A Log-Likelihood Fit for Extracting Muon Neutrino Oscillation Parameters

01707971

Imperial College London

***Abstract* — Data from T2K experiment on neutrino oscillations was analysed to produce new measurements of muon-tau neutrinos mixing parameters. Many numerical methods of optimisation were discussed, compared and implemented to minimise the negative log likelihood of the data, yielding mixing angle $\boldsymbol{\theta_{23} = (1.00 \pm 0.06)\pi/4}$, and difference in their squared masses $\boldsymbol{\Delta m^2_{23} = 3.04 \pm 0.03\ 10^{-3}\ eV^2}$. These parameters perfectly resembled data with p-value $\boldsymbol{0.885}$, assuming muon neutrinos' cross section linearly grew with energy with proportionality constant $\boldsymbol{\alpha}$ measured as $\boldsymbol{1.12 \pm 0.05\ GeV^{-1}}$.**

I. Introduction

The Standard Model encapsulates our knowledge of three of the four fundamental interactions (strong, weak and electromagnetic) and all known elementary particles.

A major advancement in modern physics was the discovery of neutrino oscillations, whereby neutrinos can change their lepton family number. Violating the well-established lepton number conservation law and the assumption neutrinos are massless, this represents a significant challenge to the Standard Model, and strong evidence for new physics.

T2K ("Tokai to Kamioka") experiment measures neutrino oscillations after they travelled 295 km [1]. In this report, data from T2K experiment is analysed and fitted with different numerical methods to derive estimates of oscillation parameters of muon neutrinos ν_μ.

The data consisted of m_i observations of ν_μ events at different energies arranged in 200 equally spaced bins, as plotted red in Fig 1.1. A simulation was also run of the expected number of ν_μ events assuming no neutrino oscillations and constant cross section of ν_μ events, plotted (blue) in Fig 1.1.

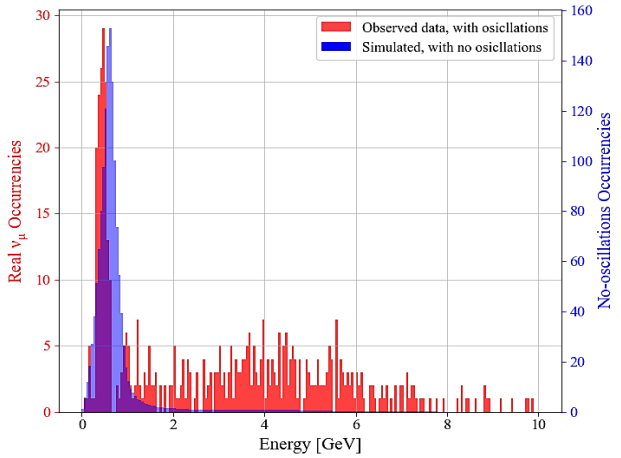

Figure 1.1 Observed data (red) compared to the simulated with no oscillations (blue). Note how the relative height of the tail (E > 1 GeV) with respect to the peak is much bigger in the real data than the simulated.

The shapes of the datasets highly differ in their tail, i.e. for $E \geq 1$ GeV, with the simulated dataset having much lower values compared to its peak. To explain such discrepancy, an alternative hypothesis is discussed in section V, which is the cross section of ν_μ events is not constant, but linearly increases with energy.

II. Theory

A. Neutrino Oscillations

The survival probability of a muon neutrino with energy E [GeV] after having travelled a distance L [km] is theoretically approximated by

$$P(\nu_\mu \to \nu_\mu) = 1 - \sin^2(2\theta_{23}) \sin^2\left(1.267 \frac{\Delta m^2_{23} L}{E}\right), \quad (1.1)$$

being θ_{23} the muon-tau mixing angle and Δm^2_{23} [eV^2] the difference between the squared masses of tau and muon neutrinos [1]. This probability is plotted in Fig 2.1, presenting many oscillations at very low energies $E < 1$ GeV.

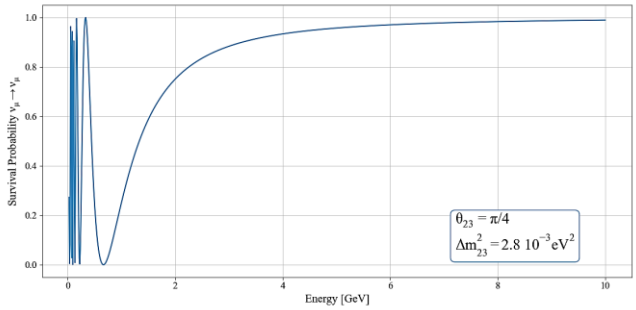

Figure 2.1 The survival probability $P(\nu_\mu \to \nu_\mu)$ as a function of energy. It presents many oscillations at very low energy E < 1 GeV.

B. The Statistics

Multiplying the results of the simulated data by the survival probability (1.1) yields the expected number $\lambda_i \equiv \lambda(E_i; \theta_{23}, \Delta m^2_{23})$ of ν_μ events to which compare our data.

To fit the oscillation parameters, we first define the likelihood $\mathcal{L}$ of the n measurements $\{m_i\}$

$$\mathcal{L} = \prod_{i=1}^{n} \mathcal{P}(\lambda_i; m_i), \quad (2.1)$$

being $\mathcal{P}(\lambda_i; m_i)$ the probability of obtaining the observed m_i knowing its expectation λ_i: given the low statistics, a Poisson distribution is used, with form

$$\mathcal{P}(\lambda; m) = \frac{\lambda^m e^{-\lambda}}{m!}. \quad (2.2)$$

Being λ_i a function of the oscillation parameters, so is $\mathcal{L}$.

The best fit of the parameters is then the one maximising the likelihood, or, alternatively, minimising the Negative Log Likelihood NLL, with

$$\mathrm{NLL} = -\ln(\mathcal{L}) = -\sum_{i=1}^{n} \ln[\mathcal{P}(\lambda_i;\, m_i)]. \qquad (2.3)$$

Substituting (2.2) in (2.3) gives

$$\mathrm{NLL}(\vec{u}) = \sum_{i=1}^{n} [\lambda_i(\vec{u}) - m_i \ln \lambda(\vec{u}) + \ln(m_i!)], \quad (2.4)$$

where the dependence of the NLL on the parameters' vector $\vec{u} = (\theta_{23}, \Delta m^2_{23})$ is made explicit. As shown in Fig 1.1, the data m_i are all lower than 30, with many datapoints being lower than 5: this suggests application of Stirling's approximation [1] should be avoided, as being, for these relatively small values, a poor approximation.

The NLL as a function of $\vec{u}$ can be observed in the contour plot in Fig 2.2. It must be noted the NLL is symmetric with respect to $\theta_{23} = \pi/4$, as (1.1) is. Considering that (1.1) is periodic in $\pi/2$, we can consider θ_{23} in the range $[0, \pi/4]$ only, as fully representing the problem.

Minimising the NLL to $\mathrm{NLL}_{\mathrm{min}}$, the 1σ (~68%) confidence limit is found by determining the points $\vec{u}$ where $\mathrm{NLL}(\vec{u}) = \mathrm{NLL}(\hat{u}) + 0.5$, hence solving equation

$$\mathrm{NLL}(\vec{u}) - \mathrm{NLL}_{\mathrm{min}} - 0.5 = 0. \qquad (2.5)$$

For all numerical calculations in later sections, a set of scaled units was used, with θ_{23} being expressed as a fraction of $\pi/4$, and Δm^2_{23} in terms of $m_0^2 = 10^{-3}$ eV^2. For clarity, in this report the parameters will always be expressed in their full non-scaled form.

III. 1-Dimension Optimisation

In this section, a 1D minimisation of the NLL is conducted over θ_{23}, keeping Δm^2_{23} constant. A parabolic minimiser is used: three initial coordinates are first taken in proximity of the minimum and interpolated to a parabola; the minimum of the interpolation is then found and substituted with the largest of previous points, repeating until convergence. The minimiser I wrote was first tested on 2 analytically known functions: a parabola and a cosine.

The estimate of Δm^2_{23} was chosen from Fig 2.1 (b): the NLL is substantially lower in the 2.6-3.1 $\times 10^{-3}$ eV^2 interval, so $\Delta m^2_{23} = 2.8 \times 10^{-3}$ eV^2 was chosen.

Then, three initial θ_{23} coordinates were picked by evaluating the NLL at 500 different points in the $[0, \pi/4]$ interval, taking the value leading to the minimum NLL, and the adjacent ones. This ensured closeness of the initial guesses to the real minimum. It is in fact of prime importance for the minimiser that the initial points lie in a parabolic-like region of the function, so that the interpolation can effectively resemble the real function.

This procedure led to $\theta_{23} = 0.926^{+0.018}_{-0.017}\, \pi/4$, within a 1σ confidence level, as defined in the end of the section II. (2.5) was solved by application of the bisection method and confirmed by brute-force scanning of NLL around the minimum.

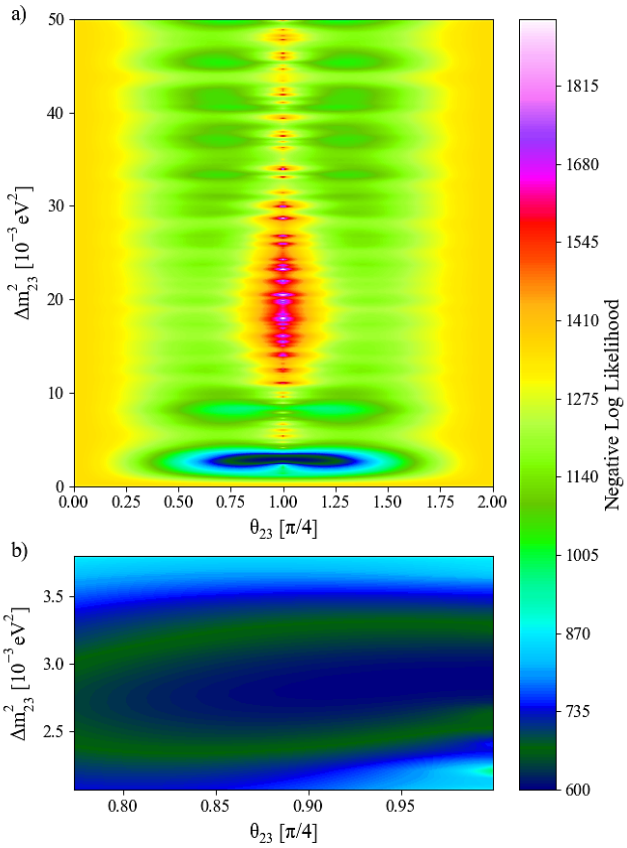

Figure 2.2. a) The NLL is plotted over all possible θ_{23} and a vast range of Δm^2_{23} values. Note the NLL is symmetric with respect to $\theta_{23} = \pi/4$. b) The NLL is plot in the main region of interest, where the NLL is lower, which is for θ_{23} in the $[0.8, 1]\pi/4$ interval and Δm^2_{23} approximately going from 2.6 to 3.1 10^{-3}eV^2.

The bisection method was chosen due to its robustness and the proximity of the problem to the minimum of the function. This, in fact, could have negatively affected the Newton-Raphson and secant methods, as the 1st derivative, which they use, is expected to tend to zero in such region.

The uncertainty could also be inferred from the curvature

$$\Sigma = \frac{\partial^2 \mathrm{NLL}}{\partial \theta_{23}^2} \qquad (3.1)$$

about the minimum. Provided the NLL about the minimum resembles a parabola, in fact, we have

$$1\sigma = \frac{1}{\sqrt{\Sigma}}. \qquad (3.2)$$

This evaluated to $0.018\, \pi/4$, showing perfect agreement with the previous method. Such agreement in turn legitimates the parabolic approximation, hence the implementation of the parabolic minimiser in the first place.

However, it must be noted it fails to capture any asymmetry of the interval. For this reason and as it cannot be assumed a priori the multidimensional NLL will show such resemblance with a quadratic, in next sections (2.5) will be used to define the 1σ interval.

IV. Multidimensional Optimisation

For a more accurate result, a 2D minimisation should be performed to fit both θ_{23} and Δm^2_{23}. In this section, first, the many different methods considered and tested are

compared; then the results of the minimisation on NLL are presented and discussed.

All derivatives involved in next sections were computed with central difference scheme and 4th order accuracy, representing a good balance between accuracy and efficiency.

A. *Comparison of Methods*

The considered methods are Univariate, Newton's, Quasi-Newton, and an original version of Metropolis with simulated annealing. These were tested in 2 and 3 dimensions, with appropriate functions:

$$f_1(x,y) = 2x^3 + 6xy^2 - 3y^3 - 150x^{[2]},$$
$$f_2(x,y,z) = -\operatorname{sinc}(x)\operatorname{sinc}(y-\pi)\operatorname{sinc}(z+\pi).$$

f_1 was chosen as presenting not only a local minimum at [5,0], but also saddles and local maximum, providing interesting obstacles to the methods being tested. f_2 goes even further: it has one global minimum $[0,\pi,-\pi]$ and countless local maxima and minima.

The shape of $f_1(x,y)$ can be observed in Fig 4.1, with critical points being the local minimum X(5,0), the maximum M(−5,0), the saddles S(3,4) and (−3,−4). It also shows the performance and results of the minimisers being tested.

The first method examined is the Univariate. It is the extension of the parabolic minimiser to multidimensional problems: it interpolates a parabola to 3 points in each direction successively, then iterates. As shown in Fig 4.1 (a) it is very robust, provided the initial guess is sufficiently close to the local minimum; oppositely, a less appropriate choice would lead to very wrong results.

The second method considered is Newton's. Starting from an initial guess, it iterates finding a new point at each step as

$$\vec{x}_{k+1} = \vec{x}_k - \mathrm{H}_k^{-1}\nabla f_k, \qquad (4.1)$$

being ∇f_k the gradient and $\mathrm{H_k}$ the curvature or Hessian matrix at previous step k. As shown in Fig 4.1 (a), as expected [4], it can be attracted to any critical point, hence having an appropriate starting point is of prime importance.

To solve this problem, I also tested a Newton-Gradient method: when (4.1) would be leading to an increase in the function, it is substituted by a gradient descent's step

$$\vec{x}_{k+1} = \vec{x}_k - \alpha\nabla f_k, \qquad (4.2)$$

with $\alpha \ll 1$ chosen as 10^{-2}. This forces the method to move to a smaller value of f, even if it may imply descending with no bounds until an N^{th} iteration is reached.

The Quasi-Newton methods has similar iteration step

$$\vec{x}_{k+1} = \vec{x}_k - \alpha_k \mathrm{G}_k\nabla f_k, \qquad (4.3)$$

where G_k approximates H_k^{-1} and is directly derived from G_{k-1}, with no 2nd derivatives being computed. α_k is derived at each step by a backtracking line search, reducing up to n times a given $\alpha_{\max}$ up to $\alpha_{\min}$, until it satisfies Armijo-Wolfe condition

$$f(\vec{x}_{k+1}) \le f(\vec{x}_k) - c_1(\mathrm{G}_k\nabla f_k)\cdot\nabla f_k, \qquad (4.4)$$

thus, ensuring a decrease in f [3]. α_{max} was chosen as 2.0, $\alpha_{\min}$ as 10^{-4}, n was set to 50 and $c_1 = 10^{-4}$. The algorithms considered to update G_k were DFP and BFGS, both having the advantage of keeping G_k positive

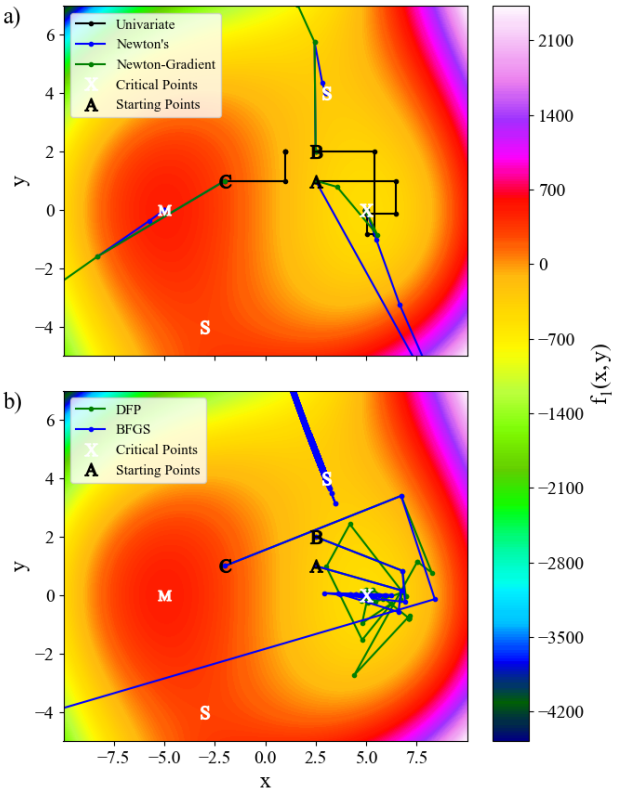

Figure 4.1. Comparison of Minimisation Methods. a) Univariate converges in few steps for initial guesses near the minimum but gets very wrong otherwise; Newton's can be attracted to any critical point; Newton-Gradient descends the function, either finding a minimum or continuing, possibly, indefinitely. b) DFP always converged to the local minimum, although with many steps; BFGS has more probability of convergence to the minimum than Newton's, but may also overshoot and/or oscillate about other critical points, such as S(3,4), without converging.

definite [3]. Fig 4.1 (b) shows BFGS, interestingly, usually overshot and retraced. It converged from starting points $A(2.5,1)$ and $B(2.5,2)$ (which is better than Newton's) but not $C(-2,1)$, in which case it overshot towards negative xs, retraced, oscillated many iterations around saddle (3,4), then repeated, never converging to such saddle, oppositely to Newton's, due to (4.4).

DFP, despite involving, on average, more steps than Newton methods, always converged to the local minimum, even when previous algorithms failed.

3D testing on f_3 confirmed the importance of initial guesses: all methods failed from inappropriate starting guesses, presumably due to the bumpiness of the function and its richness of local minima.

Monte Carlo algorithms are totally different. The Metropolis algorithm I wrote, first, performed a uniformly-random scanning of the function. After having moved to the point it found being the minimum, it iteratively took a new point $\vec{x}_{k+1}$, randomly, based on a gaussian distribution centred at the previous $\vec{x}_k$. The new point could be accepted or not, with probability

$$p_{k+1} = \begin{cases} 1 & \text{if } f(\vec{x}_{k+1}) < f(\vec{x}_k) \\ \exp\left[\dfrac{f(\vec{x}_{k+1}) - f(\vec{x}_k)}{T_{k+1}}\right] & \text{otherwise} \end{cases}, \qquad (4.3)$$

with the temperature T_{k+1} of the system being slightly reduced at every step.

The user could decide to vary the deviation of the gaussian step based on the closeness of the point to the current minimum, and to request a halving of the interval being scanned after some iterations, to improve accuracy.

Different choices of parameters were examined: the more aggressive ones (low temperature, fast annealing) improve the precision about the current minimum, however risking such current minimum is not the global; the safer choices let the algorithm scan many different regions, enabling it to approximately find the global minimum, but with relatively high uncertainty, as shown in Table 4.1.

TABLE 4.1: ACCURACY IN METROPOLIS MINIMISER

The average distance of the results of the Metropolis minimiser on f_3 from its true global minimum after 10,000 iterations and on 100 trials, for different settings.

Number of initial scans	Aggressive parameters	Safe parameters (halving)	Safe parameters (no halving)
200	1.62	0.34	4.51
500	1.33	0.25	4.23
3,000	0.76	0.26	1.96
10,000	0.12	0.25	0.69

Ultimately, the best choice is determined by the requirements of the task.

B. *Results*

Sensible choices of initial guesses for the minimisation methods were estimated from Fig 2.2, which also showed the function is smooth about the minimum. Thus, all methods could be used.

Univariate was first implemented, minimising the variables in both possible orders, using initial coordinates

$$\theta_{23} = [0.9, 0.95, 1.00]\pi/4,$$
$$\Delta m^2_{23} = [2.7, 2.9, 3.1]10^{-3}\text{eV}^2,$$

yielding same result. Then, Newton's, Newton-Gradient and DFP methods were used, and the best fit was taken. Finally, the uncertainty was estimated by solving (2.5) by means of bisection method, yielding $\theta_{23} = (1.00 \pm 0.05)\pi/4$ and $\Delta m^2 = 2.90 \pm 0.02\ 10^3\ \text{eV}^2$.

The agreement between data and tuned theoretical model is low, as shown in Fig 4.2: at low energies $E < \sim 1$ GeV the expectation gets twice as high as the data, for bigger energies it is much lower.

The poor agreement can be measured by a Chi-squared test. Being a low statistic, Williams' correction was applied [5]. This yielded reduced Chi-squared $\chi^2_{red} = 10.01$ and corresponding p-value 9.99×10^{-16}, meaning we should reject this hypothesis (please see Appendix A for more information on William's correction and calculation of p-value).

V. ALTERNATIVE HYPOTHESIS

Previous results implicitly rely on the assumption the cross section of ν_μ events, directly proportional to the observed number of such events, was constant with energy. In previous section we saw our expectation should be lower for small energy and bigger for big energy, thus

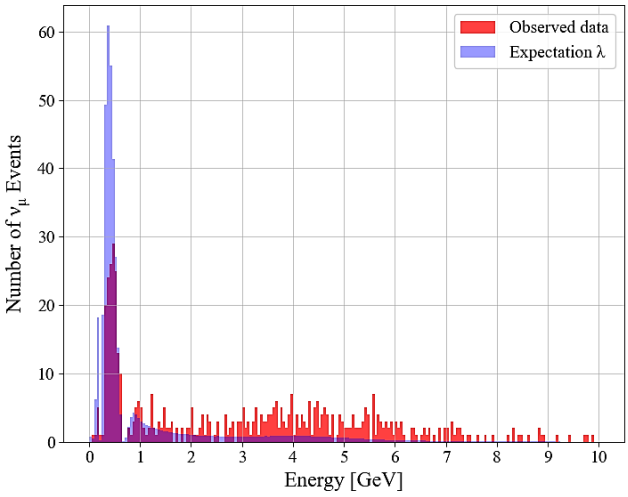

Figure 4.2. Observed number of ν_μ events (red) and theoretical expectation (blue) plotted vs energy, showing poor agreement: the expectation is twice as high for low energy and much smaller for bigger energy. Here $\theta_{23} = \pi/4$, $\Delta m^2_{23} = 2.90\ 10^{-3}\text{eV}^2$.

suggesting the cross section, and hence the expectation λ_i, should linearly increase with energy, hence

$$\lambda_i = \alpha E_i \lambda_{i_{old}}, \tag{5.1}$$

with α being the proportionality constant we ought to define. Let's observe how the NLL varies with α, shown in Fig 5.1: it is minimum at about $\alpha = 1$ GeV, then steadily increases.

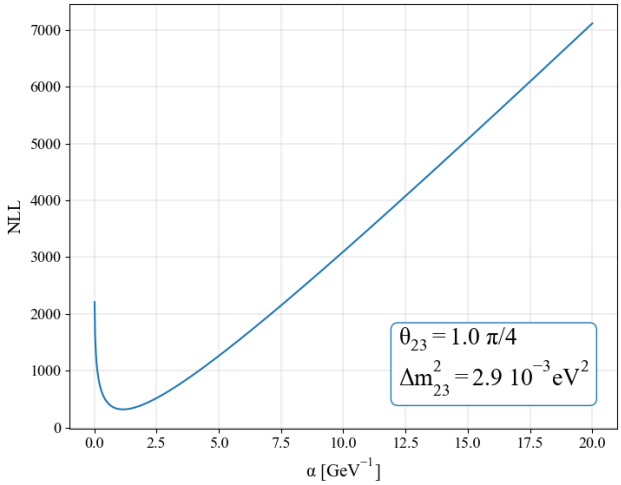

Figure 5.1 NLL vs α, with θ_{23} and Δm^2_{23} set to the values found in section IV. It is minimum at $\alpha \approx 1$ GeV^{-1}, then increases steadily.

Having introduced a new parameter, the previous must be redefined. To obtain a starting guess the Metropolis algorithm was implemented: as only wanting a rough estimate, a safe choice of parameters was used.

Then, its result was used as starting point for Newton's, Newton-Gradient and DFP minimisers. They all gave same result. Finding the 1σ uncertainty by solving (2.5), it yielded

$$\theta_{23} = (1.00 \pm 0.06)\pi/4$$
$$\Delta m^2_{23} = 3.04 \pm 0.03\ 10^{-3}\ \text{eV}^2.$$
$$\alpha = 1.12 \pm 0.05\ \text{GeV}^{-1}$$

Applying William's correction, these yield a reduced $\chi^2_{\text{red}} = 0.881$ and p-value 0.885: data and theory agree, as shown in Fig 5.2.

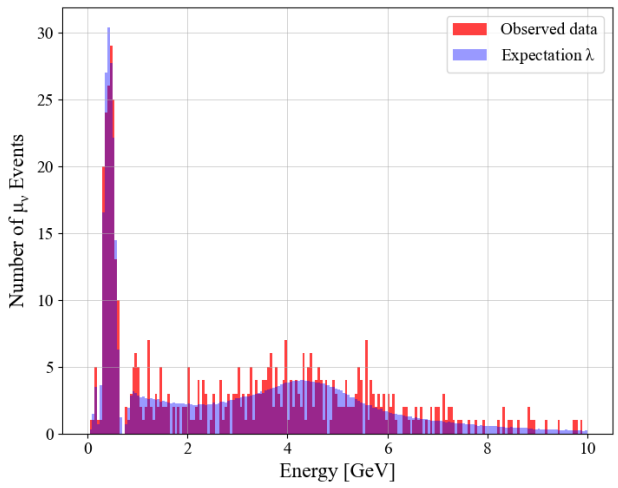

Figure 5.2. Observed number of ν_μ events (red) and the new theoretical expectation (blue) plotted vs energy, showing high agreement. Oppositely to the previous model, the heights of the peaks match, as well as the higher-energy behavior. Here $\theta_{23} = \pi/4$, $\Delta m^2_{23} = 3.04\ 10^{-3}\text{eV}^2$, $\alpha = 1.12$.

VI. Conclusion

The aim of the investigation was to determine the value of the parameters defining the $\nu_\mu \to \nu_\tau$ oscillations, by comparing data from T2K and theoretical expectations.

After having discussed and implemented many numerical optimisation methods, the oscillation angle was determined being, within a 1σ confidence level, $\theta_{23} = (1.00 \pm 0.06)\pi/4$, hence leading to an amplitude in (1.1) $\sin^2(2\theta_{23}) = 1.000_{-0.004}$.

The difference in the squared masses was estimated, within 1σ confidence level, as $\Delta m^2_{23} = 3.04 \pm 0.03\ 10^{-3}\ \text{eV}^2$.

It was observed that the data and the theoretical model are only compatible if the ν_μ cross section linearly increased with energy, with proportionality constant $\alpha = 1.12 \pm 0.05\ \text{GeV}^{-1}$.

The constant-cross section hypothesis was in fact rejected with p-value 10^{-15}, while the linearly increasing hypothesis yields a p-value 0.885.

VII. References

[1] Scott M, Dauncey P. Project 1: A log-likelihood fit for extracting neutrino oscillation parameters. Nov 2021.

[2] Max/min for functions of two variables. Available at http://personal.maths.surrey.ac.uk/st/S.Zelik/teach/calculus/max_min_2var.pdf. Last visited: 14/12/2021.

[3] Nocedal J, Wright SJ. Numerical Optimization. Springer 2006. 2nd Edition.

[4] Dauphin Y, Pascanu R, Gulcehre C, Cho K, Ganguli S, Bengio Y. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. NIPS. 2014; 27.

[5] McDonald, J.H. Handbook of Biological Statistics (3rd ed.). Sparky House Publishing, Baltimore, Maryland. 2014. Pages 86-89. http://www.biostathandbook.com/small.html.

[6] Weisstein EW. Chi-Squared Distribution. MathWorld -- A Wolfram Web Resource.https://mathworld.wolfram.com/Chi-SquaredDistribution.html. Last visited: 14/12/2021.

[7] P-Value Calculator for Chi-Square Distribution. University of Illinois at Urbana-Champaign. Department of Statistics.http://courses.atlas.illinois.edu/spring2016/STAT/STAT200/pchisq.html. Last visited: 14/12/2021.

Appendix A

An estimation of the goodness of the fit can be obtained by performing the well-known Chi-squared test.

For a low statistic, such as the one treated in this paper, with the great majority of bins having much less than 20 entries, a correction to the Chi-squared should be applied. Having more than one parameter to be fitted, William's correction was applied [5]: the standard χ^2 is divided by a factor

$$q = 1 + \frac{N^2 - 1}{6n\nu} \qquad (A.1)$$

where N is the number of bins, n is the sum of all entries of all bins, ν is the number of degrees of freedom. William's correction has the effect of slightly lowering the χ^2, hence slightly increasing the p-value.

Such p-value is calculated from

$$p = 1 - \frac{\gamma\left(\frac{\nu}{2}, \frac{\chi^2}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}, \qquad (A.2)$$

where Γ and γ are the gamma and the incomplete gamma functions [6].

Being they integrals, Extended Simpsons' Rule of integration was used with 5^{th} order accuracy, achieved by combining two successive iterations of the Trapezoidal Rule, with iterative step

$$S_j = \frac{4}{3}T_{j+1} - \frac{1}{3}T_j. \qquad (A.3)$$

The integrating method was first tested onto known p-values [7], then, applied on this problem.