We describe a new method based on the CL$_s$ approach [1] to present results in searches of new physics under the condition that the relevant nuisance parameter space is continuous. Essentially, this method uses a Gaussian approximation for the distribution of a log-likelihood ratio test statistic used in a two-hypotheses testing problem, when the sample size is large. It leads to a simple way to form exclusion sets for the parameters of interest. This Gaussian CL$_s$ method can be seen as an alternative to the traditional method of setting confidence intervals based on p-values. And one advantage of the Gaussian CL$_s$ method is that, it requires a simple procedure that is applicable in many situations where the p-value method requires computationally-intensive Monte Carlo instead of a popular thresholding rule based on a Chi-square distribution. We illustrate this method in a problem of searching for a sterile neutrino. Specifically, we derive exclusion sets in the two-dimensional parameter space of $(\sin^2 2\theta, |\Delta m^2|)$ based on the Gaussian approximation, and demonstrate that it provides adequately accurate results. Finally, we discuss the advantages and disadvantages of our method in comparison to the traditional methods of setting confidence intervals.

I. INTRODUCTION

The Standard Model of the particle physics has been extremely successful since its establishment in mid-1970s. In particular, the Higgs-like particle discovered at LHC in 2012 [2, 3] might complete the list of fundamental particles predicted by the minimal Standard Model. Nevertheless, there is a lot of experimental evidence that points to new physics beyond the Standard Model: neutrino oscillations indicate non-zero neutrino mass; various gravitational effects indicate the existence of non-baryonic dark matter; the accelerating expansion of our universe indicates the existence of dark energy; the large observed matter-anti-matter asymmetry indicates the existence of additional CP violation source beyond that in the quark mixing matrix, etc. Searches for new physics beyond the Standard Model have been and still are at the frontier of high energy particle physics.

In many occasions, search results are presented in terms of constraints in a continuous parameter space. One example is the sterile neutrino suggested by LSND [4], MiniBooNE [5], and reactor antineutrino anomalies [6]. In this case, results are generally shown in the two-dimensional phase space of $(\sin^2 2\theta, |\Delta m^2|)$, where $\theta$ is the mixing angle involving the sterile neutrino, and $|\Delta m^2|$ is mass-squared difference involving neutrino beyond three generations. Mathematically, the problem of searching for new physics is transformed into a problem of parameter estimation, such as estimating $\sin^2 2\theta$ and $|\Delta m^2|$ in the aforementioned example. Right now, for such problems, one of the most popular statistical methods is to present a set of plausible values of the parameters in a continuous space, called a confidence interval (CI). The construction of CIs is usually based on a test statistic that takes the form of $\Delta \chi^2(\beta) := \chi^2(\beta) - \chi^2_{\min}$, where $\beta$ denotes the vector of parameters of interest, such as $\beta = (\sin^2 2\theta, |\Delta m^2|)$. Then any given value of $\beta$ is included in the CI if its corresponding $\Delta \chi^2(\beta)$ is below a threshold based on a Chi-square distribution with appropriate degrees of freedom. A summary of this Chi-square method can be found in the review of particle physics by the particle data group [7], and a detailed example of such a test statistic is provided in Sec. II.

In this paper, we consider a popular alternative to the aforementioned construction of CIs in searching for new physics. The approach we consider is the CL$_s$ approach [1, 8]. Our new contribution is to propose and validate an approximation method to carry out the CL$_s$ approach that is computationally efficient and applicable in very general setups.

Recall the CI approach assigns parameter values $\beta$ to a confidence interval based on first calculating its corresponding test statistic $\Delta \chi^2(\beta)$ and then obtaining its corresponding p-value. In comparison, the CL$_s$ approach puts $\beta$ values into an exclusion set by first calculating a test statistic $\Delta T(\beta_0, \beta)$, which is usually some variation of the negative-two-log-likelihood ratio between a reference value, $\beta_0$, and the value being inspected, $\beta$, and then obtaining its corresponding CL$_s$ value, which is the ratio between the two p-values under $\beta_0$ and $\beta$, respectively. Here, the CL$_s$ value serves as a measure of the plausibility of $\beta$. We will show that, under regularity conditions that are much milder than that required by the Chi-square method, one can easily calculate CL$_s$ val-

---

* Corresponding author: xqian@bnl.gov
† Corresponding author: axin-tan@uiowa.edu
‡ Other examples include dark matter searches (the interaction cross section vs. the mass of the dark matter particle) and SUSY (super symmetry) particle searches at LHC (the interaction coupling vs. the mass scale).
uses using a Gaussian approximation to the distribution of \( \Delta \chi^2 \) at large data limit \([9, 10]\), hence avoid using the expensive MC method in setting exclusion sets. We will refer to this method as the Gaussian CL\(_s\) method. Another benefit brought by the Gaussian CL\(_s\) method is, it allows easy combination of exclusions from different experiments, given they are not correlated.

Here, we briefly compare the Gaussian CL\(_s\) method to the traditional method of constructing CIs based on \( \Delta \chi^2 \) (Chi-square CI method). The mathematical reasoning behind the Chi-square CI method is the Wilks’ theorem \([11]\), which states that \( \Delta \chi^2 \) with \( \beta \) being the true value follows approximately a Chi-square distribution under certain regularity conditions to be elaborated in details in Sec. II. Therefore, one can set approximate CIs based on \( \Delta \chi^2 \) and threshold it by quantiles of a Chi-square distribution. In analogy, under another set of regularity conditions to be elaborated in details in Sec. III A, the Gaussian CL\(_s\) method utilizes the approximate Gaussian distribution of the log-likelihood ratio \( \Delta T(\beta_0, \beta) \), which simplifies the calculation of CL\(_s\) values in setting exclusion sets. We emphasize that, the set of regularity conditions required by the Gaussian CL\(_s\) method is much more easily satisfied than that of the Wilks’ theorem. There are many cases in the search for new physics when both sets of regularity conditions hold, then the Gaussian CL\(_s\) method and the Chi-square CI method can both be used and provide complementary analysis. There are also many cases where the Gaussian CL\(_s\) method is applicable but the Chi-square CI method is not. This is one of the main advantages of our proposed Gaussian CL\(_s\) method.

When some of the regularity conditions of the Wilks’ theorem break down, a well known method to setting CIs is the Feldman-Cousins method \([12]\), which approximates the distribution of \( \Delta \chi^2 \) as accurately as desired using Monte Carlo (MC) samples. Despite the theoretical property that a large enough Monte Carlo sample size guarantees the validity of the resulting CIs, the Feldman-Cousins method can be computationally intensive and time consuming. Furthermore, unlike the popular Chi-square method, combining results from different experiments by strictly following the Feldman-Cousins method requires fitting the combined data from all those experiments, which makes it even more computationally intensive. (There are known examples of combining CIs, but they all require some forms of approximation to the Feldman-Cousins’ method.) In contrast, we will show that in many of these situations when it is difficult to set valid CIs, the Gaussian CL\(_s\) method can easily be used to set exclusion sets, as well as to combine exclusion sets from different experiments, and neither task requires intensive computation. This is particularly attractive in the application of searching for new physics.

This paper is organized as the following. In Sec. II, we briefly review the Chi-square CI method. We also describe the Feldman-Cousins method, which is commonly used when the Chi-square CI method is not applicable. In Sec. III, we describe the Gaussian CL\(_s\) method. Specifically, we present a mathematical proof that shows, when the data size increases, the distribution of the log-likelihood ratio statistic \( \Delta T(\beta_0, \beta) \), which we can simply write as \( \Delta T \) when \( \beta_0 \) and \( \beta \) are fixed, approaches the Gaussian distribution with mean \( \Delta T \) and standard deviation \( 2\sqrt{|\Delta T|} \). The mathematical definition of \( \Delta T \) is given later in Eq. (13). Basically, it is the value of the statistic \( \Delta T \) calculated using the Asimov data set \([13]\).

The Asimov data set refers to the most probable data set given the current true model without any statistical fluctuations nor systematic variations. Through an example of searching for a sterile neutrino, we compare the Gaussian CL\(_s\) method with the traditional CI method in Sec. IV. We further present the discussion and summary in Sec. V and Sec. VI, respectively.

II. CONFIDENCE INTERVALS

In this section, we briefly review the traditional method of setting confidence intervals in the context of neutrino oscillations. Consider a neutrino energy spectrum that consists of \( n \) energy bins. Assume that the expected number of counts in each bin is a function of the vector of parameters of main interest, \( \beta = (\sin^2 2\theta, |\Delta m^2|) \), and a vector of nuisance parameters (such as the overall normalization), \( \eta \). Let \( \Theta \), \( M \), and \( H \) denote the parameter space of \( \sin^2 2\theta \), \( |\Delta m^2| \), and \( \eta \), respectively. There are two further physical constraints: \( \sin^2 2\theta \geq 0 \) and \( |\Delta m^2| \geq 0 \). Then for the \( i \)-th bin, \( \mu_i(\sin^2 2\theta, |\Delta m^2|, \eta) \) and \( N_i \) represent the expected and the observed counts of neutrino induced reactions, respectively. When \( \mu_i \) is large enough, the distribution of \( N_i \) can be well approximated by a Gaussian distribution with mean \( \mu_i \) and standard deviation \( \sqrt{\mu_i} \).

Given any specific guess of the value of the parameters \( (\sin^2 2\theta, |\Delta m^2|, \eta) \), once the data \( x = \{N_i, i = 1, \ldots, n\} \) is observed, one can calculate the deviation of the data from the expected values \( \{\mu_i(\sin^2 2\theta, |\Delta m^2|, \eta), i = 1, \ldots, n\} \) to measure the implausibility of the hypothesized parameter value. A few commonly used definition of deviations can be found later in Eq. (5), (6), (7), and (8). For instance, when certain knowledge concerning the nuisance parameter \( \eta \) (e.g. knowledge of detecting efficiency and neutrino flux) is available, one useful definition of deviation is given by

$$
\chi^2(\sin^2 2\theta, |\Delta m^2|, \eta) = \chi^2_{stat}(\sin^2 2\theta, |\Delta m^2|, \eta) + \chi^2_p(\eta) = \sum_i \frac{(N_i - \mu_i(\sin^2 2\theta, |\Delta m^2|, \eta))^2}{(\delta N_i)^2} + \frac{(\eta - \eta_0)^2}{(\delta \eta)^2}. 
$$

(1)
Here, the general notation $\delta w$ represents the standard deviation of a variable $w$. So $\delta N_i = \sqrt{m_i}$, and the corresponding $\chi^2_{\text{stat}}$ term is called Pearson Chi-square.

Based on the test statistic $\chi^2(\sin^2 2\theta, |\Delta m^2|, \eta)$ in Eq. (1), confidence regions can be obtained for $\beta = (\sin^2 2\theta, |\Delta m^2|)$. Specifically, we first define

$$\sin^2 2\theta_{\text{min}} \text{ and } |\Delta m^2|_{\text{min}}$$

to be the best fit to the data in the sense that $(\sin^2 2\theta_{\text{min}}, |\Delta m^2|_{\text{min}}, \eta_{\text{min}}) = \arg \min_{\sin^2 2\theta, |\Delta m^2|, \eta} \chi^2(\sin^2 2\theta, |\Delta m^2|, \eta)$. Here, the minimum is taken over $\Theta \times M \times H$. The general notation $\arg \min_w h(w)$ denotes the value of $w$ that minimizes the given function $h$. Note that $(\sin^2 2\theta_{\text{min}}, |\Delta m^2|_{\text{min}})$, won’t be exactly the true value of the parameter $(\sin^2 2\theta, |\Delta m^2|)$, and a repetition of the experiment would yield a data set that corresponds to a different best guess. So instead of reporting only $(\sin^2 2\theta_{\text{min}}, |\Delta m^2|_{\text{min}})$, it is more rational to report a set of plausible values of $(\sin^2 2\theta, |\Delta m^2|)$ that fit the observed data not too much worse than that of the best fit, and state how trustworthy the set is. Indeed, for any given $(\sin^2 2\theta, |\Delta m^2|)$, let $\eta_{\text{min}}(\sin^2 2\theta, |\Delta m^2|) = \arg \min_{\eta} \chi^2(\sin^2 2\theta, |\Delta m^2|, \eta)$, and define

$$\Delta \chi^2_{\text{min}}(\sin^2 2\theta, |\Delta m^2|) \equiv \chi^2(\sin^2 2\theta, |\Delta m^2|, \eta_{\text{min}}(\sin^2 2\theta, |\Delta m^2|)) - \chi^2(\sin^2 2\theta_{\text{min}}, |\Delta m^2|_{\text{min}}, \eta),$$

then a level a confidence interval based on Eq. (2) is defined to be

$$C_a = \{ (\sin^2 2\theta, |\Delta m^2|) \in \Theta \times M : \Delta \chi^2_{\text{min}} \leq t_a \},$$

where we use the standard set-builder notation $\{ h(w) : \text{restriction } w \}$ to denote a set that is made up of all the points $h(w)$ such that $w$ satisfies the restriction to the right of the colon. The key in constructing Eq. (3) is to specify the correct threshold value $t_a$ for a given confidence level $a$. Most commonly examined confidence levels use $a = 68.3\%$ (1$\sigma$), 95.5$\%$ (2$\sigma$), 99.7$\%$ (3$\sigma$), which are linked to threshold values $t_a = 2.31, 5.99, 11.8$, respectively [7]. Note that these three values are the 68.3$\%$, 95.5$\%$ and 99.7$\%$ quantiles of the Chi-square distribution with two degrees of freedom, respectively.

The reason why these threshold values are used is that, $C_a$ is indeed constructed upon screening the entire parameter space by inspecting one point at a time, denoted by $(\sin^2 2\theta_0, |\Delta m^2_0|)$, and testing the pair of hypotheses, $H_0 : (\sin^2 2\theta, |\Delta m^2|) = (\sin^2 2\theta_0, |\Delta m^2_0|)$ versus $H_a :$ otherwise. For this test using the Chi-square statistic $\Delta \chi^2_{\text{min}}(\sin^2 2\theta_0, |\Delta m^2_0|)$, the full parameter space is $\Omega = (\sin^2 2\theta, |\Delta m^2|, \eta) \in \Theta \times M \times H$ and the null hypothesis space is $\Omega_0 = \{(\sin^2 2\theta_0, |\Delta m^2_0|) \} \times H$. According to the Wilks’ theorem [11], if certain regularity conditions hold, mainly

1. the parameter space $\Theta \times M \times H$ is a continuous space, and the model likelihood function is a smooth function (for example three times differentiable) in the parameters,

2. $\Omega$ contains an open neighborhood of the true value $(\sin^2 2\theta_0, |\Delta m^2_0|, \eta_0)$, and

3. the data size $N_i$ is large for each $i = 1, \ldots, n$,

then the statistic $\Delta \chi^2_{\text{min}}(\sin^2 2\theta_0, |\Delta m^2_0|)$ follows approximately a Chi-square distribution when $H_0$ is true. Further, the degree of freedom of this Chi-square distribution equals the difference between the dimension of $\Omega$ and that of $\Omega_0$, namely 2, in the current case. This procedure of constructing confidence intervals and its extensions have been successfully applied in many studies in order to constrain various parameters in the neutrino physics. One recent example is the first measurement of $\Delta m_{32}^2$, and the precision measurement of $\sin^2 2\theta_{13}$ from the Daya Bay collaboration [14].

Although this method, referred to as the Chi-square CI method, has been widely used in analyzing experimental data, it does not always produce confidence interval that has the correct coverage probability. Its limitations have been addressed by, for example, Feldman and Cousins [12]. One example is the searches for neutrino oscillations in the disappearance mode. The oscillation probability with $(\sin^2 2\theta, |\Delta m^2|)$ in a 2-flavor framework is written as:

$$P_i = 1 - \sin^2 2\theta \cdot \sin^2(1.27 \cdot |\Delta m^2| \cdot L/E_i^v),$$

where $L$ and $E^v_i$ are the distance neutrino travels and the neutrino energy at the $i$-th bin, respectively. Then the expected bin counts $\mu_i = E(N_i)$ are such that $\mu_i = a_i \cdot P_i + b_i$, where $a_i$ and $b_i$ are coefficients that may depend on the nuisance parameters $\eta$.

The reason why the Chi-square method fails for the above neutrino oscillations example is the following. A key middle step in the proof of the Wilks’ theorem is that conditions 1~3 above together ensure that the MLE (maximum likelihood estimator) of $(\sin^2 2\theta, |\Delta m^2|)$ has a sampling distribution close to that of a Gaussian distribution. This suggests two things. (1) When testing $H_0$ of the form: $(\sin^2 2\theta, |\Delta m^2|) = (0, |\Delta m^2_0|)$, condition 2 is violated, and hence the Wilks’ theorem does not apply no matter how large the data size is. (2) On the other hand, for testing $H_0$ of all other forms, conditions 1 and 2 are both satisfied, hence as the sample size grows to infinity, the sampling distribution of $\Delta \chi^2_{\text{min}}(\sin^2 2\theta_0, |\Delta m^2_0|)$ will eventually converge to a Chi-square distribution. However, for instance if the true value $\sin^2 2\theta_0$ is close to 0, then there could be
FIG. 1. (color online) Left panel (a): The phase space of \( \sin^2 2\theta \) vs. \( |\Delta m^2| \) in the Cartesian coordinate. Physical constraints are \( \sin^2 2\theta \geq 0 \) and \( |\Delta m^2| \geq 0 \). Right panel (b): Schematic illustration of the effective phase space of \( \sin^2 \theta \) vs. \( |\Delta m^2| \) taking into account the spectral difference measured by \( \chi^2 \) defined in Eq. (1). When \( \sin^2 2\theta = 0 \), points with different values of \( |\Delta m^2| \) will converge into a single point. This can be easily seen from Eq. (4). At \( \sin^2 2\theta = 0 \), \( |\Delta m^2| \) has no impact on the neutrino spectrum.

a non-ignorable probability that we observe a data set that results in a MLE of \( \sin^2 2\theta \) equaling to 0. This clearly prevents the sampling distribution of the MLE of \( (\sin^2 2\theta, |\Delta m^2|) \) from being closely approximated by a Gaussian distribution. Indeed, the closer \( \sin^2 2\theta \) (or \( |\Delta m^2| \)) is to 0, the larger the data size is needed to overcome the above phenomena. To say this in another way, given any fixed large data size, there is always a small neighborhood of zero, say \( B_0 \), such that if we test an \( H_0 \) of the form \( (\sin^2 2\theta, |\Delta m^2|) = (\sin^2 2\theta_0, |\Delta m^2_0|) \) where either \( \sin^2 2\theta_0 \in B_0 \) or \( |\Delta m^2_0| \in B_0 \), then the resulting test statistic \( \chi^2_{\text{min}}(\sin^2 2\theta_0, |\Delta m^2_0|) \) will not follow a distribution that can be adequately approximated by the Chi-square distribution.

The latter point can also be understood intuitively. Although the parameter space of \( \sin^2 2\theta \) vs. \( |\Delta m^2| \) is usually displayed as Fig. 1a and is flat, the effective phase space of \( (\sin^2 \theta, |\Delta m^2|) \) measured by \( \chi^2 \) defined in Eq. (1) is no longer flat (Fig. 1b). Due to the functional form of the oscillation formula, the effective phase space becomes more compact at smaller \( \sin^2 2\theta \), as differences between spectra with different values of \( |\Delta m^2| \) become smaller. Therefore, more data is needed to reach the large data limit required by the Wilks’ theorem. For \( \sin^2 2\theta_{\text{true}} = 0 \), the effective phase space only contains a single point, since \( |\Delta m^2| \) will have no impact on the neutrino spectrum. It is therefore impossible to reach the large data limit. Even for non-zero values of \( \sin^2 2\theta_{\text{true}} \), the required data could be well beyond the experimental reach.

When these regularity conditions are not satisfied, one can use the Feldman-Cousins method [12] to set confidence intervals. Below, we review how to produce a valid 1-\( \sigma \) (68%) confidence interval of \( (\sin^2 2\theta, |\Delta m^2|) \) using MC. This method can be easily generalized to build confidence intervals of any level.

Having observed data \( x = \{N_1, \cdots, N_N\} \), apply the following procedure to every \( (\sin^2 2\theta, |\Delta m^2|) \) in the parameter space \( \Theta \times M \) (fix one pair of \( (\sin^2 2\theta, |\Delta m^2|) \) at a time):

1. Calculate \( \Delta \chi^2_{\text{min}}(\sin^2 2\theta, |\Delta m^2|)^x \) with Eq. (2) based on the observed data.

2. Simulate a large number of MC samples, say \( \{x^{(j)}\}_{j=1}^N \), where \( x^{(j)} = \{N^{(j)}_1, \cdots, N^{(j)}_N\} \) is generated from the model with true parameter value \( (\sin^2 2\theta, |\Delta m^2|) \). For \( j = 1, \cdots, T \), calculate \( \Delta \chi^2_{\text{min}}(\sin^2 2\theta, |\Delta m^2|)^{(j)} \), that is Eq. (2) based on the \( j \)-th MC sample \( x^{(j)} \). This produces an empirical distribution of the statistic \( \Delta \chi^2_{\text{min}}(\sin^2 2\theta, |\Delta m^2|) \).

3. Calculate the percentage of MC samples such that \( \Delta \chi^2_{\text{min}}(\sin^2 2\theta, |\Delta m^2|)^{(j)} < \Delta \chi^2_{\text{min}}(\sin^2 2\theta, |\Delta m^2|)^x \). Then \( (\sin^2 2\theta, |\Delta m^2|) \) is included in the 1-\( \sigma \) confidence interval if and only if the percentage is smaller than 68%.

The key of the above procedure is to generate an empirical distribution of \( \Delta \chi^2_{\text{min}}(\sin^2 2\theta, |\Delta m^2|) \), as it no longer follows a Chi-square distribution. Feldman-Cousins method guarantees the validity of the resulting CIs. However, the procedure can be very time-consuming when the dimension of the vector of unknown parameters is high and/or when a fine grid of the phase space needs to be examined. In addition, the number of MC samples, \( T \), needed in order to produce an empirical distribution that leads to an accurate enough CI, in-
increases quickly as the required confidence level increases. The procedure could become inhibitory expensive if the minimization process used to find \((\sin^2 2\theta_{\text{min}}, |\Delta m^2_{\text{min}}|)\) is slow due to the existence of many nuisance parameters or other technical reasons. Furthermore, there is no simple recipe to strictly combine the CIs generated with the Feldman-Cousins method from different experiments to form an overall CI. To see this, consider a simple example where two experiments are carried out to probe the phase space of \((\sin^2 2\theta)\). For any phase space point \((\sin^2 2\theta, |\Delta m^2|)\), each experiment yields its own statistic \(\Delta \chi^2_{\text{min}} = \chi^2(\sin^2 2\theta, |\Delta m^2|) - \chi^2_{\text{min}}(\sin^2 2\theta, |\Delta m^2_{\text{min}}|)\). However, the minimum phase space point, \((\sin^2 2\theta_{\text{min}}, |\Delta m^2_{\text{min}}|)\), based on the first experiment is typically different from that of the second experiment. Once the two experiments are combined, a strict implementation of the Feldman-Cousins method requires the global minimum phase space point, which is in general unattainable from the experiment-wise minimum phase space points directly. Indeed, one has to redo Monte Carlo simulation under the setting of the combined data, which is extremely heavy computationally since minimization has to be done for each Monte Carlo sample.

In the next section, we introduce a new method to carry out the CLs approach, that is able to produce exclusion sets quickly in searching for new physics. Also, this method allows simple combination of results from different experiments to provide global analysis.

III. THE CLs APPROACH BASED ON THE \(\Delta T\) STATISTIC AND ITS GAUSSIAN APPROXIMATION

A. The CLs Approach Based on the \(\Delta T\) Statistic

The CLs approach [1, 8] is a popular approach to present searches for new physics beyond the Standard Model. Recent examples of using this approach in neutrino physics can be found in Ref. [16, 17]. The CLs approach is complementary to the traditional approach with confidence intervals: CL, is appropriate in setting exclusion limits, while the confidence interval is appropriate in treating established signals [1]. In this section, we will briefly review the principle of the CLs approach in a two-hypotheses testing problem.

Suppose a problem involves two hypotheses: the null hypothesis \(H_0\) (typically the Standard Model) with predicted number of events \(\mu(\eta) = (\mu_1(\eta), \cdots, \mu_N(\eta))\) and the alternative hypothesis \(H_1\) (typically a new physics model) with predicted number of events \(\nu(\zeta) = (\nu_1(\zeta), \cdots, \nu_N(\zeta))\). Here, \(\eta\) and \(\zeta\) represent the vectors of nuisance parameters under the two hypotheses respectively, and they could be of different dimension. We now introduce a test statistic, usually denoted by \(\Delta T\), for testing \(H_0\) and \(H_1\). Note that there is more than one version of the definition of \(\Delta T\) that we will list below.

To lead to the definition of \(\Delta T\), we start with either the Poisson or the Normal distribution to model the data \(x\), and use the general notation \(L(x, \lambda)\) to denote the corresponding likelihood, where \(\lambda\) equals to \(\mu(\eta)\) under \(H_0\), and \(\nu(\zeta)\) under \(H_1\), respectively. Following the convention of Ref. [18], let

\[
a(x) = 2 \log L(x, \lambda = x),
\]

then define

\[
\begin{align*}
T_{H_0}(x, \eta) &= -2 \log(L(x, \mu(\eta))) + a(x), \quad \text{and} \\
T_{H_1}(x, \zeta) &= -2 \log(L(x, \nu(\zeta))) + a(x).
\end{align*}
\]

Take the Poisson model for example, we have

\[
2 \log L(x, \lambda) = \sum_{i=1}^{n} -2 N_i \log \lambda_i + 2 \lambda_i + 2 \log(N_i!)
\]

\[
\approx \sum_{i=1}^{n} [2(\lambda_i - N_i + N_i \log(N_i/\lambda_i)) + \log N_i] + n \log(2\pi),
\]

and \(a(x) = \sum_{i=1}^{n} \log N_i + n \log(2\pi)\). Then, looking at the definition of \(T_{H_0}\) for instance, we have

\[
T_{H_0}(x, \eta) = \sum_{i=1}^{n} 2(\mu_i(\eta) - N_i + N_i \log(N_i/\mu_i(\eta))).
\]

In practice, when there are prior experiments carried out to study the nuisance parameter, an additional term that reflects deviation from this prior knowledge is added to the definition of \(T_{H_0}(x, \eta)\). We denote this term by
The green line represents \( \Delta T \), and the alternative hypothesis \( H_1 \) can be defined similarly for the alternative hypothesis \( H_0 \). Then the test statistic
\[
\Delta T(x) = T_{H_1}(x) - T_{H_0}(x)
\] (9)
is constructed in order to evaluate whether the observed data \( x \) favors \( H_0 \) or \( H_1 \). It is easy to see that a positive \( \Delta T(x) \) would favor \( H_0 \), and a negative \( \Delta T(x) \) would favor \( H_1 \). In addition, the absolute size of \( \Delta T(x) \) reflects how much one hypothesis is favored relative to the other. (An alternative way to define \( \Delta T(x) \) is to replace \( T_{H_0}(x) \) and \( T_{H_1}(x) \) by \( T_{H_0}^{\text{mag}}(x) \) and \( T_{H_1}^{\text{mag}}(x) \), respectively, which are the marginalized, or say integrated version of \( T_{H_0}(x, \eta) \) and \( T_{H_1}(x, \eta) \) over all nuisance parameters.)

Fig. 2 is a heuristic illustration of the distribution of the log-likelihood ratio \( \Delta T(X) \), where \( X \) stands for data generated from a potential repeat of the experiment. The black (red) curve stands for the density function of the distribution under the assumption that the null (alternative) hypothesis is true. The green line represents \( \Delta T(x) \), the log-likelihood ratio calculated from the observed data \( x \). The CL\( _s \) value is then defined as:
\[
\text{CL}_s(x) = \frac{1 - p_1}{1 - p_0}, \quad (10)
\]
where \( p_0 \) is the probability that a potential repeat of the experiment will yield a \( \Delta T(X) \) value smaller than (to the right of) \( \Delta T(x) \), had the null hypothesis been true, smaller order are sometimes omitted from the definition of \( T_{H_0}(x, \eta) \). After all, there are at least four common variations for \( T_{H_0}(x, \eta) \) used in practice:

\[
T_{H_0}(x, \eta) = \sum_{i=1}^{n} 2 \left( \frac{\mu_i - N_i}{N_i} \right) + \chi^2_{\text{penalty}}(\eta),
\] (5)

\[
T_{H_0}(x, \eta) = \sum_{i=1}^{n} \log \frac{\mu_i}{N_i} + \sum_{i=1}^{n} \frac{(N_i - \mu_i)^2}{\mu_i} + \chi^2_{\text{penalty}}(\eta),
\] (6)

\[
T_{H_0}(x, \eta) = \sum_{i=1}^{n} (N_i - \mu_i)^2 \mu_i + \chi^2_{\text{penalty}}(\eta),
\] (7)

\[
T_{H_0}(x, \eta) = \sum_{i=1}^{n} \frac{(N_i - \mu_i)^2}{N_i} + \chi^2_{\text{penalty}}(\eta).
\] (8)

Here, Eq. (5) and Eq. (6) origins from the likelihood function of the Poisson and the Gaussian distribution, respectively. Eq. (7) and Eq. (8) are variations of Eq. (6), and are commonly referred to as the Pearson and the Neyman Chi-square, respectively. Note that Eq. (1) is a specific example of Eq. (7). At large data limit, differences among the numerical values of \( T_{H_0}(x, \eta) \) based on these four definitions are negligible. En route to form the test statistic \( \Delta T \), \( T_{H_0}(x, \eta) \) is further minimized over all nuisance parameters (including unknown parameters and systematic uncertainties) to obtain \( T_{H_0}^{\text{min}}(x) = \min_{\eta} T_{H_0}(x, \eta) \). Also, \( T_{H_1}^{\text{min}}(x, \zeta) \) can be defined similarly for the alternative hypothesis \( H_1 \). Then the test statistic
\[
\Delta T(x) = T_{H_1}^{\text{min}}(x) - T_{H_0}^{\text{min}}(x)
\] (9)
is constructed in order to evaluate whether the observed data \( x \) favors \( H_0 \) or \( H_1 \). It is easy to see that a positive \( \Delta T(x) \) would favor \( H_0 \), and a negative \( \Delta T(x) \) would favor \( H_1 \). In addition, the absolute size of \( \Delta T(x) \) reflects how much one hypothesis is favored relative to the other. (An alternative way to define \( \Delta T(x) \) is to replace \( T_{H_0}(x) \) and \( T_{H_1}(x) \) by \( T_{H_0}^{\text{mag}}(x) \) and \( T_{H_1}^{\text{mag}}(x) \), respectively, which are the marginalized, or say integrated version of \( T_{H_0}(x, \eta) \) and \( T_{H_1}(x, \eta) \) over all nuisance parameters.)

Similarly, \( 1 - p_1 \) is the probability that a potential repeat of the experiment leads to a \( \Delta T(X) \) value larger than (to the left of) \( \Delta T(x) \), had the alternative hypothesis been true. Given the definition of CL\( _s \) in Eq. (10), one can see a CL\( _s \) value close to zero would strongly favor the null hypothesis \( H_0 \), while a CL\( _s \) value close to one would generally favor the alternative hypothesis \( H_1 \). \(^2\) In such experiments, the null hypothesis is traditionally specified to be the Standard Model. And the standard exclusion region of the parameter space is usually defined as the set of parameter values of new physics that corresponds to CL\( _s \) smaller than 0.05 [15]. Of course, other threshold values of the CL\( _s \) can be used to set exclusion regions.

The CL\( _s \) value defined in Eq. (10) is essentially a ratio of two p-values. Compared with the popular (single-sided) p-value, CL\( _s \) has clear advantages in situations

\(^2\) In the case where the null hypothesis \( H_0 \) and the alternative hypothesis \( H_1 \) are very similar (as shown in Fig. 3), a close to one CL\( _s \) value does not indicate \( H_1 \) is favored over \( H_0 \).
where the null hypothesis $H_0$ and the alternative hypothesis $H_1$ are very similar (see Fig. 3). For example, assume the data $x$ is an “extreme” measurement with respect to the alternative hypothesis $H_1$ (e.g., small $p_1$), it will also be disfavored by the null hypothesis $H_0$ (e.g., small $p_0$). If only a single p-value, either $p_0$ or $p_1$, is examined as in the CI approach, then one would draw the inappropriate conclusion of excluding $H_0$ or $H_1$ and favor the other hypothesis. However, in reality, since the hypotheses $H_0$ and $H_1$ are similar, the data does not really carry enough power to differentiate them. The CL$_s$ value that is the ratio of $1 - p_1$ and $1 - p_0$ will protect against this kind of situations.

Given the observed $\Delta T(x)$, to obtain the value of $p_0$ and $p_1$ required to calculate the CL$_s$, one needs to find the parent distribution of $\Delta T(X)$ under the null and the alternative hypothesis, respectively. While Monte Carlo simulations will always provide a valid approximation of the parent distributions of $\Delta T(X)$, simpler approximation methods are desired to lower the burden of computation. We show in the Sec. III C that, under the following set of conditions,

1. the parameter space $(H)$ of the nuisance parameters $\eta$ and $\zeta$ are both continuous and the model likelihood function is a smooth function (for example three times differentiable) in the parameters,

2. the data size $N_i$ is large for each $i = 1, \ldots, n$,

3. the best model under the null hypothesis $H_0$ and the alternative hypothesis $H_1$ are relatively close, in the sense that $|\mu^0 - \nu^0| << \mu^0 \sim \nu^0$, for $i = 1, \ldots, n$,

a simple approximation for the parent distribution of $\Delta T(X)$ under $H_j$, for either $j = 0$ or $1$, is the Gaussian distribution with mean $\Delta T_{H_j}$ and standard deviations $2\sqrt{\Delta T_{H_j}}$. We emphasize that, compared to the first regularity condition required by the Wilks’ theorem, our first regularity condition is much easier to achieve. Because we require continuity only for the parameter space $H$ instead of for the parameter space $\Theta \times M \times H$. Based on the Gaussian approximation, the CL$_s$ value is easily calculated with $\Delta T(x)$, $\Delta T_{H_0}$, and $\Delta T_{H_1}$. In case any one of the above conditions breaks down, the parent distribution of $\Delta T(X)$ is not necessarily well approximated by the Gaussian distribution, and should instead be estimated through Monte Carlo simulations. We next demonstrate in detail the procedure of using the CL$_s$ approach based on the Gaussian approximation to set exclusion sets using a specific example.

B. Setting Exclusion Sets with the Gaussian CL$_s$ Method

In this section, we illustrate the procedure of setting exclusion sets with the Gaussian CL$_s$ method for the neutrino oscillation example from Sec. II. The null hypothesis is specified to be $H_0$: $(\sin^2 2\theta, |\Delta m^2|, \eta) \in \{(0, |\Delta m^2_0|)\} \times H$ (e.g. the Standard Model with 3 neutrinos). Let $x^\text{Asimov}_{H_0}$ denote the Asimov data set under the null hypothesis $H_0$, which is exactly the expected mean counts corresponding to $H_0$ without any fluctuations of statistics nor variations of systematics. Ideally, when $H_0$ is correct, $x^\text{Asimov}_{H_0}$ would be defined as $\mu(\eta^0)$, where $\eta^0$ stands for the true value of the nuisance parameter. But in practice, we do not know $\eta^0$, so it is either approximated by the priorly known nominal value of the nuisance parameter (such as the term $\eta_0$ in Eq. (1)), or when the data size is large, by the minimizer of $T_{H_0}(x; \eta)$, denoted by $\hat{\eta}$. Then for any $(\sin^2 2\theta_1, |\Delta m^2_1|)$ from the parameter space $\Theta \times M$, one point at a time, we define $H_1$: $(\sin^2 2\theta_1, |\Delta m^2_1|, \eta) \in \{(\sin^2 2\theta_1, |\Delta m^2_1|)\} \times H$, and repeat the following procedure:

1. For the observed data $x$, calculate

$$\Delta T(x) := T_{H_1}(x) - T_{H_0}(x).$$

2. For the Asimov data $x^\text{Asimov}_{H_0}$, calculate

$$\Delta T_{H_0} := \Delta T(x^\text{Asimov}_{H_0}) = T_{H_1}(x^\text{Asimov}_{H_0}) - T_{H_0}(x^\text{Asimov}_{H_0}) = T_{H_0}(x^\text{Asimov}_{H_0}),$$

where the last step holds because $T_{H_0}(x^\text{Asimov}_{H_0}) = 0$ by the definition of $T_{H_0}$ and $x^\text{Asimov}_{H_0}$. Then according to the main result that we prove in Sec. III C, under the null hypothesis $H_0$, $\Delta T(X)$ follows approximately a Gaussian distribution with mean $\Delta T_{H_0}$ and standard deviation $2\sqrt{\Delta T_{H_0}}$. This suggests that one calculates

$$1 - p_0 = \frac{1 + \text{Erf}\left(\frac{\Delta T_{H_0} - \Delta T(x)}{\sqrt{8|\Delta T_{H_0}|}}\right)}{2},$$

where $\text{Erf}(s) = \frac{2}{\sqrt{\pi}} \int_0^s e^{-t^2} dt$ denotes the Gaussian error function for any $s \in (-\infty, \infty)$.

3. Form the Asimov data set $x^\text{Asimov}_{H_1}$, similar to how we formed $x^\text{Asimov}_{H_0}$. Calculate

$$\Delta T_{H_1} = \Delta T(x^\text{Asimov}_{H_1}) = T_{H_1}(x^\text{Asimov}_{H_1}) - T_{H_0}(x^\text{Asimov}_{H_1}) = -T_{H_0}(x^\text{Asimov}_{H_1}).$$

Then obtain

$$1 - p_1 = \frac{1 + \text{Erf}\left(\frac{\Delta T_{H_0} - \Delta T(x)}{\sqrt{8|\Delta T_{H_1}|}}\right)}{2}. \quad (14)$$


In terms of computing effort, the above CLs different definitions of the test statistic $\Delta T$ in the end of Sec. III C 2, we show that the difference for the parent distribution of $\Theta$ will be devoted to developing a Gaussian approximation namely the Pearson Chi-square statistic. For clarity, we versions of $\Delta T$ derived for each phase space point in $\Theta$ section, we focus on studying $\Delta T$ vals with $\Delta T$ standard Chi-square method of setting confidence inter-

cost of the Gaussian CLs method. In comparison, the standard Chi-square method of setting confidence intervals with $\Delta \chi_{\text{min}}(x)$ (Eq. (2)) requires the calculation of $\Delta \chi_{\text{min}}(x)$ at each phase space point. So the computing cost of the Gaussian CLs method is at the same order (about three times) as that of the standard approach. This is typically considered a mild increment in computing effort in practice, and is worthy given the benefits of using the CLs method mentioned in Sec. III A.

C. A Gaussian Approximation to the Distribution of $\Delta T(X)$ with Large Data Size

A key component in the aforementioned CLs method is a Gaussian approximation to the parent distribution of $\Delta T(X)$. We prove the validity of such an approximation in this section. Specifically, we show that by omitting terms of relatively small orders, the distribution of $\Delta T(X)$ under the null hypothesis follows approximately the Gaussian distribution with mean $\Delta T$ and standard deviation $2\sqrt{\Delta T}$.

1. Description of the mathematical problem and notations

Recall that we defined four versions of $(T_{H_0}(X; \eta), T_{H_1}(X; \zeta))$ that correspond to four different definitions of the test statistic $\Delta T(X)$. In this section, we focus on studying $\Delta T(X)$ based on Eq. (7), namely the Pearson Chi-square statistic. For clarity, we call it $D(X)$ from here on. The main part of Sec. III C 2 will be devoted to developing a Gaussian approximation for the parent distribution of $D(X)$. And in the remarks in the end of Sec. III C 2, we show that the difference between the other three versions of $\Delta T(X)$ to $D(X)$ are generally negligible, so that the approximate distribution derived for $D(X)$ can also be used for all the different versions of $\Delta T(X)$.

We now clearly state the mathematical problem con-
cerning $D(X)$. Let

$$\chi^2_{H_0}(X, \eta) := \sum_{i=1}^{n} \frac{(N_i - \mu_i(\eta))^2}{\mu_i(\eta)}$$

and

$$\chi^2_{H_1}(X, \zeta) := \sum_{i=1}^{n} \frac{(N_i - \nu_i(\zeta))^2}{\nu_i(\zeta)}$$

and let $\hat{\eta} = \arg \min_\eta \chi^2_{H_0}(X, \eta)$ and $\hat{\zeta} = \arg \min_\zeta \chi^2_{H_1}(X, \zeta)$. Then the definition of $D(X)$ is

$$D(X) = \chi^2_{H_1}(X, \hat{\zeta}) - \chi^2_{H_0}(X, \hat{\eta}).$$

Note our goal is to obtain an approximation of the distribution of $D(X)$ when the data size is large. Hence a specific quantity, say $m$, is needed that reflects the magnitude of the data, in order that we can describe how other quantities in the model change along with it. For example, $m$ could be the duration of the experiment or the total number of events. In the example of the search for a sterile neutrino in Sec. IV, $m$ is the expected number of neutrino events seen by the detector without any oscillation per bin. It is also proportional to the total number of events and total duration of the experiment. For the ease of description, let $m$ represent the duration of the experiment in this section. Then $p = \frac{X}{m}$ stands for the per unit time observed counts in a potential experiment, and it would remain stable (instead of tending to infinity or zero) as $m$ grows. So we say $p$ is of order $O_p(1)$ (with respect to $m$) 3.

In order to describe the modeling of counts rigorously, we introduce a set of notations, a summary of which is provided in Table I. Recall we employed $\mu(\eta)$ to denote the vectors of the expected counts under the correct hypothesis of the model, where $\eta$ is the vector of unknown nuisance parameters of dimension $q$. Denote the true value of $\eta$ by $\eta^0$, that is, $\mu^0 = \mu(\eta^0)$ is the true expected counts of the observation such that

$$N_i^{\text{indep}} \sim \text{Poisson}(\mu_i^0) \quad \text{for} \quad i = 1, \cdots, n.$$

When the data size is large, a very good approximation to the model above is given by

$$N_i^{\text{indep}} \approx N(\mu_i^0, \mu_i^0) \quad \text{for} \quad i = 1, \cdots, n.$$

3 The Big $O$, the small $o$, the Big $O_p$, and the small $o_p$ notation are standard mathematical symbols, such that for two sequences of random variables $\{X_m\}$ and $\{Y_m\}$, we write

- $Y_m = O_p(X_m)$ if and only if $Y_m/X_m \to 0$ in probability as $m \to \infty$, and
- $Y_m = O_p(X_m)$ if and only if $Y_m/X_m$ is bounded in probability as $m \to \infty$.

In the special case where $\{X_m\}$ and $\{Y_m\}$ are deterministic sequences, the stochastic $o_p$ and $O_p$ symbols reduce to the $o$ and $O$ symbols. See [19, sec 2.2] for details on the rules of calculus with these symbols.
Further, let $\pi := \mu/m$ denote the per unit time expected counts. To help explain these notations, take the example from Sec. III B for instance, if $H_0 : (\sin^2 2\theta, |\Delta m^2|) = (\sin^2 2\theta_0, |\Delta m^2_0|)$ is the correct hypothesis, then $q = \dim(y) = 3$, and $\pi_i = \mu_i(y)/m = a_i(y) \cdot P_i + b_i(y)$. The terms $a_i$ and $b_i$ are functions of order $O(1)$, and are determined by the configuration of the experiment. For example, $a_i$ can represent the detector efficiency, neutrino flux from reactor, target mass, etc., $b_i$ can represent the backgrounds. Also, $P_i = 1 - \sin^2 2\theta_0 \sin^2 (1.27 \cdot |\Delta m^2_0| \cdot L/E_i)$ represents the (disappearance) oscillation probability.

In addition, a competing framework, namely the collection of models that satisfy $H_1$, specifies the expected counts incorrectly as $\nu(\zeta)$, where $\zeta$ is the unknown nuisance parameter of dimension $q^*$. Also, define the per unit time expected counts under $H_1$ by $p = \nu/m$. Under $H_1$, there exists a unique $\zeta^0$, such that $\hat{\zeta}$ approaches $\zeta^0$ as $m \to \infty$. We will show in Appendix A that $\zeta^0$ has the interpretation that it corresponds to the model $\nu(\zeta)$ among all that belong to the alternative framework that is the closest to the true model $\mu^0$ in terms of the deviation $\sum^n_{i=1} (\mu^0_i - \nu(\zeta_i))^2$. Denote $\nu^0 = \nu(\zeta^0)$.

2. Approximating the distribution of the test statistic $D(X)$

In this section, we study the distribution of $D(X)$ defined in Eq. (17). For convenience, we will suppress the dependence on $X$ in the notation, and write $D = \chi^2_{H_1}(\hat{\zeta}) - \chi^2_{H_0}(\hat{\eta})$. On one hand, it's well known that the distribution of $\chi^2_{H_0}(\hat{\eta})$ approaches the Chi-square distribution with degree of freedom $(n - q)$ as $m$ increases. On the other hand, the limiting distribution of $\chi^2_{H_1}(\hat{\zeta})$ as $m$ increases does not always exist. Specifically, the behavior of $\chi^2_{H_1}(\hat{\zeta})$ for large $m$ is dependent on how far apart the expected counts of the best model under the alternative theoretical frameworks are from that of the true model. Denote the difference of per unit expected counts between the two models by $\delta = \pi^0 - \pi^0$. For all further investigation of $\chi^2_{H_1}(\hat{\zeta})$, we make the following realistic assumption on $\delta$:

\[\text{[A0]} \quad \delta = \pi^0 - \pi^0 \text{ is a constant vector that does not change with } m.\]

We write

$$\chi^2_{H_1}(\zeta) = \sum^n_{i=1} (N_i - \hat{\mu}_i)^2 = : \sum f_i^2$$ and $$\chi^2_{H_0}(\eta) = \sum^n_{i=1} (N_i - \hat{\mu}_i)^2 = : \sum e_i^2.$$

Here

$$f_i = f_i(a, b, c) = \frac{N_i - \hat{\mu}_i}{\hat{\mu}_i^2} = \sqrt{m} \frac{p_i - \pi_i}{\pi_i^2} = \sqrt{m} \frac{(p_i - \pi^0_i) - (\hat{\pi}_i - \pi^0_i)}{((\hat{\pi}_i - \pi^0_i) + \pi^0_i)^{1/2}} = \sqrt{m} \frac{a_i - c_i + \delta_i}{(c_i + \pi^0_i)^{1/2}},$$

and

$$e_i = e_i(a, b, c) = \frac{N_i - \hat{\mu}_i}{\hat{\mu}_i^2} = \sqrt{m} \frac{p_i - \pi_i}{\pi_i^2} = \sqrt{m} \frac{(p_i - \pi^0_i) - (\hat{\pi}_i - \pi^0_i)}{((\hat{\pi}_i - \pi^0_i) + \pi^0_i)^{1/2}} = \sqrt{m} \frac{a_i - b_i}{(b_i + \pi^0_i)^{1/2}},$$

where $a_i = p_i - \pi^0_i = O_p(m^{-\frac{1}{2}})$, $b_i = \hat{\pi}_i - \pi^0_i$, and $c_i = \hat{\pi}_i - \pi^0_i = O_p(m^{-\frac{1}{2}})$, for $i = 1, \cdots, n$. Then by the Taylor expansion of $f = (f_1, \cdots, f_n)^T$ and $e = (e_1, \cdots, e_n)^T$ around $(a, b, c) = (0, 0, 0)$, we have

$$f = \sqrt{m} \text{diag}(\pi^0)^{-\frac{1}{2}} \delta + \text{diag}(\pi^0)^{-\frac{1}{2}} \left[ \text{diag}(\pi^0) - \frac{1}{2} \text{diag}(\pi^0 + \pi^0) E^* \right] \sqrt{m}(p - \pi^0) + O_p(m^{-\frac{1}{2}}),$$

where the three terms in the above expression are of order $O_p(m^{\frac{1}{2}})$, $O_p(1)$ and $O(m^{-\frac{1}{2}})$ respectively. Further,

$$e = \text{diag}(\pi^0)^{-\frac{1}{2}} (I - D) \sqrt{m}(p - \pi^0) + O_p(m^{-\frac{1}{2}}),$$

where $D = B(B^T \text{diag}(\pi^0)^{-1} B)^{-1} B^T \text{diag}(\pi^0)^{-1}$, and the two terms in the above expression are of order $O_p(1)$ and $O_p(m^{-\frac{1}{2}})$ respectively. Therefore

$$D = \chi^2_{H_1}(\eta) - \chi^2_{H_0}(\hat{\zeta}) = f^T f - e^T e$$

$$= m \delta^T \text{diag}(\pi^0)^{-1} \delta + 2 \sqrt{m} \delta^T \text{diag}(\pi^0)^{-1} \left[ \text{diag}(\pi^0) - \frac{1}{2} \text{diag}(\pi^0 + \pi^0) E^* \right] \sqrt{m}(p - \pi^0) + O_p(1)$$

$$= m \delta^T \text{diag}(\pi^0)^{-1} \delta + 2 \sqrt{m} \delta^T \text{diag}(\pi^0)^{-1} \sqrt{m}(p - \pi^0)$$

$$- \frac{1}{2} \sqrt{m} \left[ \delta^T \text{diag} \left\{ \frac{\pi^0 + \pi^0\langle \pi^0 \rangle^2}{\langle \pi^0 \rangle^3} \right\} B^* \right] \left[ B^T \text{diag} \left\{ \frac{\pi^0}{\langle \pi^0 \rangle^2} \right\} B^* \right]^{-1} B^T \text{diag} \left\{ \frac{\pi^0}{\langle \pi^0 \rangle^2} \right\} \sqrt{m}(p - \pi^0) + O_p(1).$$
Under the alternative model

| General notation | Under the correct model | Under the alternative model |
|------------------|-------------------------|-----------------------------|
| Mean bin counts  | \( \mu(\eta) = (\mu_1(\eta), \ldots, \mu_N(\eta)) \) | \( \nu(\zeta) = (\mu_1(\zeta), \ldots, \mu_N(\zeta)) \) |
| Per-unit mean counts | \( \pi(\eta) = \mu(\eta)/m \) | \( \tau(\zeta) = \nu(\zeta)/m \) |

True values or their closest approximations under the give model

| nuisance parameter | \( \eta_0 \) (a \( q \)-dim vector) | \( \zeta_0 \) (a \( q \)-dim vector) |
|--------------------|-----------------------------------|-----------------------------------|
| Mean bin counts    | \( \mu_0 = \mu(\eta_0) \) | \( \nu_0 = \nu(\zeta_0) \) |
| Per-unit mean counts | \( \pi_0 = \mu_0/m \) | \( \tau_0 = \nu_0/m \) |

Estimation based on observed data

| nuisance parameter | \( \hat{\eta} = \arg \min \chi^2_{\hat{\eta}_0}(X, \eta) \) | \( \hat{\zeta} = \arg \min \chi^2_{\hat{\eta}_0}(X, \zeta) \) |
|--------------------|-------------------------------------------------|-------------------------------------------------|
| Mean bin counts    | \( \hat{\mu} = \mu(\hat{\eta}) \) | \( \hat{\nu} = \nu(\hat{\zeta}) \) |
| Per-unit mean counts | \( \hat{\pi} = \hat{\mu}/m \) | \( \hat{\tau} = \hat{\nu}/m \) |

| TABLE I. Legend of symbols used in describing the correct model and the alternative model, respectively. |

According to Eq. (27) of Lemma 1, the term in the closed bracket above reduces to 0. Hence

\[
D = m \delta^T \text{diag}\{\tau^0\}^{-1} \delta + 2 \sqrt{m} \delta^T \text{diag}\{\tau^0\}^{-1} \sqrt{m}(p - \pi^0) + O_p(1).
\]

Denote the first term of \( D \) by

\[
D_1 = m (\pi^0 - \tau^0)^T \text{diag}\{\tau^0\}^{-1} (\pi^0 - \tau^0) = \sum_{i=1}^n \frac{(\mu_i^0 - \nu_i^0)^2}{\nu_i^0} = \min_{\nu} \sum_{i=1}^n \frac{(\mu_i^0 - \nu_i)^2}{\nu_i} =: \overline{D},
\]

where the second to last equality follows from Appendix A. Note that under assumption [A0], \( D_1 = \overline{D} \) is of order \( O(m) \). Next denote the second term of \( D \) by \( D_2 \). The central limit theorem implies that as \( m \) increases to infinity, \( \sqrt{m}(p - \pi^0) \) converges in distribution to the \( N(0, \text{diag}\{\pi^0\}) \) distribution. Hence \( D_2/(2\sqrt{m}) \) converges in distribution to the \( N(\mu_2, V_2) \) where

\[
\mu_2 = \delta^T \text{diag}\{\pi^0\}^{-1} 0 = 0,
\]

and

\[
V_2 = \delta^T \text{diag}\{\pi^0\}^{-1} \text{diag}\{\tau^0\} \text{diag}\{\pi^0\}^{-1} \delta
= \delta^T \text{diag}\{\pi^0\}^{-1} \delta + \delta^T \text{diag}\{\pi^0 - \tau^0\}(\pi^0 - \tau^0)\delta
= \overline{D} + \sum_{i=1}^n \frac{(\pi_i^0 - \tau_i^0)^3}{(\pi_i^0)^2} =: \overline{D} + s.
\]

Note that both \( \overline{D} \) and \( s \) are of order \( O(1) \) under assumption [A0], hence \( D_2 = O_p(m^{\star}) \). In summary, under assumption [A0], we have \( D = D_1 + D_2 + O_p(1) \), where

\[
D_1 + D_2 \approx N(\overline{D}, 4\overline{D} + 4ms).
\]

Remarks and Implications of Eq. (19)

1. Under certain conditions, further simplification can be made to the approximating distribution, \( N(\overline{D}, 4\overline{D} + 4ms) \), that we derived for \( D(X) \). For example, one of the common situations is when the null and the alternative hypothesis are relatively close to each other, that is, when \( |\mu_0^0 - \nu_0^0| \ll \mu^0_0 \sim \nu^0_0 \). In such situations, one can safely ignore the \( ms \) term, because \( ms = \sum_i \frac{(\mu_i^0 - \nu_i^0)^2}{\nu_i^0} \ll \)
\[
\sum_i \frac{(\mu_i - N_i)^2}{\mu_i} = \overline{\mathcal{D}}. \text{ Then our main result becomes}
\]
\[
D \approx N(\overline{\mathcal{D}}, 4\overline{\mathcal{D}}). \tag{20}
\]

2. We claimed in Sec. III A that, at large data limit, the three versions of \(\Delta T(X) = T_{H_1}^{\text{min}}(X) - T_{H_0}(X)\) based on the definition of \(T_{H_0}\) (and the corresponding \(T_{H_1}\)) in Eq. (5), (6), and (8), each have negligible difference from the \(\Delta T(X)\) based on Eq. (7). And recall the last version of \(\Delta T(X)\) is the Pearson Chi-square statistic that we named \(D(X)\). Now, we validate this claim.

For the moment, we drop the penalty term \(\chi^2_{\text{penalty}}(\eta)\) from Eq. (5)–(8) for simplicity. And we will address the issue of the penalty term in the next remark.

First, for Eq. (5), we have,
\[
T_{H_0}(X) = \sum_i 2(\mu_i - N_i + N_i \log \frac{N_i}{\mu_i}),
\]
\[
= \sum_i 2(\mu_i - N_i + N_i \log (1 - \frac{\mu_i - N_i}{\mu_i})).
\]

At large data limit, \(|N_i - \mu_i| = O_p(\mu_i^{\frac{1}{2}}) = O_p(m^{\frac{1}{2}})\), we have by Taylor expansion,
\[
\log(1 - \frac{\mu_i - N_i}{\mu_i}) = \frac{\mu_i - N_i}{\mu_i} - \frac{(\mu_i - N_i)^2}{2\mu_i^2} + O_p\left(\frac{(\mu_i - N_i)^3}{\mu_i^3}\right)
\]
\[
= \frac{\mu_i - N_i}{\mu_i} - \frac{(\mu_i - N_i)^2}{2\mu_i^2} + O_p\left(m^{-\frac{3}{2}}\right).
\]

Therefore,
\[
T_{H_0}(X) = \sum_i 2\left(\frac{\mu_i - N_i}{\mu_i}\right)^2 \left(\frac{1}{2} + \frac{\mu_i - N_i}{2\mu_i} + O_p(m^{-\frac{3}{2}})\right)
\]
\[
= \sum_i 2\left(\frac{\mu_i - N_i}{\mu_i}\right)^2 \left(\frac{1}{2} + O_p(m^{-\frac{3}{2}})\right)
\]
\[
= \sum_i \left(\frac{\mu_i - N_i}{\mu_i}\right)^2 + O_p(m^{-\frac{3}{2}}).
\]

Next, for Eq. (6), we have,
\[
T_{H_0}(X) = \sum_i \left(\frac{(\mu_i - N_i)^2}{\mu_i} + \log \left(1 + \frac{\mu_i - N_i}{N_i}\right)\right)
\]
\[
= \sum_i \left(\frac{\mu_i - N_i}{\mu_i}\right)^2 + O_p(m^{-\frac{3}{2}}).\]

Finally, for Eq. (8), we have,
\[
T_{H_0}(X) = \sum_i \frac{(\mu_i - N_i)^2}{N_i}
\]
\[
= \sum_i \frac{(\mu_i - N_i)^2}{\mu_i} \left(1 + \frac{\mu_i - N_i}{N_i}\right)
\]
\[
= (\mu_i - N_i)^2 + \sum_i \frac{(\mu_i - N_i)^3}{\mu_i N_i}
\]
\[
= \sum_i \left(\frac{\mu_i - N_i}{\mu_i}\right)^2 + O_p(m^{-\frac{3}{2}}).
\]

After all, the difference between each version of \(T_{H_0}(X)\) and \(\sum_i \frac{(\mu_i - N_i)^2}{\mu_i}\) are negligible, so will be the difference between each version of \(T_{H_1}(X)\) and \(\sum_i \left(\frac{\mu_i - N_i}{\mu_i}\right)^2\). It follows that the different versions of the test statistic \(\Delta T(X)\) will behave similarly as \(D(X)\). Finally, it is easy to see that our definition for \(\overline{\mathcal{D}}\) is equivalent to \(\overline{\mathcal{T}}\) based on Eq. (7). Hence all different versions of \(\overline{\mathcal{T}}\) would bear similar numerical values as \(\overline{\mathcal{D}}\). And our main result in Eq. (20) can be stated as
\[
\Delta T \approx N(\overline{\mathcal{T}}, 4|\overline{\mathcal{T}}|). \tag{21}
\]

3. We emphasize that Eq. (17) is a general form of \(T\) in Eq. (5), (6), (7), and (8). The penalty term in \(T\) represents the constraint of systematic uncertainties, and is commonly obtained by dedicated measurements. When one includes the dedicated measurements as part of Chi-square definition, one naturally recovers Eq. (17). Therefore, our proof in Sec. III C is also valid for test statistics with the format of \(T\) in Eq. (5), (6), (7), and (8).

IV. A TOY EXAMPLE: SEARCH FOR STERILE NEUTRINO

In this section, we use a toy model to illustrate the Gaussian CL\(_s\) method.

A. Model Description

In this toy model, there are two detectors. One detector locates at 300 kilo-meters from the neutrino source and is called the near detector. The other detector locates at 300 kilo-meters from the neutrino source and is called the far detector. As shown in Fig. 4, the neutrino energy \(E_\nu\) covers from 1 GeV to 9 GeV. We further divide the entire energy range into 20 bins. A flat (energy independent) neutrino spectrum is assumed. The expected number of neutrino events seen by the near (far) detector without any oscillation is 10 k (0.9 k) per bin. We
consider two types of oscillation measurements: a disappearance measurement with oscillation formula

\[ P_{\text{dis}} = 1 - \sin^2 2\theta \cdot \sin^2 \left( 1.27 \cdot \Delta m^2 \frac{L}{E_\nu} \right), \]  

(22)

and an appearance measurement with oscillation formula

\[ P_{\text{app}} = \sin^2 2\theta \cdot \sin^2 \left( 1.27 \cdot \Delta m^2 \frac{L}{E_\nu} \right), \]  

(23)

where \( \theta \) is the neutrino mixing angle, \( \Delta m^2 \) is the neutrino mass squared difference, and \( L \) is the neutrino travel distance.

We further include a background with linear dependence on \( E_\nu \). The number of background events starts from 130 per bin for the first bin to 73 per bin for the last (20th) bin. There are three nuisance parameters, \( \epsilon \), \( \eta_n \), and \( \eta_f \). The first one is associated with the detector efficiency and the neutrino flux, which is assumed to be accurate to 5%. This uncertainty is assumed to be correlated between the near and the far detectors. The second and the third nuisance parameters are associated with the background normalization factors for the near and the far detectors, respectively. The normalization uncertainty is assumed to be 2% and uncorrelated between the two detectors. Fig. 4 shows the expected neutrino spectra. For the disappearance measurement, we compare the no-oscillation spectrum (the null hypothesis \( H_0: \sin^2 2\theta = 0 \)) with an oscillation spectrum (an alternative hypothesis \( H_1: \sin^2 2\theta = 0.06 \) at \( \Delta m^2 = 2.5 \times 10^{-3} \) eV\(^2\)). For the appearance measurement, we compare the no-oscillation spectrum (the null hypothesis \( H_0: \sin^2 2\theta = 0 \)) with two oscillation spectra (two alternative hypotheses \( H_1: \sin^2 2\theta = 0.008 \) or \( \sin^2 2\theta = 0.03 \) at \( \Delta m^2 = 2.5 \times 10^{-3} \) eV\(^2\)). Given a toy Monte Carlo (MC) sample \( N_j^i \), we use the following test statistic based on the Poisson likelihood, in line of Eq. (5):

\[
T = \sum_{j=n,f} \sum_{i=1}^{20} 2 \left( \mu_j^i(\epsilon, \eta_j, \sin^2 2\theta, \Delta m^2) - N_j^i \right) \log \frac{N_j^i}{\mu_j^i(\epsilon, \eta_j, \sin^2 2\theta, \Delta m^2)} + \frac{\epsilon^2}{0.05^2} + \frac{\eta_n^2}{0.02^2} + \frac{\eta_f^2}{0.02^2}. \]  

(24)

Here, \( i \) represents the bin number and runs from 1 to 20. \( j \) labels the near or the far detector. \( \mu_j^i \) is the expected
number of events in \(i\)-th bin and \(j\)-th detector. It depends on the oscillation parameters: \(\sin^2 2\theta\) and \(\Delta m^2\), and the nuisance parameters: \(\epsilon\) for the detector efficiency and neutrino flux, \(\eta_n\) (\(\eta_f\)) for the near (far) detector background normalization factors.

### B. Feldman-Cousins vs. Wilks

For the toy example mentioned above, we provided a theoretical argument in Sec. II stating that conditions required by the Wilks’ Theorem are violated, hence the Chi-square method is unsuitable for setting CI for the parameter \(\sin^2 2\theta\). Instead, the computationally intensive Feldman-Cousin method is always valid and need to be used to set CI in this example. The purpose of this section is to demonstrate the practical difference between the two methods. Since the Chi-square method is especially problematic when one needs to decide if 0 or values close to 0 should be in a CI for \(\sin^2 2\theta\), we consider the problem of checking the value 0. We need the parent distribution of the test statistic in Eq. (25) under the hypothesis \(H_0\): \(\sin^2 2\theta = 0\). To implement the Feldman-Cousin method, we generate a large number of MC samples assuming that \(\sin^2 2\theta = 0\). The MC samples have statistical fluctuations according to Poisson distributions, and systematic variations through randomizing the three nuisance parameters according to normal distributions. Given each MC sample, we find \(T_{\text{min}}^\text{H_0}\) which are the minimum value of \(T\) from Eq. (24) in the 5-dimensional phase space of \((\sin^2 2\theta, \Delta m^2, \epsilon, \eta_n, \eta_f)\), and the minimum value of \(T\) under the restriction, \(\sin^2 2\theta_{\text{true}} = 0\). Then we form the test statistic

\[
\Delta \chi^2_{\text{min}} = T_{\text{H_0}} - T_{\text{min}}. \tag{25}
\]

Fig. 5 shows the distribution of \(\Delta \chi^2_{\text{min}}\) which clearly does not follow a Chi-square distribution with two degrees of freedom. That is, to correctly set confidence intervals with the test statistic \(\Delta \chi^2_{\text{min}}\), the Chi-square method can not be used, and one has to resort to the computationally intensive Feldman-Cousin method.

### C. Validity of the Gaussian Approximation in the Gaussian CL\(_s\) method

Having seen that there is no fast way to implement the CI approach, we turn to the CL\(_s\) approach. In this section, we study the performance of the Gaussian CL\(_s\) method that we proposed. Specifically, we check how closely does the CL\(_s\) test statistic \(\Delta T = T_{\text{H_0}}^{\text{min}} - T_{\text{H}}^{\text{min}}\) follow the normal distribution \(N(\Delta T, 4\Delta T)\). Here, \(T_{\text{min}}\) is the value of the test statistic \(T\) from Eq. (24) under the hypothesis \(H\), minimized over the nuisance parameters (\(\epsilon, \eta_n, \eta_f\)). Fig. 6 shows the distribution of \(\Delta T\) for the disappearance measurement. The null hypothesis \(H_0\) corresponds to \(\sin^2 2\theta = 0\). The alternative hypothesis \(H_1\) corresponds to \(\sin^2 2\theta = 0.06\) and \(\Delta m^2 = 2.5 \times 10^{-3}\) eV\(^2\). The histograms on the left (right) are made from the toy MC samples assuming \(H_1\) (\(H_0\)) is true. We also compare them with the expected normal distribution \(N(\Delta T, 4\Delta T)\) from \(\Delta T_{\text{H_0}}\) and \(\Delta T_{\text{H}}\) calculated from the Asimov data sets. Good agreements are observed.

Similarly, we also check the appearance measurements. In Fig. 7, the null hypothesis \(H_0\) corresponds to \(\sin^2 2\theta = 0\), and the alternative hypothesis \(H_1\) corresponds to \(\sin^2 2\theta, \Delta m^2 = (0.008, 2.5 \times 10^{-3}\text{eV}^2)\). In Fig. 8, \(H_0\) corresponds to \(\sin^2 2\theta = 0\), and \(H_1\) corresponds to \(\sin^2 2\theta, \Delta m^2 = (0.03, 2.5 \times 10^{-3}\text{eV}^2)\). The agreement between the MCs and expectations in Fig. 7 is slightly worse than that in Fig. 6, but is still reasonably good. However, the difference between the MCs and expectations in Fig. 8 becomes large. This is because the third regularity condition “when the prediction of two hypotheses (the null hypotheses \(H_0\) and the alternative hypothesis \(H_1\) are relatively close or \(|\mu_i - \nu_i| < \mu_i \sim \nu_i\)” is no longer met. In a disappearance measurement, this condition can easily satisfied. However, this may not be true in an appearance experiment as the expected number of signal events is zero when \(\sin^2 2\theta = 0\). This could be much smaller than the expected number of signal events with oscillations. When \(H_0\) and \(H_1\) are \(\sin^2 2\theta = 0\) and \(\sin^2 2\theta, \Delta m^2 = (0.008, 2.5 \times 10^{-3}\text{eV}^2)\), respectively, this condition is still reasonably well satisfied with the existence of backgrounds. But when \(H_0\) and \(H_1\) are \(\sin^2 2\theta = 0\) and \(\sin^2 2\theta, \Delta m^2 = (0.03, 2.5 \times 10^{-3}\text{eV}^2)\), respectively, this condition is much more severely violated. Note that in such situations where \(H_0\) and \(H_1\) are very different, the experimental data is most likely able to exclude one hypothesis easily, making it less interesting to carry out an accurate statistical test. Nevertheless, we emphasize it is crucial to check with MCs whether the Gaussian approximation is valid in practice.
plotted for toy MCs (disappearance) assuming the hypothesis $H_0$ or $H_1$ is true. Here, the null hypothesis $H_0$ corresponds to $\sin^2 2\theta = 0$. The alternative hypothesis $H_1$ corresponds to $\sin^2 2\theta = 0.06$ and $\Delta m^2 = 2.5 \times 10^{-3}$ eV$^2$.

FIG. 7. (color online) The test statistic $\Delta T = T_{H_1}^{\text{min}} - T_{H_0}^{\text{min}}$ is plotted for toy MCs (appearance) assuming the hypothesis $H_0$ or $H_1$ is true. Here, the null hypothesis $H_0$ corresponds to $\sin^2 2\theta = 0$. The alternative hypothesis $H_1$ corresponds to $\sin^2 2\theta = 0.008$ and $\Delta m^2 = 2.5 \times 10^{-3}$ eV$^2$.

FIG. 8. (color online) The test statistic $\Delta T = T_{H_1}^{\text{min}} - T_{H_0}^{\text{min}}$ is plotted for toy MCs (appearance) assuming the hypothesis $H_0$ or $H_1$ is true. Here, the null hypothesis $H_0$ corresponds to $\sin^2 2\theta = 0$. The alternative hypothesis $H_1$ corresponds to $\sin^2 2\theta = 0.03$ and $\Delta m^2 = 2.5 \times 10^{-3}$ eV$^2$.

D. CL$_s$ vs. p-value

In the following, we use the disappearance experiment to compare the CL$_s$ approach with its alternatives. With the new test statistic $\Delta T = T_{H_1}^{\text{min}} - T_{H_0}^{\text{min}}$, one can in principle still follow the method of using p-value to set confidence intervals assuming the null hypothesis $H_0$. Fig. 9 compares confidence intervals set with p-values against exclusion sets obtained with CL$_s$ values, both using the test statistic $\Delta T$. The true $\sin^2 2\theta$ is assumed to be zero. The sensitivity of the Gaussian CL$_s$ method is slightly worse than that of the p-value method, because the CL$_s$ value is by construction larger than the corresponding p-value. Despite the slightly worse sensitivity, the advantage of the CL$_s$ method becomes clear when analyzing data. As shown in Fig. 9, the region that correspond to $\Delta m^2 \sim 5.5 \times 10^{-2}$ eV$^2$ and $\sin^2 2\theta < 0.01$ (also $\Delta m^2 \sim 0.1$ eV$^2$ and $\sin^2 2\theta < 0.01$) is excluded from the 95% confidence interval according to the p-value method. This is inconsistent with intuition as the expected spectrum for small $\sin^2 2\theta$ values should be very similar to that of $\sin^2 2\theta = 0$, and we do not expect to exclude regions with $\sin^2 2\theta$ values. This phenomenon can be understood as the following. With the test statistic $\Delta T$, we compare two hypotheses each time. Therefore, even when the two hypotheses are very similar, the chance of excluding one hypothesis with the p-value method can still be large as illustrated in Fig. 3. As we explained in Sec. III A, the definition of the CL$_s$ value avoid this problem, giving it an advantage over the traditional p-value method when test statistic $\Delta T$ is used.

E. The Gaussian CL$_s$ method vs. Feldman-Cousins CI method

We further compare the Gaussian CL$_s$ method with the Feldman-Cousins’ CI method. Fig. 10 compares the sensitivity of the Gaussian CL$_s$ method vs. that of the Feldman-Cousins CI method. We assumed that the true value of $\sin^2 2\theta$ is 0. The sensitivity of the 95% exclusion set from the Gaussian CL$_s$ method is slightly better than that of the 95% confidence interval from the Feldman-Cousins’ method, and is actually close that of the 90% confidence interval from the Feldman-Cousins’ method for this setup. This can be understood, since the test statistic $\Delta T = T_{H_1}^{\text{min}} - T_{H_0}^{\text{min}}$ used in the Gaussian CL$_s$ method is designed to focus on the difference between the new physics hypotheses (the alternative hypothesis: $\sin^2 2\theta > 0$) with the standard model (the null hypothesis: $\sin^2 2\theta = 0$). Therefore, when the true value of $\sin^2 2\theta$ is 0, the power of excluding new physics hypotheses with the Gaussian CL$_s$ method is stronger than that of the Feldman-Cousins’ CI method. On the other hand, when the new physics is indeed true, the Feldman-Cousins’ CI method has a clear advantage in constraining the parameter space. This is shown in Fig. 11. The toy MC sample is generated with $\sin^2 2\theta_{\text{true}} = 0.1$ and

FIG. 6. (color online) The test statistic $\Delta T = T_{H_1}^{\text{min}} - T_{H_0}^{\text{min}}$ is plotted for toy MCs (disappearance) assuming the hypothesis $H_0$ or $H_1$ is true. Here, the null hypothesis $H_0$ corresponds to $\sin^2 2\theta = 0$. The alternative hypothesis $H_1$ corresponds to $\sin^2 2\theta = 0.06$ and $\Delta m^2 = 2.5 \times 10^{-3}$ eV$^2$.
\[ \Delta m^2_{\text{true}} = 2.5 \times 10^{-3} \text{ eV}^2 \] with statistical fluctuations and systematic variations. The 90\% confidence interval of the Feldman-Cousins method clearly picked up the region close to the true value. However, the 5\%-95\% CL_{s} interval is much broader. This is due to the choice of the test statistic \( \Delta T \) in the Gaussian CL_{s} method vs. \( \Delta \chi^2_{\text{min}} \) in the Feldman-Cousins’ CI method. On the other hand, the constraint of \( \sin^2 2\theta \) for \( \Delta m^2 \) values around \( 2.5 \times 10^{-3} \text{ eV}^2 \) with the Gaussian CL_{s} method is tighter than that of the Feldman-Cousins’ method. This can also be understood since the proposed test statistic \( \Delta T \) focuses on the difference between the new physics hypothesis and the standard model, while the test statistic \( \Delta \chi^2_{\text{min}} \) takes into account all the possibilities around the tested true \( (\sin^2 2\theta_{\text{true}}, \Delta m^2_{\text{true}}) \) point. At other \( \Delta m^2 \) values, the performance of the test statistic \( \Delta T \) is clearly worse than that of the test statistic \( \Delta \chi^2_{\text{min}} \). Therefore, we confirm the conclusion from Ref. [1]: “the CL_{s} technique for setting limits is appropriate for determining exclusion sets while the determination of confidence intervals advocated by Feldman and Cousins’ method is more appropriate for treating established signals”.

V. DISCUSSION

As shown in Sec. IV, the test statistic \( \Delta \chi^2_{\text{min}}(X) \) clearly does not follow the Chi-square distribution with two degrees of freedom (Fig. 5). In Sec. II, we explained that this is because certain regularity conditions of the Wilks’ theorem are not satisfied. Firstly, \( \sin^2 2\theta = 0 \) is at the boundary of the allowed physical phase space, thus the condition on the continuity of the parameter space does not hold. Secondly, at \( \sin^2 2\theta = 0 \), due to the large coverage of \( |\Delta m^2| \) shown in Fig. 1b, the data size is not large enough to ensure that \( \Delta \chi^2_{\text{min}} \) has a parabolic shape in the close neighborhood of its minimum \( (\sin^2 2\theta_{\text{true}}, |\Delta m^2|_{\text{true}}) \). In contrast, if we use the test statistic \( \Delta T \), then its distribution agrees well with the

FIG. 9. (color online) Comparison of the exclusion sets determined by the Gaussian CL_{s} method vs. the CI determined by the p-value method, both using the test statistic \( \Delta T = T_{H_{1}}^{\text{min}} - T_{H_{0}}^{\text{min}} \). The true value of \( \sin^2 2\theta \) is 0. For the p-value (CL_{s}) method, the right side of the red (black) line has a p-value (CL_{s} value) smaller than 0.05. The sensitivity curves are generated from a large number of Monte Carlo samples. At each \( \Delta m^2 \), 50\% (50\%) of toy MC samples will have a better (worse) exclusion limit than the sensitivity curve.

FIG. 10. (color online) Comparison of the sensitivity of the 95\% Gaussian CL_{s} method vs. that of the 95\% and the 90\% Feldman-Cousins’ CI method. The true value of \( \sin^2 2\theta \) is 0.

FIG. 11. (color online) Comparison of the 90\% confidence intervals from the Feldman-Cousins’ method vs. 5\%-95\% interval from the Gaussian CL_{s} method for a toy MC sample generated with \( \sin^2 2\theta_{\text{true}} = 0.1 \) and \( \Delta m^2_{\text{true}} = 2.5 \times 10^{-3} \text{ eV}^2 \).
Gaussian distribution with mean and standard deviation of $\Delta T$ and $2\sqrt{\Delta T}$, respectively (Fig. 6 and Fig. 7). This is not surprising given that all three regularity conditions required for the Gaussian approximation in Sec. III A are satisfied. The first condition requires continuity of the phase space for the nuisance parameters, under the null and the alternative hypotheses, respectively. This is easily satisfied since the requirement concerns the nuisance parameters only, not the parameters of interest, $(\sin^2 2\theta, |\Delta m^2|)$. The second condition concerns large enough data size, which can be reached relatively easily compared to the large data size required by the Wilks’ theorem. This is because in the Gaussian CL$_s$ method we are concerned with testing a simpler pair of hypotheses, where the value of $(\sin^2 2\theta, |\Delta m^2|)$ is fixed, and one automatically avoids the situation shown in Fig. 1b that requires performing minimization over a large range of $|\Delta m^2|$ in calculating the test statistic. The third condition requires that the fixed values of $(\sin^2 2\theta, |\Delta m^2|)$ assumed by the two hypotheses are relatively close. This condition can be violated when one tries to exclude certain values of $(\sin^2 2\theta, |\Delta m^2|)$, and the performance of the Gaussian approximation does deteriorate, as seen in Fig. 8. Therefore, we emphasize that one could always use the Monte Carlo simulation method to check validity of results based on the Gaussian CL$_s$. And in situations where conditions of the Gaussian CL$_s$ method are clearly violated, one should rely on the Monte Carlo simulation results.

As discussed in Ref. [1] and shown in Sec. IV, the CL$_s$ method is complementary to the confidence interval method: CL$_s$ is appropriate to set exclusion limit while the confidence interval is appropriate in treating established signals. While the exclusion limit set by CL$_s$ method is always valid, it did not directly address the question “do we see new physics or not”. To address this question, we recommend to report the p-value calculated by the test statistic $\Delta \chi^2_{\text{min}}$ assuming the standard model is true.

Similar to the Chi-square CI method with test statistic $\Delta \chi^2_{\text{min}}$ and predefined constants, the Gaussian CL$_s$ method also allows easy combination of multiple independent experimental results that probe the same phase space. The CL$_s$ value at each alternative hypothesis $H_1$ from experiments $(k = 1, 2, \ldots, m)$ can be calculated with

$$\Delta T(x) = \sum_{k=1}^{m} \Delta T(x_k),$$

$$\Delta T(x_{H_1}^{\text{Asimov}}) = \sum_{k=1}^{m} \Delta T(x_{H_1}^{\text{Asimov}}),$$

$$\Delta T(x_{H_0}^{\text{Asimov}}) = \sum_{k=1}^{m} \Delta T(x_{H_0}^{\text{Asimov}}),$$

with $x = \sum_{k=1}^{m}$ is the combined data from all experiments. This can be easily achieved if each experiment releases their maps of $T_{x_k}$, $\Delta T(x_{H_1}^{\text{Asimov}})$, and $\Delta T(x_{H_0}^{\text{Asimov}})$.

In practice, the main challenge in combining multiple experiment results arise from the potential correlation among different experiments and requires careful examination.

VI. SUMMARY

In this paper, we describe a new method to present results in searching for new physics in a continuous parameter space. This method is based on the CL$_s$ approach that is complementary to the traditional confidence interval approach. Many pairs of hypotheses need to be tested when setting exclusion sets with the CL$_s$ approach. For each two-hypotheses testing problem, we consider the log-likelihood ratio test statistic, and certain variations of it, denoted by $\Delta T$. We provide a rigorous mathematical proof that, the parent distribution of $\Delta T$ follows a Gaussian distribution at large data limit when the two hypotheses are relatively close. This approach can be used to calculate the CL$_s$ value and thus setting exclusion limits in one or multiple dimensional parameter space without intense Monte Carlo simulations. This approach can also be used to conveniently combine results from multiple experiments.

VII. ACKNOWLEDGMENTS

We would like to thank Wei Wang for helpful discussions. This material is based upon work supported by the National Science Foundation and the U.S. Department of Energy, Office of Science, Office of High Energy Physics, Early Career Research program under contract number DE-AC02-98CH10886.

APPENDICES

A. A few basic properties of the fitted models under $H_0$ and $H_1$

Suppose $H_0$ is the correct hypothesis, that is, the data $X$ came from $H_0$. Having observed the data, the best fitting models under $H_0$ and $H_1$ have estimated nuisance parameters $\eta$ and $\zeta$ respectively, as defined in Sec. III. The corresponding per unit expected counts are denoted $\tilde{\tau}$ and $\hat{\tau}$ respectively.

We show below that there is a unique limit of $\hat{\zeta}$ as the data size increases, and that it leads to the model $\nu(\zeta)$ that is the closest model under $H_1$ to the true model $\mu^0$ under a certain criteria. Indeed, let $t_m(\zeta) = \chi^2_{H_1}(\zeta)/m = \sum_i (m_i - \tau_i(\zeta))^2$, and let $t(\zeta) = \sum_i (\tau_i - \tau_i(\zeta))^2$. Since $p$ converges almost surely (a.s.) to $\pi^0$ as $m$ increases, we have $t_m(\zeta)$ converges a.s. to $t_m(\zeta)$. Then under regularity conditions, such as $t_m$ being twice differentiable and convex in $\zeta$, $\hat{\zeta} = \arg\min_m t_m(\zeta)$ also converges...
Note by delta’s method

Lemma 1. Assuming \( A0 \), we have

\[
\sqrt{m} \left( \frac{p - \pi^0}{\hat{\tau} - \tau^0} \right) = \left( \begin{array}{c} I \\ E^* \end{array} \right) \sqrt{m} \left( p - \pi^0 \right) + O_p(m^{-\frac{1}{2}}),
\]

where

\[
E^* = B^* \left( B^{*T} \text{diag} \left\{ \frac{\pi^0 \nu}{(\tau^0)^2} \right\} B^* \right)^{-1} B^{*T} \text{diag} \left\{ \frac{\pi^0}{(\tau^0)^2} \right\} \text{ and } B^*_{n \times q^*} = \frac{\partial \pi^0}{\partial \xi}.
\]

Further,

\[
B^{*T} \text{diag} \left\{ \frac{(\pi^0 + \tau^0)}{(\tau^0)^2} \right\} \delta = 0.
\]

Proof. By definition, \( \hat{\zeta} \) is such that

\[
\frac{\partial \chi^2_{H_1}(\hat{\zeta})}{\partial \xi_j} = 0 \text{ for } j = 1, \ldots, q^*.
\]

That is,

\[
2 \sum_{i=1}^n \frac{N_i + \nu_i}{\hat{\nu}_i^2} \frac{\partial \hat{\nu}_i}{\partial \xi_j} \left( N_i - \nu_i \right) = 0
\]

\[
\sum_{i=1}^n \frac{N_i + \nu_i}{\hat{\nu}_i^2} \frac{\partial \hat{\nu}_i}{\partial \xi_j} \left( N_i - \nu_i^0 \right) = \sum_{i=1}^n \frac{N_i + \nu_i}{\hat{\nu}_i^2} \frac{\partial \hat{\nu}_i}{\partial \xi_j} \left( \hat{\nu}_i - \nu_i^0 \right)
\]

\[
\sum_{i=1}^n \frac{p_i + \tau_i}{\hat{\tau}_i^2} \frac{\partial \hat{\tau}_i}{\partial \xi_j} \left( p_i - \pi_i^0 + \pi_i^0 - \tau_i^0 \right) = \sum_{i=1}^n \frac{p_i + \tau_i}{\hat{\tau}_i^2} \frac{\partial \hat{\tau}_i}{\partial \xi_j} \left( \hat{\tau}_i - \tau_i^0 \right).
\]

Note by delta’s method

\[
\hat{\tau}_i = \tau_i^0 + \sum_k \frac{\partial \tau_i^0}{\partial \xi_k} (\hat{\zeta}_k - \zeta_k^0) + O_p(m^{-1}) \text{ and } \frac{\partial \hat{\tau}_i}{\xi_j} = \frac{\partial \tau_i^0}{\xi_j} + \sum_k \frac{\partial^2 \tau_i^0}{\xi_j \partial \xi_k} (\hat{\zeta}_k - \zeta_k^0) + O_p(m^{-1}),
\]

and

\[
\frac{p_i + \hat{\tau}_i}{(\hat{\tau}_i)^2} - \frac{\pi_i^0 + \pi_i^0}{(\tau_i^0)^2} = \frac{1}{(\tau_i^0)^2} (p_i - \pi_i^0) - \frac{1}{(\tau_i^0)^2} (2 \hat{\tau}_i - \tau_i^0) + O_p(m^{-1}).
\]
Hence, the lhs of Eq. (28) becomes

\[
\text{lhs} = \sum_{i=1}^{n} \left( \frac{\pi_i^0 + \pi_i}{(\tau_i^0)^2} \right) \left( p_i - \pi_i^0 \right) - \sum_{i=1}^{n} \left( \frac{\pi_i^0 + \pi_i}{(\tau_i^0)^2} \right) \left( \frac{\pi_i^0 + \pi_i}{(\tau_i^0)^2} \right) \left( \pi_i^0 - \pi_i^0 \right) + O_p(m^{-1}) \left( \frac{\partial \pi_i^0}{\partial C_j} \right) \left( p_i - \pi_i^0 + \delta_i \right)
\]

\[
= \sum_{i=1}^{n} \left( \frac{\pi_i^0 + \pi_i}{(\tau_i^0)^2} \right) \frac{\partial \pi_i^0}{\partial C_j} \delta_i + \sum_{i=1}^{n} \left[ \frac{\tau_i^0 + \tau_i}{(\tau_i^0)^2} + \frac{\delta_i}{(\tau_i^0)^2} \right] \frac{\partial \pi_i^0}{\partial C_j} \left( p_i - \pi_i^0 \right) - \sum_{i=1}^{n} \frac{1}{(\tau_i^0)^3} (2\tau_i^0 + 1)(\pi_i^0 - \pi_i^0) \left( \frac{\partial \pi_i^0}{\partial C_j} \right) \delta_i + O_p(m^{-1})
\]

\[
= \sum_{i=1}^{n} \left( \frac{\pi_i^0 + \pi_i}{(\tau_i^0)^2} \right) \frac{\partial \pi_i^0}{\partial C_j} \delta_i + \sum_{i=1}^{n} \frac{2\pi_i^0 + \pi_i^0}{(\tau_i^0)^2} \frac{\partial \pi_i^0}{\partial C_j} \left( p_i - \pi_i^0 \right) - \sum_{i=1}^{n} \frac{2\pi_i^0 + \pi_i^0}{(\tau_i^0)^3} \delta_i \left( \frac{\partial \pi_i^0}{\partial C_j} \right) \left( \pi_i^0 - \pi_i^0 \right) + O_p(m^{-1})
\]

\[
= \sum_{i=1}^{n} \left( \frac{\pi_i^0 + \pi_i}{(\tau_i^0)^2} \right) \frac{\partial \pi_i^0}{\partial C_j} \delta_i + \sum_{i=1}^{n} \frac{2\pi_i^0 + \pi_i^0}{(\tau_i^0)^2} \frac{\partial \pi_i^0}{\partial C_j} \left( p_i - \pi_i^0 \right) - \sum_{i=1}^{n} \frac{2\pi_i^0 + \pi_i^0}{(\tau_i^0)^3} \delta_i \left( \frac{\partial \pi_i^0}{\partial C_j} \right) \left( \pi_i^0 - \pi_i^0 \right) + O_p(m^{-1})
\]

The rhs of Eq. (28) becomes

\[
\text{rhs} = \sum_{i=1}^{n} \left( \frac{\pi_i^0 + \pi_i}{(\tau_i^0)^2} \right) \frac{\partial \pi_i^0}{\partial C_j} \left( \frac{\pi_i^0}{\tau_i^0} \right)
\]

\[
= \sum_{i=1}^{n} \frac{p_i + \pi_i^0}{\tau_i^0} \frac{\partial \pi_i^0}{\partial C_j} \left( \frac{\pi_i^0}{\tau_i^0} \right)
\]

\[
= \sum_{i=1}^{n} \left( \frac{\pi_i^0 + \pi_i}{(\tau_i^0)^2} \right) \frac{\partial \pi_i^0}{\partial C_j} \delta_i + \sum_{i=1}^{n} \frac{2\pi_i^0 + \pi_i^0}{(\tau_i^0)^2} \frac{\partial \pi_i^0}{\partial C_j} \left( p_i - \pi_i^0 \right) - \sum_{i=1}^{n} \frac{2\pi_i^0 + \pi_i^0}{(\tau_i^0)^3} \delta_i \left( \frac{\partial \pi_i^0}{\partial C_j} \right) \left( \pi_i^0 - \pi_i^0 \right) + O_p(m^{-1})
\]

Hence, equating lhs and rhs leads to, for \( j = 1, \ldots, q^* \),

\[
\sum_{i=1}^{n} \left( \frac{\pi_i^0 + \pi_i}{(\tau_i^0)^2} \right) \frac{\partial \pi_i^0}{\partial C_j} \delta_i + \sum_{i=1}^{n} \frac{2\pi_i^0 + \pi_i^0}{(\tau_i^0)^2} \frac{\partial \pi_i^0}{\partial C_j} \left( p_i - \pi_i^0 \right) - \sum_{i=1}^{n} \frac{2\pi_i^0 + \pi_i^0}{(\tau_i^0)^3} \delta_i \left( \frac{\partial \pi_i^0}{\partial C_j} \right) \left( \pi_i^0 - \pi_i^0 \right) + O_p(m^{-1})
\]

\[
= \sum_{i=1}^{n} \left( \frac{\pi_i^0 + \pi_i}{(\tau_i^0)^2} \right) \frac{\partial \pi_i^0}{\partial C_j} \delta_i + \sum_{i=1}^{n} \frac{2\pi_i^0 + \pi_i^0}{(\tau_i^0)^2} \frac{\partial \pi_i^0}{\partial C_j} \left( p_i - \pi_i^0 \right) - \sum_{i=1}^{n} \frac{2\pi_i^0 + \pi_i^0}{(\tau_i^0)^3} \delta_i \left( \frac{\partial \pi_i^0}{\partial C_j} \right) \left( \pi_i^0 - \pi_i^0 \right) + O_p(m^{-1})
\]

Note that, under assumption (A0), all the terms are \( O_p(m^{-\frac{1}{2}}) \) or smaller except for the first term on the lhs. Letting \( m \) grow to infinity in Eq. (29) implies that

\[
\sum_{i=1}^{n} \left( \frac{\pi_i^0 + \pi_i}{(\tau_i^0)^2} \right) \frac{\partial \pi_i^0}{\partial C_j} \delta_i = 0 \quad \text{for all } j,
\]

which in matrix notation becomes

\[
B^T \text{diag} \left\{ \frac{\pi_0^0 + \pi_0^0}{\tau_0^0} \right\} \delta = 0,
\]

which proves Eq. (27) of Lemma 1. Plugging this result back into Eq. (29), we have

\[
0 + 2B^T \text{diag} \left\{ \frac{\pi_0^0}{(\tau_0^0)^2} \right\} (p - \pi_0) = B^T \text{diag} \left\{ \frac{\pi_0^0 + \pi_0^0}{(\tau_0^0)^2} \right\} \left( \frac{\pi_0^0}{(\tau_0^0)^3} \right) B^* (\hat{\zeta} - \zeta^0) + O_p(m^{-1}).
\]

That is

\[
2B^T \text{diag} \left\{ \frac{\pi_0^0}{(\tau_0^0)^2} \right\} (p - \pi_0) = 2B^T \text{diag} \left\{ \frac{(\pi_0^0)^2}{(\tau_0^0)^3} \right\} B^* (\hat{\zeta} - \zeta^0) + O_p(m^{-1}).
\]
Hence
\[
\sqrt{m}(\hat{\zeta} - \zeta^0) = (B^*T \text{diag} \left\{ \frac{(\pi^0)\bar{2}}{\tau^0} \right\} B^*)^{-1} B^*T \text{diag} \left\{ \frac{\pi^0}{(\tau^0)\bar{2}} \right\} \sqrt{m} (p - \pi^0) + O_p(m^{-\frac{1}{2}})
\]
\[=: P^* \sqrt{m} (p - \pi^0) + O_p(m^{-\frac{1}{2}}). \tag{30}\]

Therefore
\[
\sqrt{m}(\hat{\tau} - \tau^0) = \left( \frac{\partial \tau^0}{\partial \zeta} + O_p(m^{-\frac{1}{2}}) \right) P^* \sqrt{m} (p - \pi^0) + O_p(m^{-\frac{1}{2}})
\]
\[= B^* P^* \sqrt{m} (p - \pi^0) + O_p(m^{-\frac{1}{2}}) \]
\[=: E^* \sqrt{m} (p - \pi^0) + O_p(m^{-\frac{1}{2}}). \tag{31}\]

Hence
\[
\sqrt{m} \left( \frac{p - \pi^0}{\hat{\tau} - \tau^0} \right) = \left( \frac{I}{E^*} \right) \sqrt{m} (p - \pi^0) + O_p(m^{-\frac{1}{2}}).
\]

[1] Alexander L. Read. Presentation of search results: The CL(s) technique. J.Phys., G28:2693–2704, 2002.
[2] G. Aad et al. Phys. Lett., B716, 2012.
[3] S. Chatchyan et al. Phys. Lett., B716, 2012.
[4] A. Aguilar-Arevalo et al. Phys. Rev., D64:112007, 2001.
[5] A. A. Aguilar-Arevalo et al. Phys. Rev. Lett., 110:16, 2013.
[6] G. Mention et al. Phys. Rev., D83:073006, 2011.
[7] K. Nakamura et al. J. Phys., G37:075021, 2010.
[8] Alexander L. Read. Modified frequentist analysis of search results (The CL(s) method). 2000.
[9] X. Qian, A. Tan, W. Wang, J.J. Ling, R.D. McKeown, et al. Statistical Evaluation of Experimental Determinations of Neutrino Mass Hierarchy. Phys.Rev., D86:113011, 2012, 1210.3651.
[10] Mattias Blennow, Pilar Coloma, Patrick Huber, and Thomas Schwetz. Quantifying the sensitivity of oscillation experiments to the neutrino mass ordering. JHEP, 1403:028, 2014, 1311.1822.
[11] S. S. Wilks. The Annals of Mathematical Statistics, 9, 1938.
[12] Gary J. Feldman and Robert D. Cousins. A Unified approach to the classical statistical analysis of small signals. Phys.Rev., D57:3873–3889, 1998, physics/9711021.
[13] G. Cowan et al. Eur. Phys. J., C71:1554, 2011.
[14] F.P. An et al. Spectral measurement of electron antineutrino oscillation amplitude and frequency at Daya Bay. Phys.Rev.Lett., 112:061801, 2014, 1310.6732.
[15] Louis Lyons. Open statistical issues in Particle Physics. Ann. Appl. Stat., 2:887–915, 2008.
[16] A. Anokhina, A. Bagulya, M. Benettoni, P. Bernardini, R. Brugnera, et al. Prospects for the measurement of muon-neutrino disappearance at the FNAL-Booster. 2014, 1404.2521.
[17] F.P. An et al. Search for a Light Sterile Neutrino at Daya Bay. 2014, 1407.7259.
[18] Steve Baker and Robert D. Cousins. Clarification of the Use of Chi Square and Likelihood Functions in Fits to Histograms. Nucl.Instrum.Meth., 221:437–442, 1984.
[19] A. W. Van der Vaart. Asymptotic Statistics. The Press Syndicate of the University of Cambridge, 1998.
[20] H. White. Maximum Likelihood Estimation of Misspecified Models. Econometrica, 50, 1982.