e = KE_e + mc^2$ is the total energy and $pc = \sqrt{E^2 - (mc^2)^2}$ is the momentum. And this expression is correct for non-relativistic betas, $Q \ll m_e c^2$. It is correct for the tritium beta spectrum where $Q = 18.57\,keV$.

As a result, the shapes of the beta spectra are well studied and described. The shape of the spectrum changes depending on the source, this is determined by the Fermi function which depends on the charge number of the atom and the kinetic energy of the electrons. This allows the spectrum to be recovered.

The Troitsk nu-mass experiment had used tritium, so its spectrum (Fig. 4) was used in this work. The equation for the tritium beta decay is

$$H_1^3 \rightarrow He_2^3 + e_{-1}^0 + \bar{\nu}_0^0 \tag{21}$$

so $Z' = 2$.

To account for the influence of a systematic error of the experiment, one can add to the formula (19) the decomposition of the spectrum into additional effects. It is sufficient to consider only the first, linear term. For the Troitsk nu-mass experiment, these effects are the trapping effect and the source thickness. The task of determining which effects cause systematic errors and accounting for them into account in a formula of the spectrum is not easy. However, it is possible and a solution will provide a great opportunity. This is just another reason why the data-driven approach is flawed.

Since the analytical formula of the physical phenomenon is well known, the data simulation is quite simple. It contains the following steps::

- Write an analytical equation described a phenomenon with linear approximation of some additional effects.
- Choose the most appropriate distribution. In this thesis, it is the Poisson distribution
- Use a pseudorandom number generator. For example, it might be the number generator of NumPy [15] as used in this work.

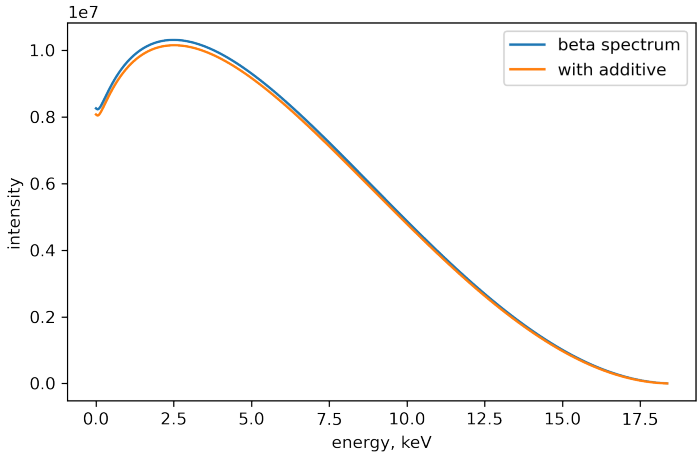


Figure 4: The comparison of the beta spectrum of tritium and the spectrum with the linear additive that represents systematic error

3.1.2 Strategy Basis

The time strategy can be decomposed into the basis. Hence, the optimization is an optimization of basis weights. And the basis consists of cubic B-splines (Fig. 5),

because they do not contain polynomials of large degree, therefore it allows the use of large dimensions without causing computational problems, and describes well an arbitrary smooth function that is supposed to be the time strategy.

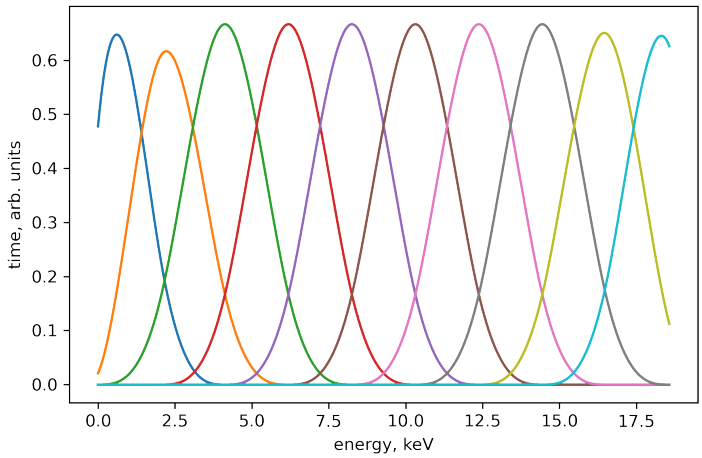


Figure 5: The cubic B-spline basis

There is also the question of how large the dimension of the basis should be. Increasing the dimension leads to problems because the iterations needed to converge increase very quickly. On the other hand, the dimension should be large enough to represent spectrum features, but not too large so as not to make the strategy too sharp. The strategy can not be too sharp because there are restrictions on the resolution in a real experiment. Therefore, the strategy should be convoluted with a normal distribution. This leads to a smearing of the strategy and all the advantages of the large dimension of the basis are leveled. Therefore, the dimension of the basis was chosen to be equal to 10.

3.1.3 Optimization Problem

The task is to minimize the target parameter, the maximum energy in this case, error using expressions (16, 17). Since the time strategy is decomposed into the basis, the problem is to the optimize basis weights. There are also two conditions:

- Maintain the total measurement time constant.
- Limit time to zero from below.

The first point represents that increasing the time leads to a decrease in errors, but the interest is in finding the optimal way not just collecting as much data as possible. Without this condition, the time strategy just inflates like a balloon. However, even in this case, there are still more important parts of the spectrum that grow faster. Another important consequence of such a retention is local minima. Moreover, the larger the dimension of the basis, the deeper these local minima are. So this is another reason why the dimension should not be too large. Moreover, to avoid local minima, a stochastic method have to be used, which is why the annealing algorithm was chosen.

The second point is necessary because negative time does not make sense, but it can occur during the optimization. It means that some parts of the spectrum contain so much useful information that it is more profitable to "measure" others negative time.

3.1.4 Optimal Strategy

The optimization described above gives a strategy of this kind (Fig. 7). It consists of two spikes:

- The first peak is directly related to the decrease in fit errors.
- The second peak tunes the parameters to reduce correlation.

The shape of the optimal strategy shows that there are parts that contain less and more information. The spikes contain enough information so some parts of the spectrum do not needed to be measured. This is illustrated in Fig. 8.

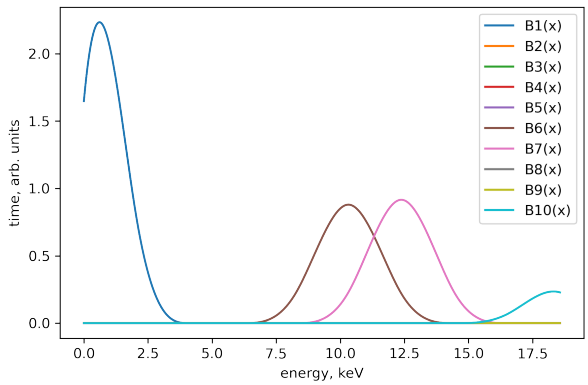


Figure 6: B-spline basis for the beta spectrum of tritium. There are only 4 splines with non-zero weights.

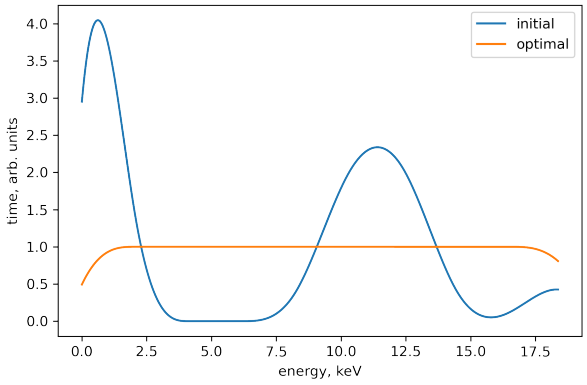


Figure 7: Optimal strategy for the beta spectrum of tritium. There are two peaks. One is likely to reduce fit errors. And the other one causes the reduction in correlations.

The B-spline basis (Fig. 6) shows that only four of the ten splines have non-zero weights, and they also do not cross each other very much. All this says that the basis of ten dimensions is close to the basis of the optimal number of dimensions.

Thus, this optimal strategy improves the target parameter error by about 1.6 times, as shown in Table 1. However, this significant improvement does not explain the Troitsk nu-mass phenomenon and requires an explanation.

3.1.5 Correlation Reduction

This error optimization can be explained by considering the change in a parameter space. The figure consists of two parts. The last or diagonal plots show the projections of the multivariate distribution of the parameters onto the level of the individual parameters. The plots in the center represent the confident regions of each pair of parameters. In

Parameter	initial error	optimal error
max energy	$2.83 \cdot 10^{-4}$	$1.83 \cdot 10^{-4}$
noise	22	31
calibration	$6.62 \cdot 10^{-5}$	$9.30 \cdot 10^{-5}$
additive	60	54

Table 1: Parameters errors for the initial and optimal strategies. The maximum energy parameter error is reduced by about 1.6 times while the errors of other parameters increased or remained almost unchanged.

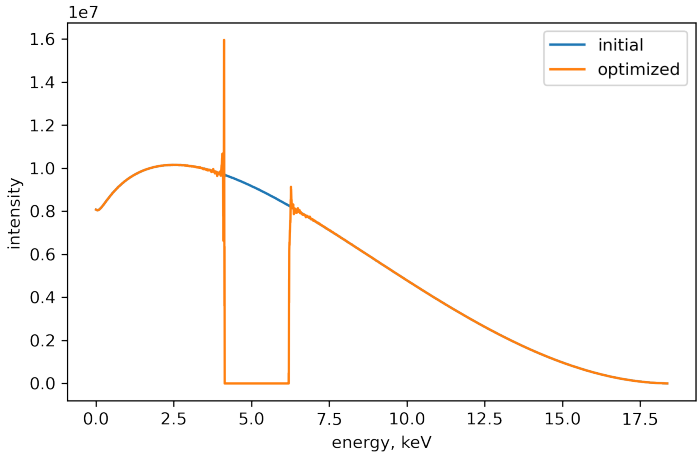


Figure 8: The beta spectrum of tritium for the optimal strategy. One part is not measured at all.

other words, they are the same distributions but from a different angle, from above. In addition, the confident regions show more information than the simple one-dimensional distribution because they also represent correlations between parameters. The more the ellipse is stretched, the greater the correlation. Thus, one can think of confident regions as representations of covariance.

Comparing the parameter space for the initial strategy (Fig. 9a) and for the optimal strategy (Fig. 9b), one can observe that the confidence regions become rounder, this represents the correlation reduction shown in Table 2. This is important because this reduction means that systematic and nuisance parameters have less influence on the estimation of the target parameter. And this explains the reduction in error. It can also be seen that the areas narrowed relative to the target parameter but increased relative to the other parameters. This is the difference between this approach and others where the entire area is minimized. This illustrates that the target parameter, for example, the mass of the particle, is the most important in the experiment, so that information about other parameters can be neglected.

3.1.6 Comparison with Common Strategies

It is also interesting to compare our optimisation based on the maximum parameter error with the common strategies described in the first chapter. The comparison is shown

Parameter	initial correlation	optimal correlation
noise	0.076	-0.16
calibration	0.54	0.14
additive	0.84	0.42

Table 2: Correlations between the maximum energy parameter and other parameters for the initial and optimal strategies for the beta spectrum of tritium. The decreasing in the correlation with the linear additive parameter might explain the improvement in the error. Since the parameter represents additional effects that cause systematic errors.

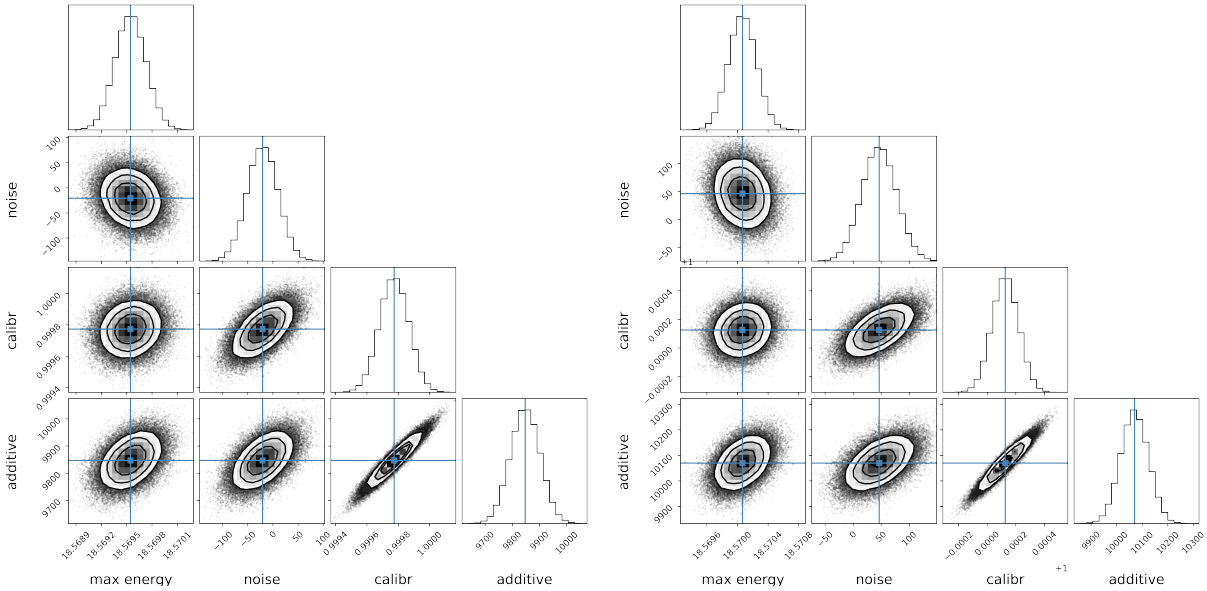


Figure 9: The beta spectrum of tritium parameters spaces for the initial (a) and optimal (b) strategies. The space alteration represents the decrease in the error of the target parameter and its correlations.

in Fig. 10. As expected, different optimal designs produce different optimal strategies. For example, A- and D-optimal designs, respectively minimising the trace and the determinant of the inverse FIM, also produce strategies with two peaks. However, the second peak is shifted and there is a much stronger peak at the end. The result of the E-optimal design or minimisation of the maximum eigenvalue of the FIM is the most different. It has only one, but very strong peak at the end.

The comparison of the maximum energy parameter error is shown in Table 3. The error is indeed the smallest in the case of error optimisation. However, its sum of errors taking into account correlations, i.e., the square root of the sum of all elements of the covariance matrix, is expected to be greater than for the initial strategy and for the strategies obtained from A- and D-optimal designs, as expected. In addition, the correlation between the maximum energy parameter and the linear additive parameter is also the smallest for the error optimisation. So it might also show that the key effect here is the correlation reduction. And the worst case is the E-optimal design. It may not be suitable for this kind of task.

Strategy	MME, 10^{-4}	errors' sum
initial	2.8	73
error optimisation	1.8	75
E-optimal design	13330.1	24594
D-optimal design	2.4	49
A-optimal design	2.2	45

Table 3: The maximum energy error (MME) obtained from minimising the MME and the optimisation strategies described in the first chapter. The last column is the sum of the errors taking into account correlations, i.e. the square root of the sum of all elements of the covariance matrix, or the inverse FIM.

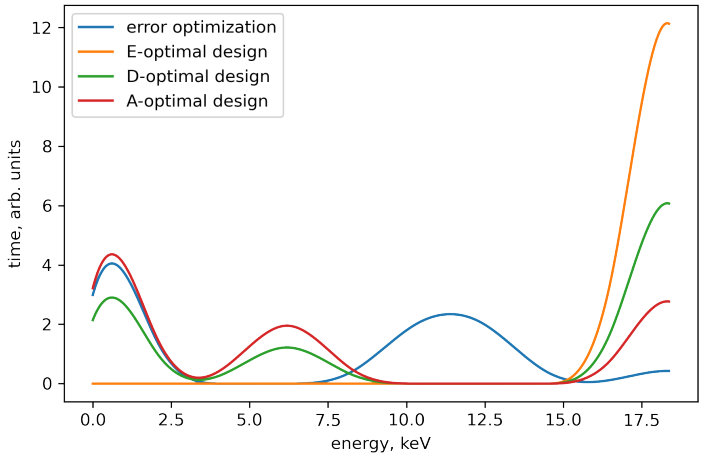


Figure 10: The comparison of optimal strategies for the tritium spectrum obtained from the optimisation based on the maximum parameter error and the common strategies described in the first chapter.

3.2 Axion Spectrum

3.2.1 Axion Physics

The method applies to experiments in any physics. So let us consider another field – axion physics. The axion is a hypothetical elementary particle and a candidate for a cold dark matter component. The interesting thing about axions is that they are involved in the well-known solution of the strong CP problem. It is obtained that in the case of a strong magnetic field, photons can transform into axions. This process is called the Primakoff process and the expression (22) represents the axion photon conversion probability. From this probability it is clear that L is actually a coherence length and that there are oscillations of photons and axions. Hence, the conversion probability is maximized if the axion and photon remain in phase over the length of the magnet and satisfy the coherence condition $qL < \pi$. As result, it determines the parameters of the experiment, the size of the tube and the amplitude of the magnetic field.

$$P_{a\to\gamma} = \left(\frac{g_{a\gamma}B}{q}\right)^2 \sin^2\left(\frac{qL}{2}\right) \tag{22}$$

Where L is the magnet length, B is the magnetic field, and $q = m_a^2/(2E_a)$ is the axion-photon momentum transfer, m_a and E_a are the mass and energy of axion.

3.2.2 Solar Axions

The uncertainties in the observation of some fundamental astrophysical quantities allow the Sun to emit additionally half of its energy. It means that there may be many new unknown particles born in the Sun. For instance, it may be axions. However, it exacerbates the solar neutrino problem. Nevertheless, this work considers not which particles are more likely, but the application of the developed method to experiments from different fields of physics.

So the solar core has suitable conditions for the creation of axions and the Sun can be considered as a factory, generating axions. Moreover, only from this fact, one could get that the upper bound of the axion mass is 20 eV [16]. And there are two most important processes of the axion production:

- The Primakoff proccess (Φ_P) that dominants in hadronic axion models like the KimShifman-Vainshtein-Zakharov
- The processes involving electrons: Compton scattering (Φ_C), bremsstrahlung (Φ_B), and atomic recombination and deexcitation (Φ_A)

These two types of processes are described by different models and can be considered separately. The difference is in the differential flux at Earth due to these mechanisms. It can be parameterized as

$$\frac{d\Phi_P}{dE_a} = \Phi_{P10} \left(\frac{g_{a\gamma}}{10^{-10} GeV} \right)^2 \frac{E_a^{2.481}}{e^{E_a/1.205}} \tag{23}$$

where E_a is always in units of keV $\Phi_{P10} = 6.02 \cdot 10^{10}\, cm^2 s^1 keV^1$, and

$$\frac{d\Phi_C}{dE_a} = \Phi_{C13} \left(\frac{g_{ae}}{10^{-13}} \right)^2 \frac{E_a^{2.987}}{e^{0.776 E_a}} \tag{24}$$

where $\Phi_{C13} = 13.314 \cdot 10^6\, cm^2 s^1 keV^1$, and

$$\frac{d\Phi_B}{dE_a} = \Phi_{B13} \left(\frac{g_{ae}}{10^{-13}} \right)^2 \frac{E_a}{1 + 0.667 E_a^{1.278}} e^{-0.77 E_a} \tag{25}$$

where $\Phi_{B13} = 26.311 \cdot 10^8\, cm^2 s^1 keV^1$. However Φ_A can not be efficiently or accurately parametrized. So it should be simulated and that is why these mechanisms are not considered in the thesis.

3.2.3 Helioscope

This part of the work is devoted to the IAXO experiment (Fig. 11) [17]. It is worth noting that there is no certainty that the method can be applied to it. IAXO is a

helioscope consisting of a long magnetic bore pointed at the Sun with a collecting x-ray detector at one end. The expected number of photons reaching a detector placed at the end of the bore is given by the integral

$$N_\gamma = St \int dE_a \varepsilon_D(E_a)\varepsilon_T(E_a)\frac{d\Phi_i}{dE_a}P_{a\to\gamma}(E_a) \tag{26}$$

where S is the total cross-sectional area of the helioscope, and t is the measurement time. $\frac{d\Phi_i}{dE_a}$ is the axion flux due to process i (i.e., P or $A+B+C$), $P_{a\to\gamma}$ is the axion-photon conversion probability (eq. 22). And ε_D, ε_T are two efficiency functions for the detector and the telescope.

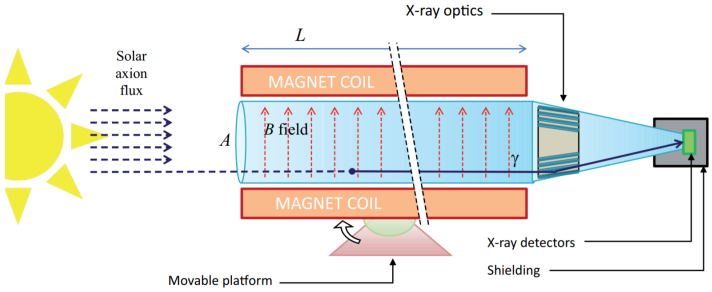


Figure 11: The scheme of IAXO. Axions come to the Earth from the Sun. A strong magnetic field in a tube of the helioscope makes them produce photons. These photons are then measured by X-ray detectors. The length of the tube determines what axion masses are possible to detect.

The systematic uncertainties come from the imperfect understanding of plasma screening effects. However, it was unclear how to take them into account therefore the application to axion spectra does not include any parameter related to systematic errors. Nevertheless, the results are quite interesting even without it.

3.2.4 Primakoff Axions Spectrum

Primakoff Axions have a more smooth spectrum so they were chosen for the analysis. And the figure 12 shows differential x-ray spectra for different axion masses. One could see that, for a relatively big mass, there are oscillations and the spectra of small masses are indistinguishable. For comparison, the spectra were normalized.

The optimized strategies for different masses are slightly different (Fig. 14). However, a strategy from one mass or even from neutrinos can be applied to another mass and there will be optimization but not so great. Also, there is no second peak. Thus, this may be the result of the absence of systematic errors in the model. And in contrast to the case with the beta spectrum, correlations of the parameters do not decrease. This can be considered as an argument that the second peak is indeed responsible for reducing correlations.

If one apply the optimal strategies (Fig. 14) to measure the spectrum, the measurement spectra will have interesting feature. There will be a threshold point after which the

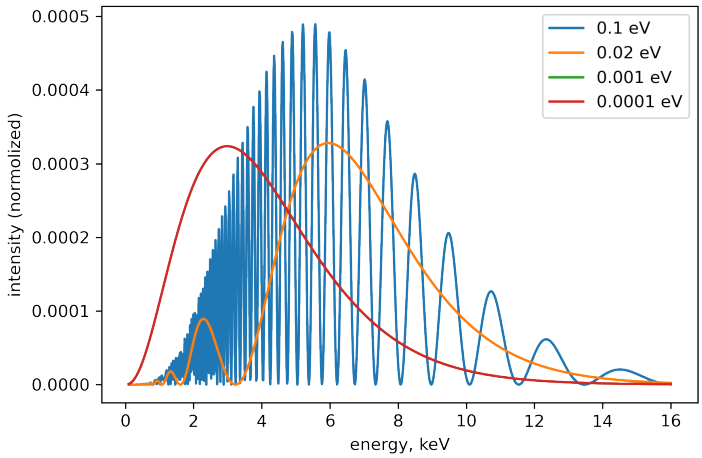


Figure 12: The axion spectra for different masses. The largest mass is close to the upper limit (0.016 eV) which is based on the coherence condition. And the smallest masses are difficult to distinguish.

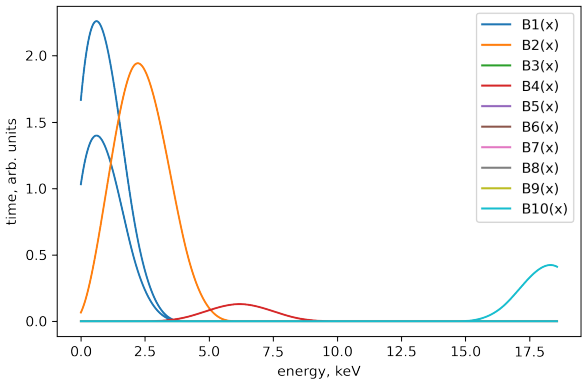


Figure 13: The B-spline basis for axions. There is not much difference for different masses. Thus, the strategy from one mass may be applied to another.

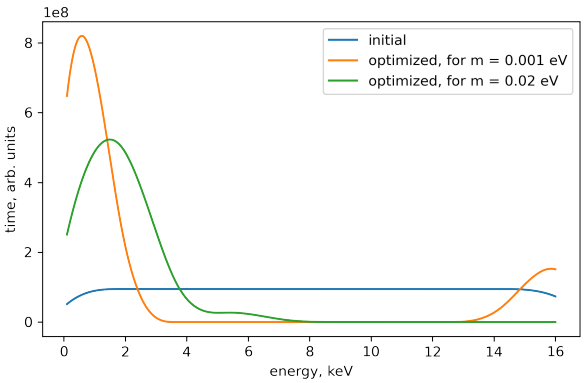


Figure 14: The optimal strategy for axions. There is no second peak since there is no parameter related to systematic errors so correlations are not optimized.

spectrum is not measured. It is especially true for the spectrum of 0.02 eV axions (Fig. 16). However, for 0.001 eV axions (Fig. 15), there is also some measurements at the end. These features may help to apply the method to the IAXO experiment using a trigger to set up the strategy.

3.2.5 Parameters Space

The transformations of the parameter spaces are more significant in this case. Moreover, they are quite different for different masses. For instance, the parameter space for 0.001-eV axion was asymmetric for the initial strategy. But for the optimal strategy, it is almost symmetric. Thus, the optimization of the strategy might be understood as a kind of normalization of the projection of the parameter space projection onto a target parameter. In other words, this optimization makes the target parameter distribution more normal and confident regions with the target parameter more circular with respect

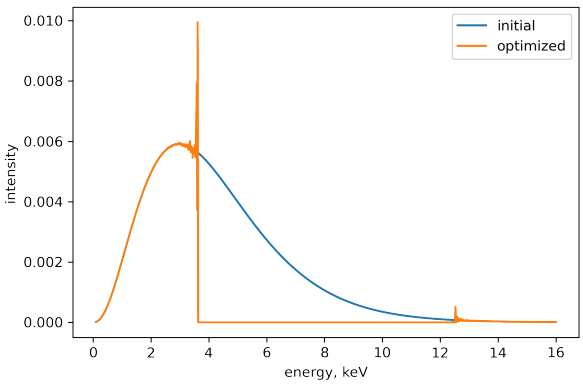


Figure 15: The spectrum of 0.001 eV axions for the optimal strategy. A small part of the spectrum end is measured however it may still be possible to simply cut off the end.

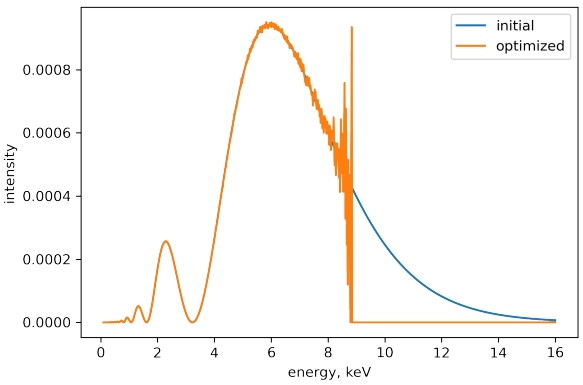


Figure 16: The spectrum of 0.02 eV axions for the optimal strategy. The end of the spectrum is not measured so it is possible to use a trigger to set up the strategy.

to the target parameter. However, it does not mean that a confident region becomes more circular because the errors of the nuisance parameters might become larger. Nevertheless, the more similar the distribution is to the normal distribution, the easier it is to work with.

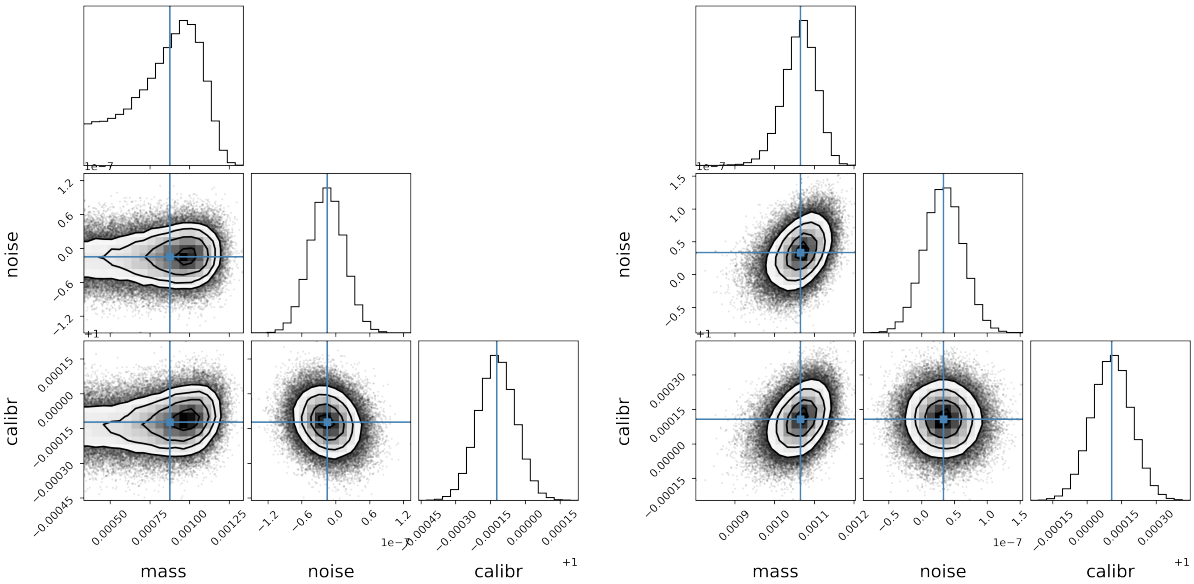


Figure 17: The 0.001 eV axion parameter space for the initial (a) and optimal (b) strategies. The optimisation altered the space making it more symmetric or Gaussian-like.

This logic is also demonstrated with an example involving the 0.02 eV axion. This situation is interesting because the initial parameter space (Fig. 18a) was strongly asymmetric with respect to the noise parameter. Also, the distribution of the axion mass has a small ledge near the end. And for the optimal strategy (Fig. 18b), there is no ledge in the distribution. Although the distribution of the noise parameter is still asymmetric, it now looks like a half normal distribution. Thus, the error of the noise parameter has increased

to remove the ledge in the distribution of the axion mass. It is also characteristic that the confident region between the calibration and axion mass parameters has become less elongated.

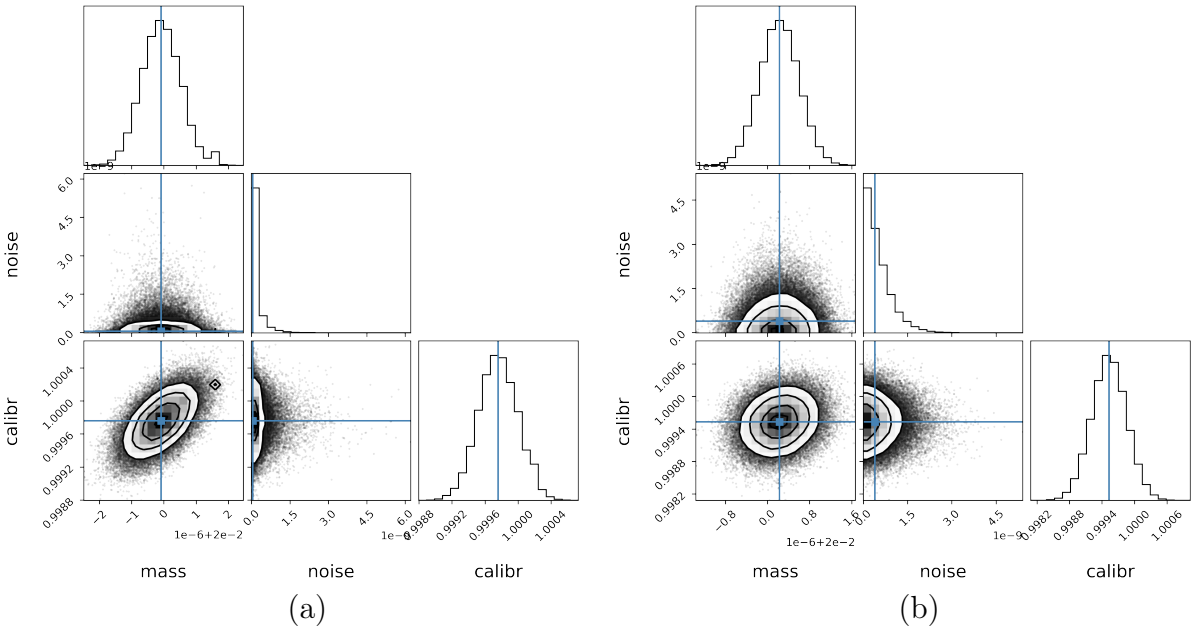


Figure 18: The 0.02 eV axion parameter space for the initial (a) and optimal (b) strategies. The resulted space is still not symmetric but very close. Most importantly, an additional small peak in the mass distribution is removed for the optimal strategy.

Finally, 0.0001 eV axion is under consideration. As was described above (Fig. 12), axion spectra are indistinguishable for small masses. It is illustrated in the parameter space (Fig. 19a). There is an almost flat distribution near the true value. However, about 0.0006 eV point is a threshold point after which the probability is decreasing. So, it may be considered as the maximum mass that can be measured, or in the other words, masses less than 0.0006 eV are indistinguishable. And the optimal strategy (Fig. 19b) makes a peak around this threshold point.

3.2.6 Error Comparison

The flat distributions described above have an interesting consequence. Because they are flat, standard methods such as Python's $curvefit$ function do not estimate the parameters correctly. The errors are underestimated. However, the error estimates from the Fisher information matrix are quite large. As a result, these two error estimates are not equal for small masses.

The difference in error estimates might be an indicator that the distribution of the target parameter has become uniform. And the uniform distribution means that one can not make a good estimate. For this reason, we considered the ratios of the errors obtained from the Fisher information matrix and the $curvefit$ function with simulation. The results are shown in Figure 20.

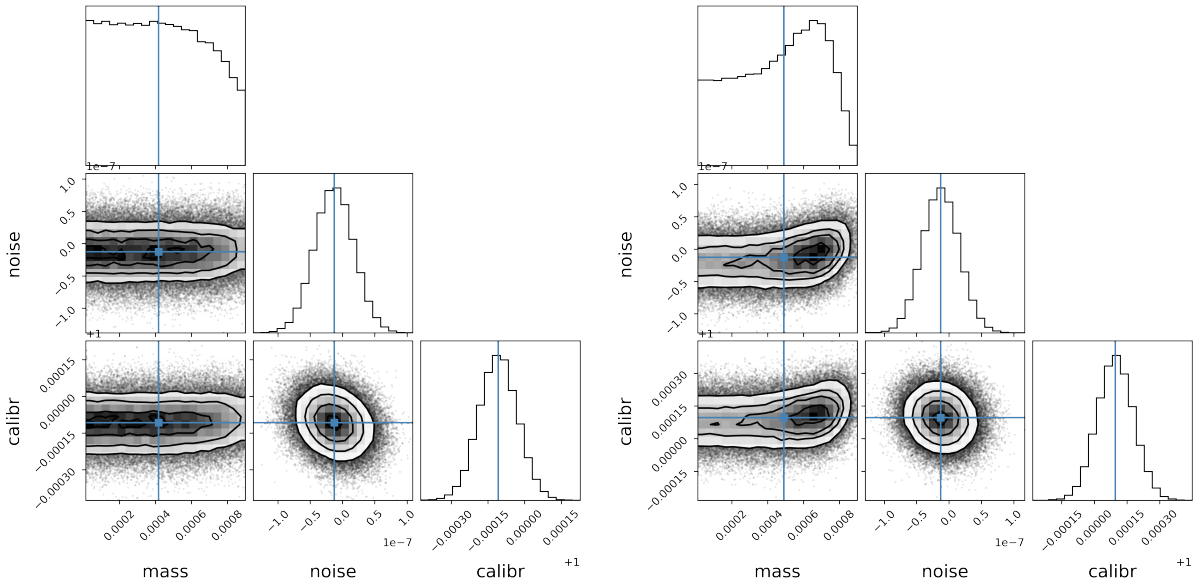
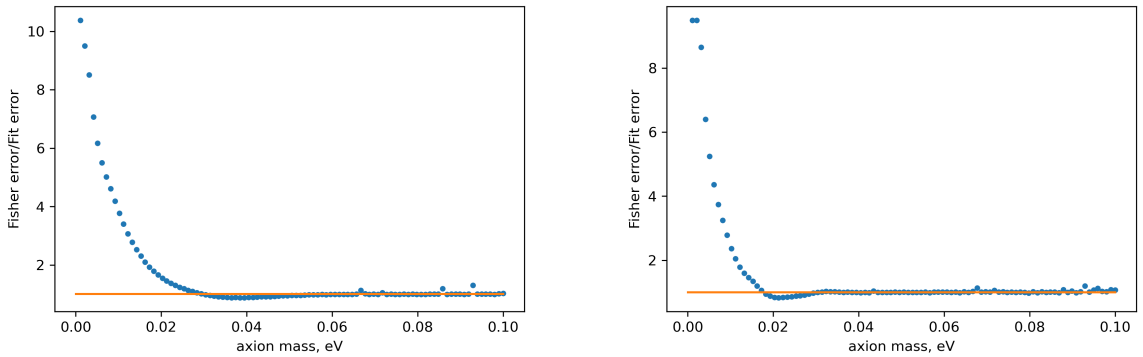


Figure 19: The 0.0001 eV axion parameter space for the initial (a) and optimal (b) strategies. The left uniform distribution shows the impossibility of distinguishing between small masses. And the optimal strategy may shift the lowest limit for discovering the axion mass.

Thus, the method may help to solve the small mass problem and push the lower bound of the distinguishability of axion masses.



(a) (b)

Figure 20: The ratios of the errors obtained from the Fisher information matrix and the function *curvefit* with the simulation for the initial (a) and optimal (b) strategies. A point at which the ratio becomes almost equal to 1 might be considered as the threshold point that defines the lower limit of the axion mass. Hence, its shift towards lower masses means a descent in the lower boundary.

3.2.7 Mass Variety

The axion is a hypothetical particle, so there is a variety of possible masses. However, the method is applied to a specific experiment, not to a spectrum itself. And this produces bounds. First, IAXO will measure solar axions and their masses have an upper bound

from uncertainties in solar observations, as described above. Then there are bounds from the experiment:

- The upper bound is determined by the coherence length or a tube length of the helioscope.
- The lower bound is determined by the absence of differences in the spectra for small masses.

Both cases are shown in Figure (12). So the variety is actually not that great. Hence, one can just pick a median mass and use its optimal strategy, and there will be an improvement event if that mass was a wrong choice. Moreover, one can calculate the convolution of the masses to get more universal results. In addition, the method is more useful for cases where some measurements have already been made, and the next chapter is devoted to this problem.

3.3 Some Useful Tools

3.3.1 Global Optimization

Many statistic tasks are reduced to multidimensional optimization. Hence, we do not actually know the shape of our function it may have many local minima. So, we can assume the worst case or reason that local minima might exist like in the thesis. As a result, we should find the global optimum. For instance, we can start multiple initial guesses. If the optimal solution remains almost constant, it might indicate the optimiser works well and the solution is at least very close to the global one. Another way is to use one of the global optimisation methods like the following:

- Simulated Annealing (described below).
- Evolutionary Computation [18].
- Ant Colony Optimization [19].

There are also many other methods and ways to group them. We describe only the simulated annealing method since it was used in the thesis. One can find descriptions of others in the article [20]. However, the general properties of such a stochastic approach are that they are very flexible and easy to implement since a numerical simulation considered as black box. In addition, their randomness leads to that an optimization result can change in every run and the computational cost, number of objective function evaluation, is high. Moreover, there is no guarantee for convergence.

So, simulated annealing is a random search method that simulates the physical process of metal annealing, finding low-energy states of a solid in a heat bath via heating and cooling of a material. As a result, it avoids getting trapped in local minima by accepting

not only transitions corresponding to an improvement in an objective function value but also transitions corresponding to a worse value. Hence, the method is specified by an acceptance probability function that determines the probability of making the transition from the current state to a candidate new worse state. It was originally the Gibbs measure since it describes the state distribution. This function depends also on a global time-varying parameter called the temperature. When the temperature tends to zero, the probability must also tend to zero if a new state has a greater energy, a worse value, and to a positive value otherwise. Thus, the temperature plays a crucial role in controlling the evolution of the system. The higher the temperature, the more sensitive the evolution of the state to coarser changes in energy. While it is sensitive to finer energy variations when the temperature is low. As result, the advantage of simulated annealing is that it is very easily implementable, robust, and applicable to a very general class of global optimization problems. One of the best illustrations of this evolution process is shown on the corresponding Wikipedia's page.

3.3.2 Convolution

When one works with distributions, the most important operation is convolution of two functions that expresses how the shape of one is modified by the other. On the other hand, convolution is just the sum of random variables generated by these two distributions. So, if one wants to work with distributions directly, there is a definition:

$$(f * g)(t) = \int_{-\infty}^{\infty} f(\tau)g(t-\tau)d\tau \tag{27}$$

However, there may appear some problems with the realisation of convolution. So, We provide a sample Python code using lambda functions:

```
convolution = lambda f, g, a, b:
                numpy.vectorize( lambda x: integrate(
                        lambda y: f(x-y)*g(y), a, b
                        ) )
```

And there are also some problems. There is a function at the output of the interpolator, so we are using pure Python, which is extremely slow. To solve that, we can use another language, for example Kotlin, which can do it almost for free. Also, we need to calculate the integral and do the interpolation with splines when we have an array and not a function. This may lead to additional errors and makes the process slower. A faster way is to use the numpy.convolve function, which computes the discrete convolution of two finite sequences. To get a correct result, one should make sure that the grids of the two distributions match. Furthermore, to ensure that the method used computes a correct convolution, one can try the method on the example of two normal distributions.: $N(0, \sigma_1^2)$ and $N(\mu, \sigma_2^2)$. Since convolution is just the sum of random variables, the result should be $N(\mu, \sigma_1^2 + \sigma_2^2)$ if we assumed that the variables are independent.

4 Bayesian Statistics

This chapter is devoted to Bayesian statistics, one of the most promising areas of statistics.

4.1 Frequentist versus Bayesian

4.1.1 Key Points

There are two schools of statistics: Bayesian and frequentist. Both approaches allow one to evaluate parameters and make predictions. There is also a debate about which approach is better. This question is posed incorrectly. However, to understand better the features of both approaches, it is important to highlight differences between them. The key points are shown in the Table 4.

Frequentist	Bayesian
long-term frequencies	degrees of belief/logical support
θ is fixed	θ is random
X is random	X is fixed
$\|X\| >> \|\theta\|$	any X
Maximum Likelihood	Bayes theorem

Table 4: Comparison of frequentist and Bayesian statistics that shows the main features and differences

So, the first difference is in the definition of probability. For the frequentist, it is long-term frequencies, e.g., a coin toss, i.e. some lengthy process

$$P(x) = \lim_{n_t \to \infty} \frac{n_x}{n_t} \tag{28}$$

Where n_x is the number of trials where the event x occurred, and n_t is the total number of trials. Thus, for the frequentist, the repeatability of an experiment is the key concept, since the number of trials should approach infinity. For the Bayesian, on the other hand, probability is the degree of belief or logical support (Eq. 32). This leads to the fact that the frequentist approach requires that the sample size of the data be large enough. The "large enough" means that the data sample size (X) is larger than the number of parameters (θ). In contrast, the Bayesian approach works with arbitrary data sample sizes. It is even able to deal with one event that will never occur again. In addition, as described in the second chapter, the parameter θ is fixed in the frequentist method. And in Bayesian statistics, it is treated as a random variable, blurring the boundaries between data and parameters.

Another difference is in the method by which estimates are made. In frequentist statistics, maximum likelihood estimation is used. In Bayesian statistics, it is the maximum a posteriori estimation or Bayes theorem. However, the first method is only a special case

of the second method when the prior distribution is a uniform distribution, as shown in Eq. 29 - 31. And an additional constant does not change the results.

$$\theta_{MAP} = \underset{\theta}{\operatorname{argmax}} P(\theta|X) = \underset{\theta}{\operatorname{argmax}} P(X|\theta)P(\theta) = \underset{\theta}{\operatorname{argmax}} \sum_i log\, P(x_i|\theta) + log\, P(\theta) \quad (29)$$

But since

$$\theta_{MLE} = \underset{\theta}{\operatorname{argmax}} \sum_i log\, P(x_i|\theta) \quad (30)$$

the right part of eq. (29) can be rewritten as

$$\theta_{MAP} = \theta_{MLE} + constant \quad (31)$$

The second equality in eq. (29) is from Bayes theorem (eq. 32). Moreover, the denominator $P(X)$ should be neglected since it does not affect the result but requires a lot of resources on the calculation.

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)} \quad (32)$$

The prior $P(\theta)$ is the probability that θ is true before the data is considered. The posterior $P(\theta|X)$ is the probability that θ is true after the data is considered. The likelihood $P(X|\theta)$ is the evidence about θ provided by the data X. And $P(X)$ is the total probability of the data taking into account all possible hypotheses, that is, a normalization constant.

So the Bayesian approach may seem like a kind of generalization of the frequentist approach, just as quantum mechanics generalizes classical. However, the implementations of the methods have subtle but fundamental differences. Therefore, for some problems the Bayesian approach is preferable, for others the frequentist approach.

4.1.2 Bayesian Billiards Problem

To illustrate the differences, we consider the "Bayesian Billiards Problem" [21]. There are two players, Alice and Bob, they can not see the billiard table. Carol rolls a ball down the table and marks where it lands. Once this mark is in place, Carol begins rolling new balls down the table. If it comes to rest to the left of the initial mark, Alice wins the point; to the right of the mark, Bob wins the point. The first person to reach six points wins the game. We assume that Alice is leading with 5 points and Bob has 3 points. The question is what the expected probability that Alice will win is.

Frequentist (naive) approach is following:

- The frequency at which Alice has won so far is 5/8

- Hence a maximum likelihood estimate of θ that any given roll lands in Alice's favor is $\theta_{MLE} = 5/8$.
- The probability that Bob will win $P(B) = (1-\theta)^3 \approx 0.05$. Since, he needs to get all 3 balls, otherwise Alice wins.

In Bayesian approach, the posterior probability is

$$P(B|D) = \frac{\int P(B|\theta, D)P(D|\theta)P(\theta)\, d\theta}{\int P(D|\theta)P(\theta)\, d\theta} \tag{33}$$

Where we used the chain rule $P(X, Y, Z) = P(X|Y, Z)P(Y|Z)P(Z)$. B represents that Bob wins, $D = (5, 3)$ is observed data and θ is unknown probability that a ball lands in Alice's side during the current game. The prior $P(\theta)$ is assumed to be uniform so it is just a constant. $P(D|\theta)$ is a binomial $\frac{8!}{5!3!}\theta^5(1-\theta)^3$ since 8 games have been played and Alice has 5 wins. In addition, to calculate integrals, we use a gamma function $(n+1) = n!$ and that $\int_0^1 x^{n-1}(1-x)^{m-1}\, dx = \frac{\Gamma(n)\Gamma(m)}{\Gamma(n+m)}$ Thus,

$$P(B|D) = \frac{\int_0^1 \theta^5(1-theta)^3\, d\theta}{\int_0^1 \theta^5(1-theta)^3\, d\theta} \approx 0.09 \tag{34}$$

Using Monte Carlo simulations, we can obtain that the true answer is 0.09. So, Bayesian approach is right. There is still a question of why the frequentist result is wrong. The problem is not the estimate itself. In fact, in this example, the Bayesian maximum a posterior probability is the same as the frequentist MLE. The difference is that Bayesian posterior contains all of the information we have about θ, whereas the frequentist result discards a large part of that information. The result we are interested in, the probability of winning the match, is a non-linear transform of θ, and in general for a non-linear transform f, the expectation $E[f(\theta)]$ does not equal $f(E[\theta])$. The Bayesian method computes the first, which is right; the frequentist method approximates the second, which is wrong.

To summarise, Bayesian methods are better not just because the results are correct, but more importantly because the results are in a form, the posterior distribution, that lends itself to answering questions and guiding decision-making under uncertainty. That is vital for Physics since we have theories and old experiments as prior information and want to obtain results as distributions like Bayesian methods yield.

4.1.3 Critiques and Defenses

To finish the comparison, we consider critiques and defenses of both kinds of estimators. The main critique of Bayesian approach is that a subjective prior is subjective. Different people produce different priors and hence may obtain different results and figure different conclusions. In addition, there are philosophical objections to assigning probabilities to hypotheses, since hypotheses do not constitute outcomes of repeatable experiments in which one can measure long-term frequency. Rather, a hypothesis is either true or false,

regardless of whether it is known to be so. The coin is either fair or unfair; the sun will or will not rise tomorrow. In addition, Bayesian methods are computationally intensive. This is the reason why Bayesian methods are only now experiencing an enormous renaissance in fields like machine learning and medicine.

The defense is that the probability of hypotheses is vital to make decisions. A patient wants to know the probability of his diagnosis. And it is easy to report a result formulated in terms of probabilities of hypotheses. In addition, by trying different priors one can see how sensitive his results are to the choice of prior. Moreover, even though the prior may be subjective, one can specify the assumptions used to arrive at it allowing other people to challenge it or try other priors. An important thing is that data can be used as it comes in. It gives one a few opportunities. Thus, Bayesian approach addresses the question everyone is interested in, by using assumptions no one believes.

Frequentist approach is ad-hoc and does not carry the force of deductive logic. Moreover, experiments must be fully specified in advance. This can lead to paradoxical seeming results. In addition, in hypothesis testing, the p-value and significance level are known to be often misinterpreted. And this misinterpretation led to a huge amount of wrong conclusions and worthless research. However, frequentist approach is objective, it uses impeccable logic to deal with an issue with no interest of anyone.

4.2 Bayesian Fisher Information

4.2.1 Definition

As described above, the transition to Bayesian statistics is done by replacing likelihood with

$$\mathcal{L} = L(\theta)\pi(\theta_0) \tag{35}$$

where θ_0 can be any parameter.

So one could get

$$\frac{\partial \mathcal{L}}{\partial \theta_0}\frac{\partial \mathcal{L}}{\partial \theta_j} = \left(\frac{\partial L}{\partial \theta_0} + \frac{\partial \pi}{\partial \theta_0}\right)\frac{\partial L}{\partial \theta_j},\ j \neq 0 \tag{36}$$

and

$$\frac{\partial \mathcal{L}}{\partial \theta_0}\frac{\partial \mathcal{L}}{\partial \theta_0} = \left(\frac{\partial L}{\partial \theta_0} + \frac{\partial \pi}{\partial \theta_0}\right)^2,\ j \neq 0 \tag{37}$$

So, new the Fisher information matrix can be written as

$$\tilde{I}_{0j} = E\left(\frac{\partial L}{\partial \theta_0}\frac{\partial L}{\partial \theta_j}\right) + E\left(\frac{\partial \pi}{\partial \theta_0}\frac{\partial L}{\partial \theta_j}\right) = I_{0j},\ j \neq 0 \tag{38}$$

since

$$E\left(\frac{\partial\pi}{\partial\theta_0}\frac{\partial L}{\partial\theta_j}\right) = E\left(\frac{\partial\pi}{\partial\theta_0}\right)E\left(\frac{\partial L}{\partial\theta_j}\right),\ j \neq 0 \tag{39}$$

and

$$E\left(\frac{\partial L}{\partial\theta_j}\right) = 0 \tag{40}$$

It should be noted that, to calculate the expected value, likelihood L was used. And for $j = 0$ one could get

$$\tilde{I}_{00} = E\left(\frac{\partial L}{\partial\theta_0}\frac{\partial\pi}{\partial\theta_0}\right)^2 = E\left(\frac{\partial L}{\partial\theta_0}\right)^2 + E\left(2\frac{\partial L}{\partial\theta_0}\frac{\partial\pi}{\partial\theta_0}\right) + E\left(\frac{\partial\pi}{\partial\theta_0}\right)^2 = I_{00} + \delta_{00} \tag{41}$$

Thus, the prior information matrix δ has only one nonzero element associated with the parameter θ_0. It means that δ is a singular matrix. And one can assume that π is normal distribution

$$log\pi(\theta_0) \sim \frac{(\vartheta - \theta_0)^2}{2\sigma^2} \tag{42}$$

So, the prior information matrix equals

$$\delta_{00} = \frac{(\vartheta - \theta_0)^2}{\sigma^4} \tag{43}$$

Thus, $\tilde{I}(\vartheta;\theta)$ is a distribution. And one should calculate the convolution of it and normal distribution to obtain an estimate

$$\int p(\vartheta)\left(I_{00}(\theta) + \delta_{00}(\vartheta;\theta)\right)d\vartheta = I_{00} + \int \frac{(\vartheta - \theta_0)^2}{\sigma^4} \times \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(\vartheta-\theta_0)^2}{2\sigma^2}} d\vartheta = I_{00} + \frac{1}{\sigma^4} \tag{44}$$

As result, prior information about target or event nuisance parameters might improve the target parameter error.

4.2.2 Application

Only the application of the method to the tritium beta spectrum is considered. The errors obtained using the formula (44) and the errors obtained using numerical methods, including Monte Carlo methods, are equal. And they are shown in the Table 5

Thus, prior information about the target parameter obviously reduces its error. Moreover, such information about nuisance parameters improves the target parameter error too but slightly. However, there is almost no improvement for the optimal strategy. This is a consequence of the correlation reduction. For the initial strategy, there are enough big correlations so nuisance parameters are able to affect the target parameter. But for

Parameter	initial MEE, 10^{-4}	optimal MEE, 10^{-4}	improvement
maximum energy	1.90	1.49	1.27
noise	2.83	1.82	1.55
calibration	2.74	1.82	1.50
additive	2.55	1.79	1.42

Table 5: Errors of the maximum energy parameter (MEE) for the initial and optimal strategies and their improvements with prior information about different parameters. Without prior information, MEE for the initial strategy is $2.82 \cdot 10^{-4}$ for the initial strategy and $1.83 \cdot 10^{-4}$ for the optimal strategy. So, the improvement is 1.54.

the optimal strategy, these correlations are so small that there is no longer any influence of nuisance parameters.

4.3 Regularisation of Inverse Problems

4.3.1 Inverse Problems in Physics

There are several definitions of inverse problems. One possible definition is that inverse problems are concerned with determining causes for a desired or observed effect. The cause is, for example, an unknown parameter, and the effect might be some data or an observation. Inverse problems are among the most important mathematical problems in science because they tell us something about parameters that we cannot directly observe. They have broad applications that can occur almost wherever.

The simplest case, a linear problem, might be inverse. For instance, one can consider the Earth's gravitational field. It is determined by the density distribution of the Earth in the subsurface. Due to significant changes in the Earth's lithology, we can observe tiny differences in the Earth's gravitational field at the Earth's surface of the Earth. From Newton's law of gravity, we know that the mathematical expression for gravity is

$$a = \frac{Gm}{r^2} = Fm \tag{45}$$

here a is a measure of the local gravitational acceleration, G is the universal gravitational constant, m is the local mass of the rock in the subsurface, and r is the distance from the mass to the observation point. So, a and m are vectors of the dimension of the number of observations, and F is a matrix.

To find the model parameters that fit the data, we can invert the matrix F to convert the measurements directly into our model parameters:

$$m = F^{-1}a \tag{46}$$

However, even in this case, the matrix F may have no inverse. It may be not square or have zero eigenvalues so the solution is not unique. Also, noise may corrupt our observations making m possibly outside the space $F(P)$ of possible responses to model

parameters so that solution of the system [54] may not exist. In addition, a large dimension of the matrix F can also cause such problems.

In the thesis, we are trying to inverse the Fisher information matrix that is a much harder task. And for the case of high correlations, it might produce too large errors that make no sense and is a difficulty. An example of such a situation is the beta spectrum with explicit allowance for the neutrino mass.

To take the neutrino mass into account in Eq. 19, we first consider the law of energy conservation

$$Q = M_H - M_{He} - m_e = E_e + E_\nu + m_\nu \tag{47}$$

Where Q is the maximum energy of the spectrum if neutrino is zero, M_H, M_{He}, m_e, m_ν are masses of tritium, helium, electron and neutrino. Neutrino energy can be expressed as follows

$$E_\nu = \sqrt{p_\nu^2 + m_\nu^2} - m_\nu \tag{48}$$

So, one can obtain

$$p_\nu = \sqrt{E_\nu^2 + 2E_\nu m_\nu} = \tag{49}$$

And then using (19)

$$N(E) \sim \sqrt{(Q-E)^2 - m_\nu^2} \tag{50}$$

Thus, the neutrino mass shifts the spectrum end. So, precise knowing the end makes it possible to determine the mass.

As result, there are huge correlations between parameters so common methods of fitting do not work and relative errors obtained using the Fisher information are in the millions. It is a bad situation but on other hand, this behaviour of the Fisher information predicts the behaviour of a fitting function in a sense. Nevertheless, it is an inverse problem and it might be solved with regularisation.

4.3.2 Common Solutions

Using regularisation to solve inverse problems is a part of functional analysis and well studied for some cases. So, we only mention some interesting solutions. One can find more detail description in [22]. One of them is Tikhonov regularization [23]. Suppose that for a known matrix A and vector b , we wish to find a vector x such that

$$Ax = b \tag{51}$$

In order to give preference to a particular solution with desirable properties, a regu-

larization term can be included as a penalty in the ordinary least squares minimisation:

$$x_{Tik} = \text{argmin}\left(||Ax - b||^2 + \alpha||\Gamma x||^2\right) \tag{52}$$

Here Γ is Tikhonov matrix and it monitors the growth of instability. A choice of the regularization operator Γ requires a priori belief about the solution because it could be anything. In many cases, this matrix is chosen as a multiple of the identity matrix, giving preference to solutions with smaller norms; this is known as L_2 regularization.

Another method is to use the singular value decomposition (SVD) [24]. SVD of a $m \times n$ matrix A is a factorization of that matrix into three matrices:

$$A = U\Sigma V^* \tag{53}$$

Where U is a $m \times m$ unitary matrix, V is a $n \times n$ unitary matrix and Σ is a $m \times n$ diagonal matrix. And Σ_{ii} are known as the singular values of A. As result, SVD can be used for computing the pseudoinverse of a matrix A:

$$A^+ = V\Sigma^+ U^* \tag{54}$$

Where Σ^+ the pseudoinverse of Σ, which is formed by replacing every non-zero diagonal entry by its reciprocal and transposing the resulting matrix. So, it is one way to solve linear least-squares problems.

Thus, regularisation helps to improve or obtain results in inverse problems we face every day. And in the simplest case, the problem of a near-singular matrix is alleviated by adding positive elements to the diagonals, thereby decreasing its condition number.

4.3.3 Bayesian Fisher Information as Regularisation

Regularisation is a useful technique. But a more natural and understandable approach is Bayesian since regularisation is Bayes theorem. To illustrate it, we consider L_2 regularisation described above that is popular in machine learning for the linear regression $f(x) = w^T x + w_0$:

$$w_{reg} = argmin_w \left(\sum_{n=1}^{N} (y_n - f(x_n))^2 + \lambda \sum_{k=1}^{K} w_k^2 \right) \tag{55}$$

Then we consider MAP (eq. 29) and assume that the likelihood and the prior distribution are normal:

$$w_{MAP} = arg\, max_w \prod_{n=1}^{N} \mathcal{N}(y_n; f(x_n), \sigma_y^2)\mathcal{N}(w; 0, \sigma_w^2) \tag{56}$$

So we get

$$w_{MAP} = argmin_w \left(\sum_{n=1}^{N} (y_n - f(x_n))^2 + \frac{\sigma_y^2}{\sigma_w^2} \sum_{k=1}^{K} w_k^2 \right) \tag{57}$$

As a result, we obtain the equation that repeats eq. 55. Moreover, we immediately determine the hyperparameter $\lambda = \sigma_y^2/\sigma_w^2$, so we do not need to optimise it. So we need some prior information to perform the regularisation, hence the regularisation is just Bayes theorem.

This consideration leads to the idea that the Bayesian Fisher information might be a regularisation for the inverse problem of the Fisher information matrix. We have considered an application of this idea to the large correlations problem for the spectrum with neutrino mass described above. The Bayesian Fisher information does indeed reduce the correlation and error for a given parameter. In addition, this reduction is particularly significant with respect to the parameter about which we have prior information. Thus, like the entire method, this Bayesian Fisher information allows one to make regularisation with respect to the parameter about which we have information. In physics, such a parameter may be the result of previous experiments or, more interestingly, knowledge about systematic errors.

5 Conclusion and Future Work

5.1 Conclusion

This thesis is devoted to the development of experiment design optimization based on the Fisher information. The essence of the method is to minimize the error of the target parameter. The application of this technique to the spectra of tritium and axions shows the following results

- The strategy optimisation for the beta spectrum of tritium results in reducing the target parameter error by ~ 1.6 times. One can say that the required time is hence reduced by ~ 2 times. This means that an experiment might be performed 2x faster to achieve the same level of accuracy. However, whether this is true is an open question, since the maximum time in the optimal strategy increased 4-fold
- The strategy optimization for the axion spectrum leads to a reduction of the target parameter error by ~ 1.6 - 2.8 times for different axion masses. This result may be too large due to the lack of systematic errors
- Parameter spaces showed that this improvement might be explained in terms of normalization of the multivariate distribution of parameters relative to the target parameter. In other words, the correlations between the target and nuisance parameters are reduced and the distribution of the target parameter is more normal. It means that systematic effects have less influence. And it is the most important result that it is possible to remove systematic errors

The development of the Bayesian approach deserves special attention. The result that prior information about a nuisance parameter may reduce the error of a target parameter is very interesting. And the fact that this effect is not present in an optimal strategy represents the reduction in correlations, i.e., the influence of other parameters.

Thus, the method can be used in two ways:

- To evaluate errors and correlations or even to predict features of the fitting.
- To optimize the time strategy of an experiment that measures a spectrum enables to measure some parts of it longer than others and has a target parameter.

5.2 Future Work

Although the Fisher information is a powerful tool, it is only local information since the Fisher information is defined in one point. A real interest is to use global information rather than local information. This means that one should work with the whole distribution at once. Such global information will give more stable and more informative results.

This leads to some problems, for instance, convolution, a mathematical operation that helps to sum two distributions, is quite computationally intensive. In the thesis, convolution was used to make the strategies closer to the real situation when the instrument accuracy does not allow the time strategy to be set correctly. So the strategy was convoluted with the normal distribution to simulate a more realistic situation. However, the development of quantum computer might solve this problem.

The Bayesian approach is to work with a distribution that helps to include more information in a model. However, limitations in computing power have hindered development in this area. Therefore, developing a new methodology for Fisher information in Bayesian statistics is the most promising area for further work. In addition, it is interesting to design an optimal setup scheme so that we can develop both the best strategy and the best setup. In conclusion, this method has been applied to a "toy" example and not to real data. Therefore, applying it to real data may yield new interesting results.

References

[1] A. Belesev et al. "The search for an additional neutrino mass eigenstate in the 2-100 eV region from 'Troitsk nu-mass' data: A detailed analysis". In: *Journal of Physics G Nuclear and Particle Physics* (July 2013). DOI: 10.1088/0954-3899/41/1/015001.

[2] Gary J. Feldman and Robert D. Cousins. "Unified approach to the classical statistical analysis of small signals". In: *Phys. Rev. D* 57 (7 Apr. 1998), pp. 3873–3889. DOI: 10.1103/PhysRevD.57.3873. URL: https://link.aps.org/doi/10.1103/PhysRevD.57.3873.

[3] Alexander Etz. "Introduction to the Concept of Likelihood and Its Applications". In: *Advances in Methods and Practices in Psychological Science* 1.1 (2018), pp. 60–69. DOI: 10.1177/2515245917744314. eprint: https://doi.org/10.1177/2515245917744314. URL: https://doi.org/10.1177/2515245917744314.

[4] Jonas Zmuidzinas. "Cramér–Rao sensitivity limits for astronomical instruments: implications for interferometer design". In: *J. Opt. Soc. Am. A* 20.2 (Feb. 2003), pp. 218–233. DOI: 10.1364/JOSAA.20.000218. URL: http://josaa.osa.org/abstract.cfm?URI=josaa-20-2-218.

[5] Tracianne B. Neilsen et al. "Optimal experimental design for machine learning using the Fisher information matrix". In: *The Journal of the Acoustical Society of America* 144.3 (2018), pp. 1730–1730. DOI: 10.1121/1.5067675. eprint: https://doi.org/10.1121/1.5067675. URL: https://doi.org/10.1121/1.5067675.

[6] J. Sourati et al. "Intelligent Labeling Based on Fisher Information for Medical Image Segmentation Using Deep Learning". In: *IEEE Transactions on Medical Imaging* 38.11 (2019), pp. 2642–2653. DOI: 10.1109/TMI.2019.2907805.

[7] Burr Settles. *Active Learning Literature Survey*. Computer Sciences Technical Report 1648. University of Wisconsin–Madison, 2009.

[8] Sinno Jialin Pan and Qiang Yang. "A Survey on Transfer Learning". In: *IEEE Transactions on Knowledge and Data Engineering* 22 (2010), pp. 1345–1359.

[9] Jerry Chao, E. Sally Ward, and Raimund J. Ober. "Fisher information theory for parameter estimation in single molecule microscopy: tutorial". In: *J. Opt. Soc. Am. A* 33.7 (July 2016), B36–B57. DOI: 10.1364/JOSAA.33.000B36. URL: http://josaa.osa.org/abstract.cfm?URI=josaa-33-7-B36.

[10] B. Roy Frieden, Raymond J. Hawkins, and Joseph L. D'Anna. "Financial Economics from Fisher Information". In: *Exploratory Data Analysis Using Fisher Information*. Ed. by B. Roy Frieden and Robert A. Gatenby. London: Springer London, 2007, pp. 42–73. ISBN: 978-1-84628-777-0. DOI: 10.1007/978-1-84628-777-0_2. URL: https://doi.org/10.1007/978-1-84628-777-0_2.

[11] V.A. Nastasiuk. "Fisher information and quantum potential well model for finance". In: *Physics Letters A* 379.36 (2015), pp. 1998–2000. ISSN: 0375-9601. DOI: https://doi.org/10.1016/j.physleta.2015.06.052. URL: https://www.sciencedirect.com/science/article/pii/S037596011500571X.

[12] A Baranov et al. "Optimising the Active Muon Shield for the SHiP Experiment at CERN". In: *Journal of Physics: Conference Series* 934 (Dec. 2017), p. 012050. DOI: 10.1088/1742-6596/934/1/012050. URL: https://doi.org/10.1088/1742-6596/934/1/012050.

[13] Gaia Franceschini and Sandro Macchietto. "Model-based design of experiments for parameter precision: State of the art". In: *Chemical Engineering Science* 63.19 (2008). Model-Based Experimental Analysis, pp. 4846–4872. ISSN: 0009-2509. DOI: https://doi.org/10.1016/j.ces.2007.11.034. URL: https://www.sciencedirect.com/science/article/pii/S0009250907008871.

[14] Ali Shahmohammadi and Kimberley B. McAuley. "Sequential Model-Based A-Optimal Design of Experiments When the Fisher Information Matrix Is Noninvertible". In: *Industrial Engineering Chemistry Research* 58.3 (2019), pp. 1244–1261. DOI: 10.1021/acs.iecr.8b03047.

[15] Charles R. Harris et al. "Array programming with NumPy". In: *Nature* 585.7825 (Sept. 2020), pp. 357–362. DOI: 10.1038/s41586-020-2649-2. URL: https://doi.org/10.1038/s41586-020-2649-2.

[16] Joshua A. Frieman, Savas Dimopoulos, and Michael S. Turner. "Axions and stars". In: *Phys. Rev. D* 36 (8 Oct. 1987), pp. 2201–2210. DOI: 10.1103/PhysRevD.36.2201. URL: https://link.aps.org/doi/10.1103/PhysRevD.36.2201.

[17] Theopisti Dafni et al. "Weighing the solar axion". In: *Phys. Rev. D* 99 (3 Feb. 2019), p. 035037. DOI: 10.1103/PhysRevD.99.035037. URL: https://link.aps.org/doi/10.1103/PhysRevD.99.035037.

[18] Zbigniew Michalewicz Raymond Chiong Thomas Weise. *Variants of Evolutionary Algorithms for Real-World Applications*. Springer-Verlag Berlin Heidelberg, 2012, pp. XIV, 466. DOI: https://10.1007/978-3-642-23424-8.

[19] Marco Dorigo and Thomas Stützle. *Ant Colony Optimization*. MIT Press, 2004.

[20] Panos M. Pardalos, H.Edwin Romeijn, and Hoang Tuy. "Recent developments and trends in global optimization". In: *Journal of Computational and Applied Mathematics* 124.1 (2000). Numerical Analysis 2000. Vol. IV: Optimization and Nonlinear Equations, pp. 209–228. ISSN: 0377-0427. DOI: https://doi.org/10.1016/S0377-0427(00)00425-8. URL: https://www.sciencedirect.com/science/article/pii/S0377042700004258.

[21] S. Eddy. "What is Bayesian statistics?" In: *Nat Biotechnol* 22 (2004), pp. 1177–1178. DOI: 10.1038/nbt0904-1177.

[22] Christian Clason. *Regularization of Inverse Problems*. 2021. arXiv: 2001.00617 [math.FA].

[23] Andrey N. Tikhonov and Vasiliy Y. Arsenin. *Solutions of ill-posed problems*. Translated from the Russian, Preface by translation editor Fritz John, Scripta Series in Mathematics. Washington, D.C.: John Wiley & Sons, New York: V. H. Winston & Sons, 1977, pp. xiii+258.

[24] Per Christian Hansen. "The truncatedSVD as a method for regularization". In: *BIT Numerical Mathematics* 27 (4 Dec. 1987), pp. 534–553. DOI: 10.1007/BF01937276. URL: https://doi.org/10.1007/BF01937276.