Importance sampling method for Feldman-Cousins confidence intervals

Lukas Berns[*]
Tohoku University, Sendai, Miyagi, Japan

(Received 11 February 2024; accepted 4 April 2024; published 8 May 2024)

In various high-energy physics contexts, such as neutrino-oscillation experiments, several assumptions underlying the typical asymptotic confidence interval construction are violated, such that one has to resort to computationally expensive methods like the Feldman-Cousins method for obtaining confidence intervals with proper statistical coverage. By construction, the computation of intervals at high confidence levels requires fitting millions or billions of pseudoexperiments, while wasting most of the computational cost on overly precise intervals at low confidence levels. In this work, a simple importance sampling method is introduced that reuses pseudoexperiments produced for all tested parameter values in a single mixture distribution. This results in a significant error reduction on the estimated critical values, especially at high confidence levels, and simultaneously yields a correct interpolation of these critical values between the parameter values at which the pseudoexperiments were produced. The theoretically calculated performance is demonstrated numerically using a simple example from the analysis of neutrino oscillations. The relationship to similar techniques applied in statistical mechanics and p-value computations is discussed.

DOI: 10.1103/PhysRevD.109.092002

I. INTRODUCTION

An essential part of any experiment is the statistical analysis to extract information about the model parameters, such as physics constants, from the measurement outcome. As measurements inherently include statistical fluctuations, one often reports these constraints in the form of confidence intervals (or confidence regions in higher dimensions). These are intervals over the parameter space calculated from the observed data, which are constructed in such a way that, for any true value of the parameters, at least a predefined percentage of the possible experimental outcomes would produce an interval that covers the true parameter value. The predefined percentage over possible experimental outcomes is called the confidence level (CL).

For the rest of this paper, we shall use the following notation: x denotes the experimental outcome, which can be a vector of many observations within the single experiment. θ denotes the model parameters, which can contain one or higher-dimensional continuous degrees of freedom and may contain discrete degrees of freedom as well. $p(x|\theta)$ denotes the probability distribution function for the experimental outcomes given some model parameters. $p(x|\theta)$, seen as a function of θ for a given experimental outcome, is called the likelihood function and denoted $L(\theta|x) \coloneqq p(x|\theta)$. The parameter value for which the likelihood is maximized is denoted $\hat{\theta}(x) \coloneqq \mathrm{argmax}_\theta L(\theta|x)$, and the difference of the log-likelihood at some parameter value to the maximum likelihood is denoted as $\Delta\chi^2(\theta|x) \coloneqq -2\log L(\theta|x)/L(\hat{\theta}(x)|x)$. The confidence level is denoted $1-\alpha$.

In many cases, a useful theorem by Wilks [1] can be applied, which greatly simplifies the construction of such confidence intervals. The theorem says that, in the asymptotic limit, $\Delta\chi^2(\theta|x)$ evaluated at the true parameter value is distributed as a χ^2 distribution with k degrees of freedom, where k is the dimension of the parameter space θ, which has to be continuous. The theorem holds under suitable conditions that ensure that a maximum likelihood value can be found in the neighborhood of the true parameter value with a quadratic Taylor expansion of the likelihood. Given this asymptotic distribution, one can thus construct a confidence interval by all values of θ that satisfy $\Delta\chi^2(\theta|x) \leq \Delta\chi^2_c$, where the critical value $\Delta\chi^2_c$ is easily computed from the quantile function of the χ^2 distribution.

Because of the necessary assumptions, confidence intervals based on Wilks's theorem are not suitable if the number of observations is small or the parameter space is unsuitable because of physical boundaries (such as $\theta \geq 0$), discrete degrees of freedom, or periodicities that cannot be captured by the quadratic expansion. Neutrino-oscillation experiments, for example, suffer from all of these deficiencies, for which we will present an example later. In this situation, one has to resort to actually producing ensembles of pseudoexperiments for selected

[*] lukasb@epx.phys.tohoku.ac.jp

parameter values to study the distribution of a suitable statistic to be used for the construction of the confidence interval.

A commonly used method is the Feldman-Cousins (FC) method [2], where for each pseudoexperiment x' generated assuming a true value θ_t, the $\Delta\chi^2(\theta_t|x')$ value at the true parameter value is computed to obtain its distribution. Then the critical value $\Delta\chi^2_c$ is obtained by the empirical $1-\alpha$ percentile of this distribution. Since the distribution of $\Delta\chi^2(\theta_t|x')$ will, in general, be different for each true parameter value, the critical values are now a function of the true value at which they are computed, which we denote as $\Delta\chi^2_c(\theta_t)$. Finally, the confidence interval for the actually observed data x is constructed by choosing $\Delta\chi^2(\theta|x) \leq \Delta\chi^2_c(\theta)$. In practice, it is only possible to compute $\Delta\chi^2_c(\theta)$ at selected parameter values, which need to be interpolated, for example, linearly, in order to compute the confidence intervals.

The Feldman-Cousins method is very inefficient for obtaining high-CL intervals because, by definition, only a small fraction of pseudoexperiments contribute to the quantile computation. For example, in particle physics the threshold for "discovery" is commonly chosen at $\alpha = 5.7 \times 10^{-7}$ (the "5σ" threshold), in which case only one in 1.7×10^6 pseudoexperiments would (by definition) have a $\Delta\chi^2(\theta_t|x')$ value larger than the critical value. As a result, one easily ends up with millions of pseudoexperiments to be fitted in order to obtain the necessary critical values, while simultaneously "wasting" most of this computation time for overprecise critical values at lower CL. In practice, FC confidence intervals are often computed only up to 2σ ($\alpha = 4.6 \times 10^{-2}$) or 3σ CL ($\alpha = 2.7 \times 10^{-3}$) for such reasons.

In this work, we show that it is actually extremely easy to introduce an alternative sampling distribution that generates high-CL pseudoexperiments much more frequently: one simply reuses the pseudoexperiments generated at the values of the parameters in the form of a mixture distribution. By appropriate reweighting, this results in an exponential reduction in the errors on critical values for high CL. The method also introduces a method for correctly interpolating the critical values between the subset of true parameter values, thus removing the need for naive interpolation methods that are commonly employed.

The paper is organized as follows. First, we review the conventional-FC method. Next, we define the new method, deriving it from a discussion of an ideal importance sampling distribution. Bounds for the importance sampling weights are calculated, which are used to calculate the reduction of errors on the estimated critical values compared to the conventional-FC method. The ability to interpolate critical values and the calculation of errors and other diagnostics are discussed. Next, a toy example from the analysis of neutrino oscillations is used to compare the two computation methods and the improvement is checked against the theoretical upper bounds from the previous section. Finally, we discuss the relationship to similar techniques in statistical mechanics and p-value calculations, the relationship to Bayesian marginalized likelihoods, and the limit of applicability in the presence of nuisance parameters.

II. CRITICAL VALUES IN THE CONVENTIONAL FELDMAN-COUSINS METHOD

To prepare the notation, we briefly review the computation of critical values in the conventional Feldman-Cousins method. First, we make a choice of S points in the parameter space, which we denote θ_s with s going from 1 to S. At each θ_s, we now generate an ensemble of n_{exp} pseudoexperiments $\{x\}_s$ by sampling from $p(x|\theta_s)$. While all pseudoexperiments are assumed to live in the same space, the s suffix on the curly brackets representing the ensemble keeps track of the distribution that generated the experiments. For each pseudoexperiment $x \in \{x\}_s$, we now compute $\Delta\chi^2(\theta_s|x)$ and find the $1-\alpha$ quantile $\Delta\chi^2_{c,s}$ through any suitable estimator. For example, one may simply sort the $\Delta\chi^2(\theta_s|x)$ values and take the $\lfloor \alpha \times n_{\mathrm{exp}} \rfloor$ largest value as $\Delta\chi^2_{c,s}$, in which case we have

$$\sum_{x \in \{x\}_s} I(\Delta\chi^2(\theta_s|x) \geq \Delta\chi^2_{c,s}) = \lfloor \alpha \times n_{\mathrm{exp}} \rfloor. \tag{1}$$

Here, $I(\cdot)$ is the indicator function returning 1 if the logical statement in the parentheses is true, and 0 otherwise. $\lfloor \cdot \rfloor$ denotes the floor function.

Finally, the critical value function $\Delta\chi^2_c(\theta)$ is obtained by some interpolation scheme. For example, one may set $\Delta\chi^2_c(\theta_s) \coloneqq \Delta\chi^2_{c,s}$ and linearly interpolate for any θ values in between. To reduce the interpolation error, one typically has to either manually or automatically [3] adjust the choice of sampling parameter values $\{\theta\}_S$ in an iterative scheme.

The asymptotic variance on the critical values is proportional to the binomial error $\alpha(1-\alpha)/n_{\mathrm{exp}}$, so high CL ($\alpha \ll 1$) generally means that one needs $n_{\mathrm{exp}} \gg 1/\alpha$ for reliable critical values. Since the whole process is repeated for all S points in the parameter space, the total number of generated (and fitted) pseudoexperiments is $S \times n_{\mathrm{exp}} \gg S/\alpha$.

III. THE MIXTURE FELDMAN-COUSINS METHOD

A. Definition

Our new method, which we shall refer to as the "mixture Feldman-Cousins" method, differs from the conventional method mainly in the reuse of *all* generated pseudoexperiments for the critical-value computation of *each* target parameter space point θ_t with an additional weight

$$w(x|\theta_t) \coloneqq \frac{p(x|\theta_t)}{\frac{1}{S}\sum_{s=1}^{S} p(x|\theta_s)} = \frac{1}{\frac{1}{S}\sum_{s=1}^{S} \exp\left[-\frac{1}{2}\{\Delta\chi^2(\theta_s|x) - \Delta\chi^2(\theta_t|x)\}\right]}. \quad (2)$$

The value in the denominator is the asymptotic sampling probability distribution of $x \in \{x\}_{\mathrm{mix}} \coloneqq \cup_{s=1}^{S}\{x\}_s$, which is the mixture distribution of $p(x|\theta_s)$ for all θ_s values (see Appendix A). A mixture distribution is the distribution one obtains by randomly sampling from several separate probability distributions to form a single ensemble, which results in a probability distribution function equal to the weighted sum of the individual probability distribution functions. Since the weights from Eq. (2) are based on the sampling probabilities, which are nothing but the likelihood function, they are computable using the same procedure that calculates the $\Delta\chi^2(\theta|x)$ for each pseudoexperiment. Because of taking the difference of two $\Delta\chi^2$ values, the contribution from the minimum χ^2 at $\hat{\theta}(x)$, as well as any θ-independent offsets (e.g., the $n!$ factor in the Poisson likelihood) vanishes in the denominator and hence does not need to be known accurately. While in the conventional method one only needs to compute $\Delta\chi^2(\theta_s|x)$ for the θ_s value at which the pseudoexperiment was generated, here we need it for all $\theta_{s'}$ (including $s' \neq s$) and θ_t.

Now we can define the critical value $\Delta\chi^2_{c,t}$ as the w-weighted $1-\alpha$ quantile of $\Delta\chi^2(\theta_t|x)$ for $x \sim \{x\}_{\mathrm{mix}}$, for example,

$$\sum_{x\in\{x\}_{\mathrm{mix}}} w(x|\theta_t) I(\Delta\chi^2(\theta_t|x) \geq \Delta\chi^2_{c,t}) \lesssim \alpha \times S n_{\mathrm{exp}}, \quad (3)$$

where the $\lesssim$ is meant to represent that we take the smallest $\Delta\chi^2_{c,t}$ that satisfies the inequality.

B. Derivation

In order to obtain more pseudoexperiments at large $\Delta\chi^2$ values, which would yield more precise high-CL critical values, we use an importance sampling approach: instead of directly sampling from the target distribution $p(x|\theta_t)$, we sample from a different distribution and weight the sampled pseudoexperiments by the ratio of probability distributions to calculate the relevant quantities under the target distribution (the critical values). The question therefore becomes: what is the ideal sampling distribution to generate the desired pseudoexperiments?.

Note that it is important to find a sampling distribution that is as close as possible to the target distribution apart from generating high-$\Delta\chi^2$ pseudoexperiments with higher probability. In particular, if each experiment x consists of m measurements, the experiments are points in an m-dimensional space and there are m dimensions in which we can stretch or shrink the sampling distribution. Instead of thinking about estimating quantiles, let us think of estimating the probability density $p(Y(x)|\theta_t)$ using histograms for $Y(x) \coloneqq \Delta\chi^2(\theta_t|x)$. When using reweighting, in addition to the binomial error $n_{\mathrm{exp}}p(1-p)$ for the number of pseudoexperiments falling into a bin, there will be an additional contribution due to the variance of weights (Appendix B).

We therefore want to increase the number of pseudoexperiments falling into a high-$Y(x)$ bin to reduce the binomial error, while at the same time keeping the weight variance within the bin as small as possible. This means the ideal case of 0 variance would be for the weights to depend on x through $Y(x)$ alone. Or equivalently, since the weights are the ratio of the target and sampling distribution, we want to use a sampling distribution that differs from the target distribution only by a functional factor of $\Delta\chi^2(\theta_t|x)$.

The key idea is to think about the meaning of a high-$\Delta\chi^2(\theta_t|x)$ value. The likelihood $L(\theta|x)$ is the probability to sample the given pseudoexperiment x from θ. A high $\Delta\chi^2(\theta_t|x) = -2\log L(\theta_t|x)/L(\hat{\theta}(x)|x)$ means there exists a value $\hat{\theta}(x)$ where it is more likely to sample the given pseudoexperiment than at the "target" θ_t value. Thus by using pseudoexperiments generated at $\theta \neq \theta_t$, we can more efficiently obtain ones with high $\Delta\chi^2(\theta_t|x)$.

The naive choice of simply using pseudoexperiments generated at some θ' ($\neq \theta_t$) weighted by the ratio of sampling probabilities $p(x|\theta_t)/p(x|\theta')$, however, will do worse than before. This is because $\hat{\theta}(x)$ depends on the pseudoexperiment x, such that for some pseudoexperiments it may be preferable to sample x from $p(x|\theta_t)$ than from $p(x|\theta')$, resulting in an exponentially large (often unbounded) variance of weights.

The solution is simple: by using a mixture distribution $p_{\mathrm{sample}}(x) = \frac{1}{S}\sum_{s=1}^{S} p(x|\theta_s)$ over a set $\{\theta\}_S \coloneqq \{\theta_1, \theta_2, \ldots, \theta_S\}$ which includes θ_t, we can guarantee the weights to be bounded from above, because $p_{\mathrm{sample}}(x) \geq \frac{1}{S} p(x|\theta_t)$ and hence

$$w(x|\theta_t) \leq S. \quad (4)$$

C. Bounds on pseudoexperiment weights for a good grid

If we choose the grid $\{\theta\}_S$ dense *and* wide enough (a "good" grid) such that we may assume to have a good minimum $\hat{\theta}_S(x)$ on $\{\theta\}_S$ in the following sense, we can put a much stricter bound on the weights than Eq. (4). Given the minimum $\Delta\chi^2$ *on the grid*,

$$\Delta\chi^2(\hat{\theta}_S(x)|x) \coloneqq \min_s \Delta\chi^2(\theta_s|x), \quad (5)$$

which corresponds to the difference between the smallest χ^2 over $\theta \in \{\theta\}_S$ and over the full continuous parameter space of θ, we assume this difference to be bounded from above by

$$\Delta\chi^2(\hat{\theta}_S(x)|x) \le \begin{cases} \epsilon & \text{if } \Delta\chi^2(\theta_t|x) \le \Delta\chi^2_{\max} \\ \Delta\chi^2(\theta_t|x) & \text{otherwise} \end{cases} \tag{6}$$

for all possible x. This upper bound is characterized by the two parameters $\epsilon \lesssim 1$ and $\Delta\chi^2_{\max}$ related to the grid spacing and the range covered by the grid, respectively. The upper bound due to ϵ applies to pseudoexperiments for which the true minimum $\hat{\theta}(x)$ is within the range covered by the grid, with ϵ being smaller for denser spacing of $\{\theta\}_S$. $\Delta\chi^2_{\max}$ is a parameter that limits the range to which the ϵ bound can be applied and has to be introduced because it is not possible to fully cover the parameter space of θ with a finite grid unless the parameter space is bounded. It gives the fraction of pseudoexperiments whose minimum $\hat{\theta}(x)$ lies outside of the grid range. With larger fluctuations of this minimum away from the grid points, the minimum over the grid $\Delta\chi^2(\hat{\theta}_S(x)|x)$ can become arbitrarily large, such that it would not be possible to impose an upper bound in the most general sense. However, since in this case $\Delta\chi^2(\theta_t|x)$ will be large as well, we can keep the ϵ bound by restricting its applicability to pseudoexperiments satisfying $\Delta\chi^2(\theta_t|x) \le \Delta\chi^2_{\max}$. For pseudoexperiments that violate this inequality, we still have an upper bound of $\Delta\chi^2(\theta_t|x)$ by definition of $\hat{\theta}_S(x)$, if $\theta_t \in \{\theta\}_S$. Since the upper bound due to ϵ is significantly stronger, it is preferable to increase the value of $\Delta\chi^2_{\max}$ by preparing a grid $\{\theta\}_S$ that covers a wider range of values in θ. In the later example, we will briefly show how ϵ and $\Delta\chi^2_{\max}$ can be estimated for practical problems. In order to extend the following discussion to the case of $\theta_t \notin \{\theta\}_S$ as well, we define a symbol C which is 1 if $\theta_t \in \{\theta\}_S$ and 0 otherwise. Note that to guarantee Eq. (6) under $C = 0$ one generally needs to have parameter values in $\{\theta\}_S$ that surround θ_t sufficiently well. For example, with a one-dimensional continuous θ parameter, one needs $\min_s \theta_s \le \theta_t \le \max_s \theta_s$.

First, we focus on the pseudoexperiments with $\Delta\chi^2(\theta_t|x) \le \Delta\chi^2_{\max}$, which are our primary interest, and note that

$$\begin{aligned}\frac{p(x|\hat{\theta}_S(x))}{p(x|\theta_t)} &= \exp\left[\frac{1}{2}\{\Delta\chi^2(\theta_t|x) - \Delta\chi^2(\hat{\theta}_S(x)|x)\}\right] \\ &\ge \exp\left[\frac{1}{2}\Delta\chi^2(\theta_t|x) - \frac{\epsilon}{2}\right].\end{aligned} \tag{7}$$

The sum of probability ratios is now bounded from below by

$$\begin{aligned}\sum_{s=1}^{S}\frac{p(x|\theta_s)}{p(x|\theta_t)} &\ge \frac{p(x|\hat{\theta}_S(x))}{p(x|\theta_t)} \\ &\ge \exp\left[\frac{1}{2}\Delta\chi^2(\theta_t|x) - \frac{\epsilon}{2}\right]\end{aligned} \tag{8}$$

with first inequality following from $\hat{\theta}_S(x) \in \{\theta\}_S$ and nonnegativity of probability. This means that, for any pseudoexperiment with $\epsilon < \Delta\chi^2(\theta_t|x) \le \Delta\chi^2_{\max}$, it is more likely to be sampled in Sn_{exp} samples from $\frac{1}{S}\sum_{s=1}^{S} p(x|\theta_s)$ than in n_{exp} samples from the target distribution $p(x|\theta_t)$. The sum of probability ratios is further bounded from above by

$$\begin{aligned}\sum_{s=1}^{S}\frac{p(x|\theta_s)}{p(x|\theta_t)} &= C \times \frac{p(x|\theta_t)}{p(x|\theta_t)} + \sum_{s(\neq t)}\frac{L(\theta_s|x)}{L(\theta_t|x)} \\ &\le C + (S - C)\exp\left[\frac{1}{2}\Delta\chi^2(\theta_t|x)\right]\end{aligned} \tag{9}$$

because $L(\theta_s|x) \le L(\hat{\theta}(x)|x)$. This means the weights are bounded by

$$\begin{aligned}&\frac{S}{C + (S - C)\exp\left[\frac{1}{2}\Delta\chi^2(\theta_t|x)\right]} \\ &\le w(x|\theta_t) \le \frac{S}{\exp\left[\frac{1}{2}\Delta\chi^2(\theta_t|x) - \frac{\epsilon}{2}\right]}.\end{aligned} \tag{10}$$

We see that the bounds depend on the pseudoexperiments through $\Delta\chi^2(\theta_t|x)$ only and also note that, for sufficiently large $\Delta\chi^2(\theta_t|x)$, the ratio of upper $w_{\max}$ to lower bound $w_{\min}$ converges to

$$\frac{w_{\max}}{w_{\min}} \to (S - C)e^{\epsilon/2}, \tag{11}$$

which indicates a small relative variance of weights as long as the number of grid points S is not a very large number.

For pseudoexperiments with $\Delta\chi^2(\theta_t|x)$ above the threshold $\Delta\chi^2_{\max}$, we have

$$\begin{aligned}\frac{p(x|\hat{\theta}_S(x))}{p(x|\theta_t)} &= \exp\left[\frac{1}{2}\{\Delta\chi^2(\theta_t|x) - \Delta\chi^2(\hat{\theta}_S(x)|x)\}\right] \\ &\ge 1\end{aligned} \tag{12}$$

by Eq. (6) and hence an upper bound on the weights

$$w(x|\theta_t) \le \frac{S \times p(x|\theta_t)}{p(x|\hat{\theta}_S(x))} \le S. \tag{13}$$

D. Critical value estimator performance with a good grid

Since quantiles (the critical values) are just the inverse function of the cumulative distribution function (CDF), we can estimate the relative reduction of the quantile estimation variance by the reduction of the CDF estimation variance. The relationship for an observable $y \sim f(y)$ is given by $\mathrm{Var}[\hat{y}(P)] = f(y)^{-2}\mathrm{Var}[\hat{P}(y)]$ where $\hat{y}(P)$ is the quantile function estimator, $\hat{P}(y)$ is the CDF estimator, and $f(y)$ is the probability distribution function.

Following Eq. (3) and using the shorthand notation $Y(x) \coloneqq \Delta\chi^2(\theta_t|x)$, our CDF estimator is

$$\hat{P}(y|\theta_t) = \frac{1}{Sn_{\mathrm{exp}}} \sum_{x\in\{x\}_{\mathrm{mix}}} w(x|\theta_t) I(Y(x) \geq y) \tag{14}$$

with $x \sim \frac{1}{S}\sum_{s=1}^{S} p(x|\theta_s)$. This is an unbiased estimator for the target CDF $P(y|\theta_t)$,

$$\mathbb{E}[\hat{P}(y|\theta_t)] = \mathbb{E}[w(x|\theta_t)I(Y(x) \geq y)] \tag{15}$$

$$= \mathbb{E}_t[I(Y(x) \geq y)] \tag{16}$$

$$= P(y|\theta_t), \tag{17}$$

where $\mathbb{E}[\cdot]$ means to take the expectation with $x \sim \frac{1}{S}\sum_{s=1}^{S} p(x|\theta_s)$, and $\mathbb{E}_t[\cdot]$ means to take the expectation with $x \sim p(x|\theta_t)$. Now, defining

$$y_{\mathrm{max}} \coloneqq \max\{\Delta\chi^2_{\mathrm{max}}, y\}, \tag{18}$$

the variance from a single pseudoexperiment is

$$\mathrm{Var}[w(x|\theta_t)I(Y(x) \geq y)] \tag{19}$$

$$= \mathbb{E}[w(x|\theta_t)^2 I(Y(x) \geq y)^2] - \mathbb{E}[w(x|\theta_t)I(Y(x) \geq y)]^2 \tag{20}$$

$$= \mathbb{E}_t[w(x|\theta_t)I(Y(x) \geq y)] - P(y|\theta_t)^2 \tag{21}$$

$$\leq \mathbb{E}_t\left[\frac{S \times I(y \leq Y(x) \leq y_{\mathrm{max}})}{\exp\left[\frac{1}{2}(Y(x)-\epsilon)\right]}\right] + S \times \mathbb{E}_t[I(Y(x) \geq y_{\mathrm{max}})] - P(y|\theta_t)^2 \tag{22}$$

$$\leq \frac{S}{\exp\left[\frac{1}{2}(y-\epsilon)\right]} \mathbb{E}_t[I(y \leq Y(x) \leq y_{\mathrm{max}})] + S \times P(y_{\mathrm{max}}|\theta_t) - P(y|\theta_t)^2 \tag{23}$$

$$= S \times \left[\frac{P(y|\theta_t) - P(y_{\mathrm{max}}|\theta_t)}{\exp\left[\frac{1}{2}(y-\epsilon)\right]} + P(y_{\mathrm{max}}|\theta_t)\right] - P(y|\theta_t)^2, \tag{24}$$

where in going from the second to the third line, we used $I(\cdot)^2 = I(\cdot)$ and reduced a power of w to replace $\mathbb{E}[\cdot]$ by $\mathbb{E}_t[\cdot]$; in going to the fourth line, we used the upper bound from Eqs. (10) and (13); in going to the fifth line, we used $Y(x) \geq y$ from the argument of the indicator function. The variance of the CDF estimator is therefore

$$\mathrm{Var}[\hat{P}(y|\theta_t)] \tag{25}$$

$$= \frac{1}{Sn_{\mathrm{exp}}} \mathrm{Var}[w(x|\theta_t)I(\Delta\chi^2(\theta_t|x) \geq y)] \tag{26}$$

$$\leq \frac{1}{n_{\mathrm{exp}}}\left(\frac{P(y|\theta_t) - P(y_{\mathrm{max}}|\theta_t)}{\exp[\frac{1}{2}(y-\epsilon)]} + P(y_{\mathrm{max}}|\theta_t) - \frac{P(y|\theta_t)^2}{S}\right), \tag{27}$$

where we note that the S factors in the first two terms were canceled thanks to being able to reuse the pseudoexperiments generated at all S values for the CDF estimation of each θ_t value.

For reference, the variance on the CDF estimator in the conventional-FC method (denoted in the following equations by “conv”) is given by the binomial error

$$\mathrm{Var}_t[\hat{P}_{\mathrm{conv}}(y)] = \frac{1}{n_{\mathrm{exp}}}(P(y|\theta_t) - P(y|\theta_t)^2), \tag{28}$$

with $\mathrm{Var}_t[\cdot]$ being the variance under $x \sim p(x|\theta_t)$. Hence the variance on the estimated critical values $\hat{y}(P|\theta_t)$ in the mixture-FC method is smaller by the factor

$$\gamma \coloneqq \frac{\mathrm{Var}[\hat{y}(P|\theta_t)]}{\mathrm{Var}_t[\hat{y}_{\mathrm{conv}}(P|\theta_t)]} \tag{29}$$

$$= \frac{\mathrm{Var}[\hat{P}(y|\theta_t)]}{\mathrm{Var}_t[\hat{P}_{\mathrm{conv}}(y|\theta_t)]} \tag{30}$$

$$\leq \frac{A(y) + B(y)P(y_{\mathrm{max}}|\theta_t)/P - \frac{1}{S}P}{1-P}, \tag{31}$$

$$A(y) \coloneqq \frac{1}{\exp\left[\frac{1}{2}(y-\epsilon)\right]}, \tag{32}$$

$$B(y) \coloneqq 1 - A(y), \tag{33}$$

where y is the true P quantile satisfying $P(y|\theta_t) = P$, and the important behavior will be the exponential decrease of $A(y)$ as a function of y. The typical functional shape of the upper bound is shown in Fig. 1(a). Let us first consider the case of $P(y_{\mathrm{max}}|\theta_t) \ll P$. For the $P \leq 1/2$ values one is typically interested in, the mixture-FC method obtains more precise critical values than the conventional method (i.e., $\gamma \leq 1$) for all $y \geq \epsilon + 2\ln 2$. As y increases, the relative variance starts to decrease exponentially as $\gamma \lesssim \exp(-y/2)$. As y further increases toward $\Delta\chi^2_{\max}$, and $P(y_{\max}|\theta_t) \ll P$ fails to hold anymore, the $B(y)P(y_{\max}|\theta_t)/P$ term becomes dominant, which saturates to $P(y_{\max}|\theta_t)/P = 1$ for $y \geq \Delta\chi^2_{\max}$. Hence, the improvement flattens out to $\gamma \leq 1$ for $y \geq \Delta\chi^2_{\max}$ to leading order in P, which is still at least as good as the conventional-FC method. By choosing suitable parameter points $\{\theta\}_S$ and thus a suitable $\Delta\chi^2_{\max}$, critical values of the desired precision can be calculated. As the exponential reduction in variance cancels the typically exponential dependence of the CDF on the test statistic [$\exp(-y/2)$ in the case of a χ^2 distribution], the relative error on the estimated CDF becomes approximately flat over a wide range of test-statistic values [Fig. 1(b)], which is much more efficient than for the conventional FC where low-CL become overprecise with more pseudoexperiments, while high-CL still suffers from large errors.

E. Interpolation

While for the conventional-FC method one can only compute the critical values at the parameter value θ_s where the pseudoexperiments were generated, in the mixture-FC method it is sufficient to guarantee that the target parameter value θ_t is sufficiently close and surrounded by the sampling points $\{\theta\}_S$ such that condition (6) holds. Considering that for a typical setup the pseudoexperiments to be generated are the same as those used in the conventional-FC method, this means that the mixture-FC method not only reduces the uncertainty on the critical values at the sampling points $\{\theta\}_S$, but also allows interpolating the critical values between these points with similar performance.

F. Diagnostics and error estimation

As the mixture-FC method exploits the relationship of the $\Delta\chi^2$ statistic to the probability of sampling pseudoexperiments, it is essential that the calculation of $\Delta\chi^2$ matches the process used to generate the pseudoexperiments. It is, for example, not allowed to sample from a Poisson random number generator while using an approximation like Pearson's χ^2 for $\Delta\chi^2$ instead of the Poisson log-likelihood. A simple diagnostic is to calculate the average weight across all pseudoexperiments in $\{x\}_{\text{mix}}$ and check that this is equal to 1 up to statistical fluctuations. Since the same pseudoexperiments will be used for all target parameter values θ_t (of which the weights are a function), the statistical fluctuations of these average weights will be correlated for different θ_t values.

To estimate the error of the computed critical values, we recommend using resampling methods such as the nonparametric bootstrap [4] or jackknife [5] instead of simple methods like binomial errors, in order to capture not only the statistical fluctuations in the number of pseudoexperiments that fall into a range of $\Delta\chi^2$ values, but also the statistical fluctuations in their weights.

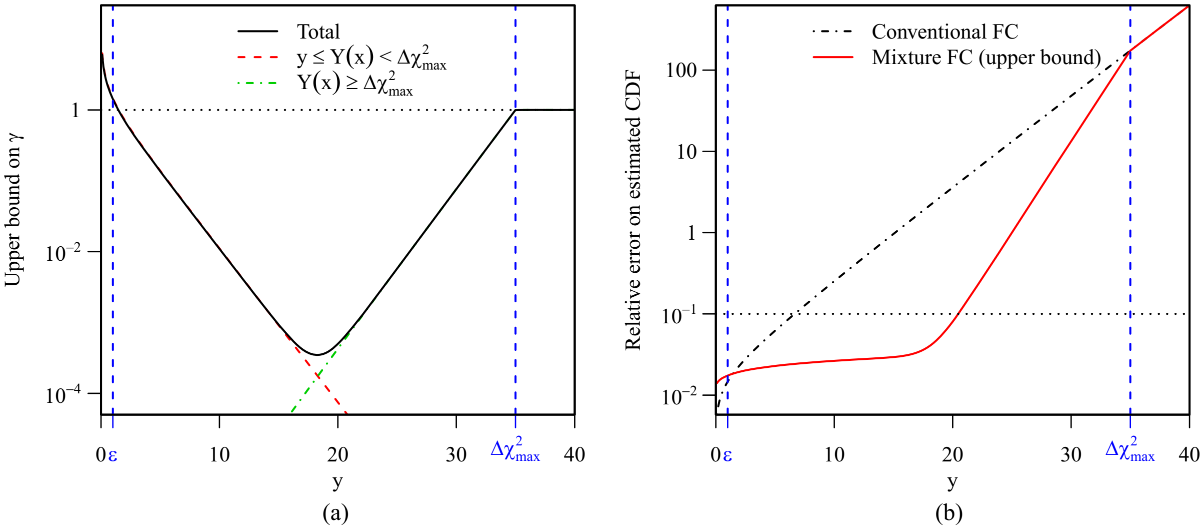


FIG. 1. (a) Example functional shape of upper bound on the ratio of estimated critical value variance from Eq. (31). The red line indicates the error contribution from pseudoexperiments with $y \leq Y(x) < \Delta\chi^2_{\max}$ [first term with $A(y)$], which is responsible for the exponential reduction of total uncertainty until the contribution from pseudoexperiments with $y \geq \Delta\chi^2_{\max}$ [second term with $B(y)$ shown by green line] takes over for very high-CL critical values. (b) The relative error on the calculated CDF estimator $\hat{P}(y|\theta_t)$ assuming $n_{\text{exp}} = 10{,}000$ pseudoexperiments at each sampling value θ_s. A reference 10% error threshold is indicated by the dotted line. The example used for both plots is constructed assuming $\epsilon = 1$, $S = 10$, $\Delta\chi^2_{\max} = 35$, and the true $Y(x)$ CDF is assumed to be χ^2 with 1 degree of freedom. In (a), the exponential growth factor for the green line depends on the assumed CDF, unlike the red line whose decay factor is given by Eq. (31).

IV. EXAMPLE WITH A SINGLE CYCLIC PARAMETER

We consider a simple example that uses a binned-Poisson model, inspired by the search for CP violation in a long-baseline neutrino-oscillation experiment, here, in particular, the T2K experiment [6]. The model has a single angular parameter called the "CP violation phase" $\delta_{CP} \in [-\pi, \pi]$ which is constrained by $B = 10$ Poisson-distributed observations $n_b \sim \text{Poisson}(\lambda_b)$ with the predicted event rate

$$\lambda_b(\delta_{CP}) \coloneqq 10 \times (1 - \phi_b^2) \times \left(1 - \frac{1}{4}\sin(\delta_{CP} + \phi_b)\right), \quad (34)$$

$$\phi_b \coloneqq \frac{b - 5.5}{10} \quad (35)$$

for each bin with index $b = 1, 2, \ldots, 10$ (Fig. 2). The main feature of this model is that one is mostly sensitive to $\sin\delta_{CP}$ through the overall normalization of approximately 100 total observations ($\sum_b n_b$) and weakly sensitive to the $\cos\delta_{CP}$ component through the "shape" of the observations as a function of b (meant to represent bins of increasing neutrino energy). Deviations from Wilks's theorem are caused by $\sin\delta_{CP}$ having physical boundaries at ± 1 (resulting in reduced critical values around $\sin\delta_{CP} = \pm 1$), the sign of $\cos\delta_{CP}$ acting as an effectively discrete degree of freedom (resulting in increased critical values at some $\sin\delta_{CP} \neq \pm 1$ values), as well as the Poisson nature of the observations. In an actual experiment, one would have further continuous and discrete physics parameters degenerate with δ_{CP} as well as various systematic uncertainties treated as nuisance parameters. For simplicity and clarity, however, we focus on δ_{CP} alone, which for continuity with the earlier sections will be referred to as $\theta = (\delta_{CP})$ and the observations as $x = (n_1, n_2, \ldots, n_{10})$.

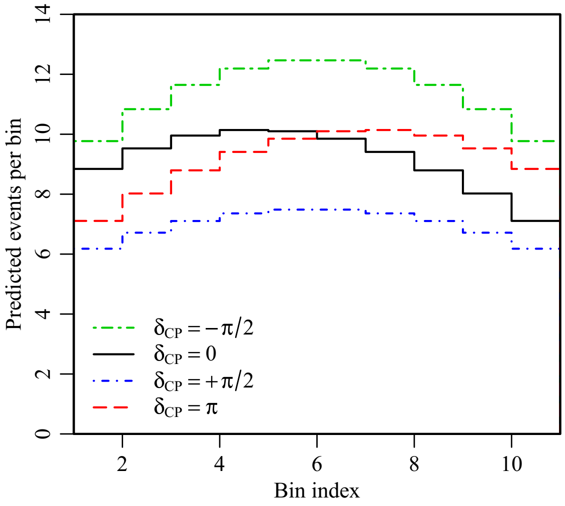


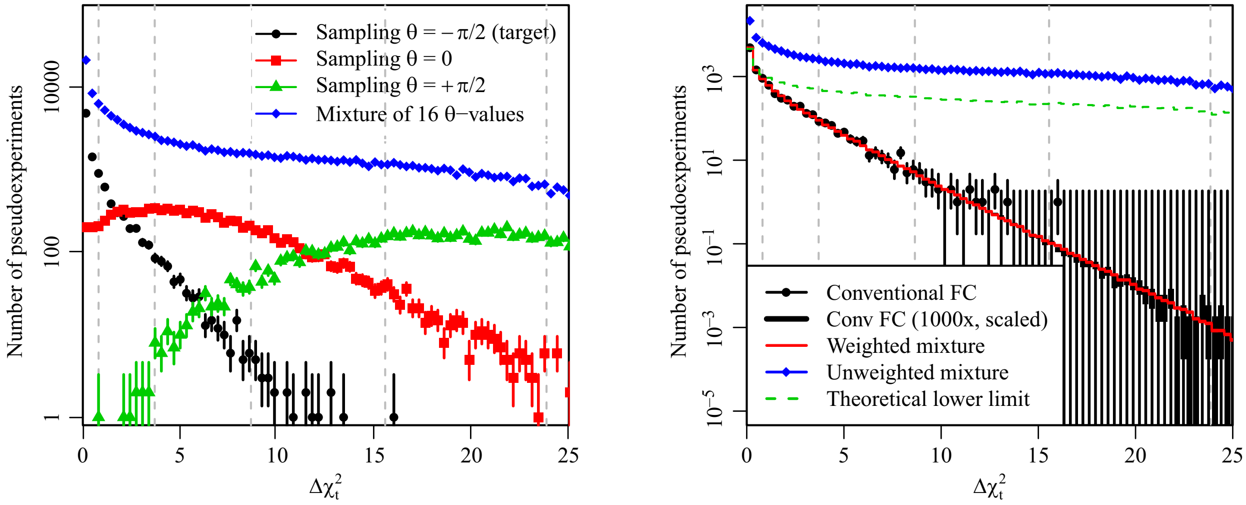
FIG. 2. Predicted number of events $\lambda_b(\delta_{CP})$ for each bin index b as used in the example.

We generate $n_{\text{exp}} = 10,000$ pseudoexperiments at each of $S = 16$ values of θ evenly distributed in the parameter range $[-\pi, \pi]$. We first focus on the target value of $\theta_t = -\pi/2$. Figure 3(a) shows the distribution of $\Delta\chi^2(\theta_t|x)$ obtained for pseudoexperiments x sampled from different θ_s values. In the conventional-FC method, only those generated at $\theta_s = \theta_t$ are used, which correspond to the black histogram that falls off quickly for large $\Delta\chi_t^2 \coloneqq \Delta\chi^2(\theta_t|x)$. In the mixture-FC method, we further make use of the pseudoexperiments generated at all other θ_s values, of which $\theta_s = 0$ and the other extreme of $\theta_s = \pi/2$ are shown by the red and green histograms, respectively. Clearly, the pseudoexperiments sampled from the shifted θ_s values have a significantly higher fraction of large $\Delta\chi_t^2$ values. At the same time, one can see one of the problems arising from using only the pseudoexperiments generated at $\theta_s = \pi/2$, in that one would need to apply very large weights for the small $\Delta\chi_t^2$ region where $\theta_s = \pi/2$ has a very small sampling probability. The mixture of pseudoexperiments generated at all 16 θ values, however, shown by the blue histogram, is able to generate more pseudoexperiments for all $\Delta\chi_t^2$ values, with the difference in slope compared to the black target histogram showing the exponential increase is pseudoexperiments for larger $\Delta\chi_t^2$ values. This is even clearer to see in Fig. 3(b), where the mixture distribution was reweighted using the assigned weights. Good agreement with the target distribution as simulated by the conventional-FC method is seen, and the total number of unweighted pseudoexperiments in the mixture-FC method exceeds the theoretical lower bound.

In order to gauge the consistency of the test-statistic distribution for large $\Delta\chi_t^2$ values, the conventional-FC calculation was repeated with 1000 times more pseudoexperiments (10×10^6 in total for the shown θ_t value), and is shown by the thick black error bars with appropriate scaling to allow comparison. We see good agreement within statistical errors ($\chi^2 = 65$ over 62 degrees of freedom for $\Delta\chi_t^2 \leq 20$). Despite the significantly larger computational overhead, the errors for the high-statistics conventional FC at large $\Delta\chi_t^2 > 20$ are still much larger compared to the mixture-FC method.

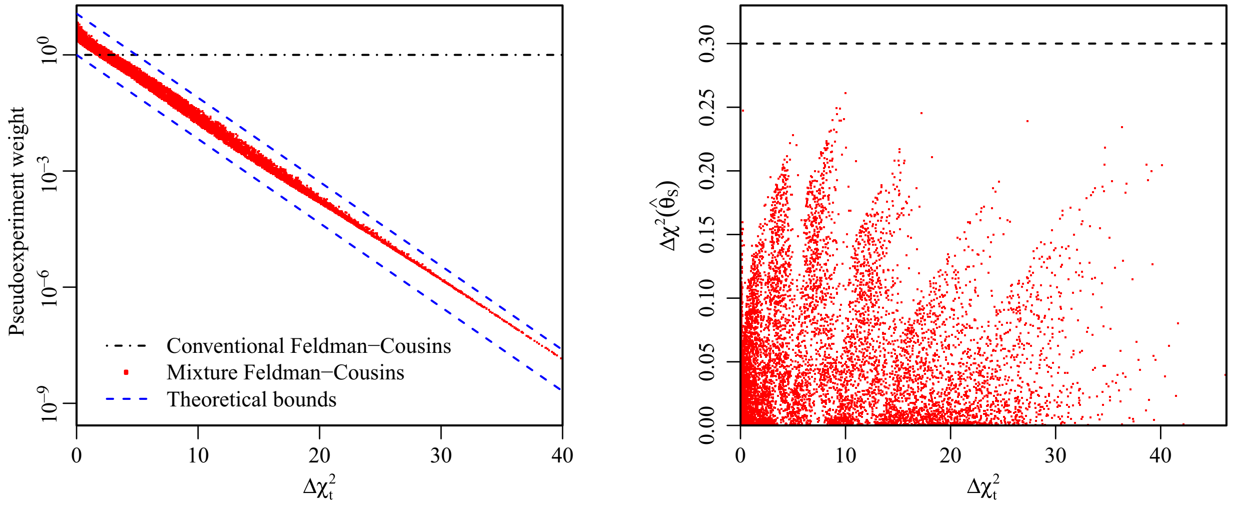
We now check some of the diagnostics for the mixture-FC method. The distribution of importance sampling weights $w(x|\theta_t)$ are shown in Fig. 4(a) and are found to be mostly a function of $\Delta\chi_t^2$ with small additional variance. The weights are found to be well contained by the theoretical bounds from Eq. (10), which were drawn assuming a $\epsilon = 0.3$ value by looking at the $\Delta\chi^2(\hat{\theta}_S(x)|x)$ distributions in Fig. 4(b). The sum of weights is found to be consistent with 1 (Fig. 5).

Next, we look at the critical values. Figure 6 shows the critical values as function of the (true/target) parameter



(a) (b)

FIG. 3. $\Delta\chi^2_t \coloneqq \Delta\chi^2(\theta_t|x)$ distributions with target $\theta_t = -\pi/2$ for (a) various sampling parameter values θ_s and (b) comparison of estimated distributions at θ_t obtained using conventional-FC and mixture-FC methods. In both plots, error bars indicate 1σ binomial confidence intervals. Vertical dashed lines indicate $1, 2, 3, 4, 5\sigma$ confidence level critical values obtained by the mixture-FC method. (a) Error bars are omitted for bins with zero entries for clarity. (b) The red "weighted mixture" histogram is also drawn with boxes representing the error from number of pseudoexperiments in each bin and their weight variance, but these errors are smaller than the linewidth and not visible. The "theoretical lower limit" on the total number of pseudoexperiments in the mixture distribution is obtained by multiplying the lower bound in Eq. (8) to the red "weighted mixture" histogram assuming $\epsilon = 0.3$.



(a) (b)

FIG. 4. Distribution of (a) weights $w(x|\theta_t)$ and (b) minimum $\Delta\chi^2$ values over $\{\theta\}_S$ for each pseudoexperiment as function of $\Delta\chi^2_t \coloneqq \Delta\chi^2(\theta_t|x)$. (a) Dashed blue lines indicate theoretical limits assuming $\epsilon = 0.3$ and $C = 0$. (b) The dashed line indicates $\epsilon = 0.3$.

value θ using both the conventional-FC method (black error bars) and the mixture-FC method. Despite using the same set of pseudoexperiments, the critical values obtained with the mixture-FC method have significantly smaller uncertainty, especially at higher CL, and also provide access to details of the functional shape between the 16 sampling values of θ.

For the 1σ critical values [Fig. 6(b)], we see that, despite the relatively fine spacing of sampling values, the interpolation error as indicated by the nonoverlap of red and gray error bands next to the $\theta = \pm\pi/2$ values is larger than the size of the binomial error band in the conventional method. As these binomial error bands do not capture the interpolation error, their smallness can be misleading and render the interpolation feature of the mixture-FC method very useful.

For the 2σ [Fig. 6(c)] and 3σ critical values [Fig. 6(d)], we see good consistency between the two methods, while

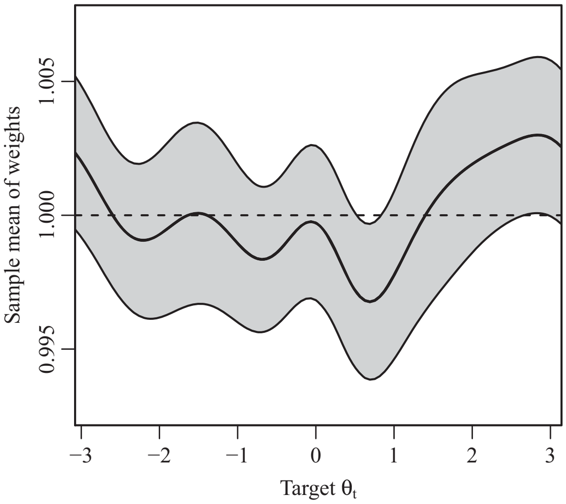


FIG. 5. The sample mean of mixture-FC pseudoexperiment weights $w(x|\theta_t)$ as a function of the target θ_t value. The error bands indicate the 1σ standard error on the mean, which are correlated between different target θ_t values.

also noting the significantly smaller errors in the mixture-FC calculation. For 3σ CL [Fig. 6(d)], we see the errors in the conventional method are already so large that some of the features of the critical values are not recognizable, such as the bumps at $\theta = 0, \pi$ and the asymmetry of critical values for a flip of the $\sin\delta_{CP}$ sign, caused by Poisson statistics.

For 4σ and higher CL [Fig. 6(a)], the conventional-FC method is unable to determine the critical values except for a lower limit. The mixture-FC method, on the other hand, still produces critical values with comparable relative error sizes to the lower CL critical values.

The estimated relative errors are plotted in Fig. 7 and are consistent with the typical shape from theoretical arguments [Fig. 1(b)]. To draw the upper bound from γ in Eq. (31), we conservatively assume $\Delta\chi^2_{\max} = 32$ based on Fig. 4(b), i.e., we will only assume $\Delta\chi^2(\hat{\theta}_S(x)|x) \leq \epsilon$ up to $\Delta\chi^2_t \leq 32$. In this example, the actual mixture-FC error estimated with the bootstrap is smaller than the theoretical upper limit from γ by about a factor of 2 for $\Delta\chi^2_t < 16$. This can be interpreted as more than one sampling value θ_s contributing to the sampling of each pseudoexperiment, rather than the assumption in the theoretical upper limit that only $\hat{\theta}_S(x)$ would contribute. For $\Delta\chi^2_t > 16 = \Delta\chi^2_{\max}/2$, on the other hand, the theoretical upper limit starts to increase significantly, whereas the actual error estimated with the bootstrap only grows slowly. This can be interpreted as our choice of $\Delta\chi^2_{\max} = 32$ being overly conservative: with the present example, the chosen sampling grid $\{\theta\}_S$ appears to be effective up to significantly higher $\Delta\chi^2$ values. This is partly due to the convenient situation of having a parameter $\theta = (\delta_{CP})$ with a bounded parameter space $\delta_{CP} \in [-\pi, \pi]$.

V. DISCUSSION

A. Relation to techniques in statistical mechanics

The presented method is similar in spirit to the "multiple histogram reweighting" (multihistogram) method [7] in statistical mechanics, where statistical ensembles are simulated for various parameter values and combined by reweighting to the desired parameter value. In the multihistogram method, the ensembles are combined with an additional per-ensemble weight, which is adjusted to minimize the overall error on the variable to be estimated. A similar per-ensemble weighting could be applied in the presented mixture-FC method as well, where these additional weights would be allowed to depend on the target $\Delta\chi^2_t \coloneqq \Delta\chi^2(\theta_t|x)$ value as well, in order to reduce the variance on the critical value estimator as much as possible.

One difference to the multihistogram method, however, is that, because we do not resort to Markov chain Monte Carlo techniques to sample the pseudoexperiments, the sampling distribution of pseudoexperiments x at each parameter value θ is known exactly including the normalization constant. Hence, the iterative procedure that is required at the end of the multihistogram method to self-consistently determine these normalization constants (the free energies) is not necessary in the mixture-FC method.

B. Relation to the marginal distribution

The sampling distribution constructed as a mixture over several parameter values $\{\theta\}_S$ can be considered a marginal probability distribution with prior $\pi(\theta) = \frac{1}{S}\sum_{s=1}^{S}\delta(\theta - \theta_s)$, where $\delta(\cdot)$ is the Dirac δ function. Additional per-ensemble weights as discussed in the previous paragraph would correspond to an alternative prior $\pi(\theta|\Delta\chi^2_t) = \sum_{s=1}^{S} r_s(\Delta\chi^2_t)\delta(\theta - \theta_s)$, where $r_s(\Delta\chi^2_t)$ can be optimized to reduce errors subject to the condition $\sum_s r_s(\Delta\chi^2_t) = 1$ for all $\Delta\chi^2_t$. One can even generalize the discussion to continuous priors $\pi(\theta|\Delta\chi^2_t)$, where in order to preserve the arguments on efficiency reduction, we would need to extend the single-point condition from Eq. (6) to a condition on a finite-size region on $\pi(\theta|\Delta\chi^2_t)$.

Unlike in the conventional-FC method, where one needs a large number of pseudoexperiments at each target parameter value, it can be preferable in the mixture-FC method to generate less pseudoexperiments at each sampling value, but instead increase the number of considered sampling points S. If Sn_{exp} is held fixed, this results in a reduction of the variance of critical values by reducing the variance in weights bounded from above by $\exp(\epsilon/2)$.

Given this relation to the marginal distribution, let us now consider the computation of $\Delta\chi^2(x|\theta_t) = -2\log L(x|\theta_t)/L(x|\hat{\theta}(x))$ as being approximated by $-2\log L(x|\theta_t)/L_m(x)$, where in the denominator, the profiling operation was replaced by a marginalization over θ with some prior over θ. We have, therefore, a simple

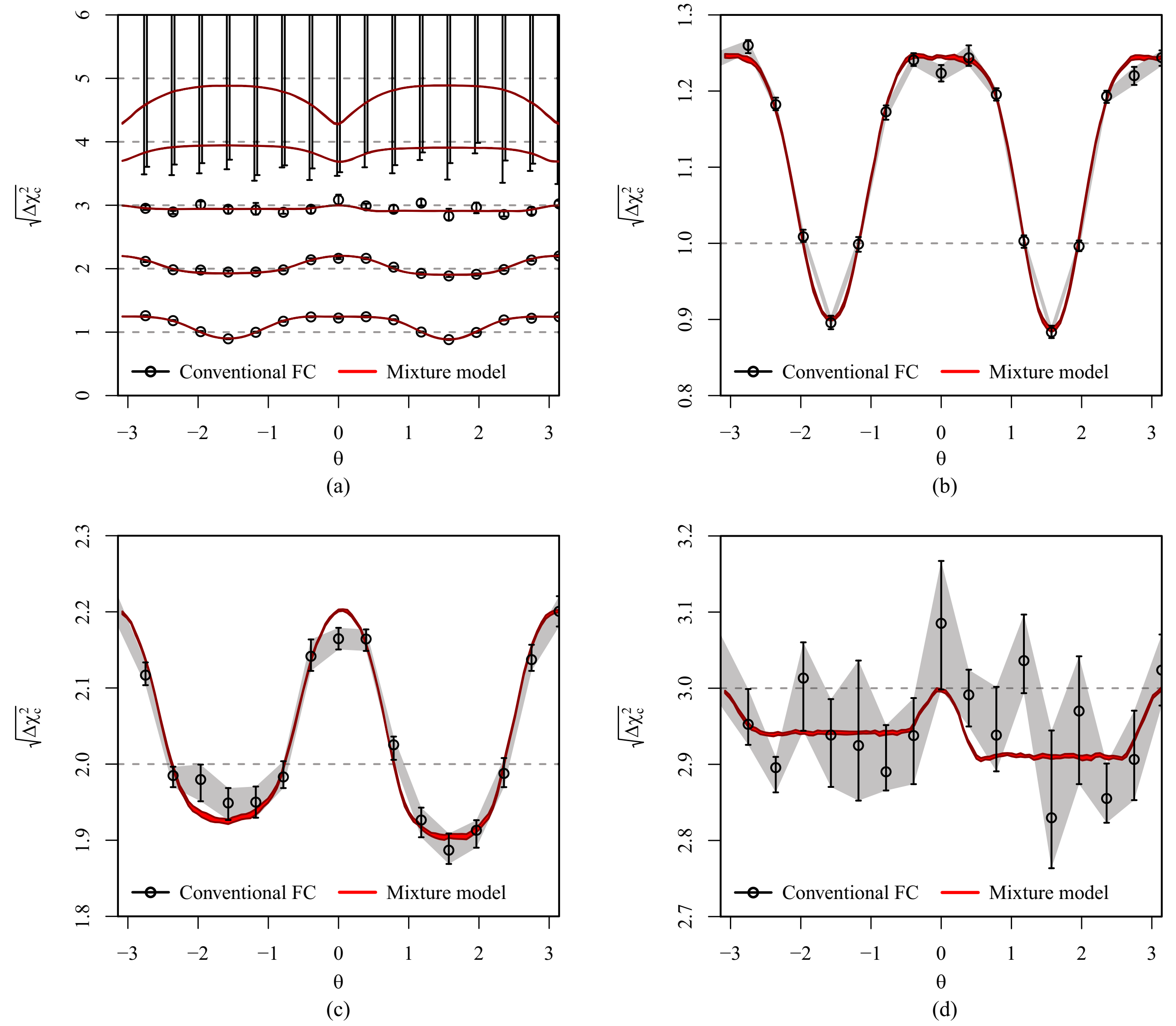


FIG. 6. $1, 2, 3, 4, 5\sigma$ confidence level critical values (given as $\sqrt{\Delta\chi_c^2}$) obtained from the same set of pseudoexperiments with the conventional-FC (black error bars) and mixture-FC method (red error bands). In both cases the error bars/bands indicate the 1σ error on the critical values obtained with binomial/bootstrap errors for the standard-/mixture-FC method, respectively. Dashed lines indicate critical values by Wilks's theorem, which are not valid here, but still drawn for reference. (a) Overview over all 1–5σ CL critical values. (b)–(d) Zoomed in view of the 1, 2, and 3σ CL critical values respectively.

likelihood-ratio test between $p(x|\theta_t)$ and $p_m(x) \coloneqq \int d\theta\pi(\theta)p(x|\theta)$ and it now becomes evident that, in order to efficiently generate pseudoexperiments with small p values under the null hypothesis $p(x|\theta_t)$, one should simply generate the pseudoexperiments from the alternative hypothesis $p_m(x)$, which is what is being done in the mixture-FC method.

In practice, it will be easier to use the discrete "prior" over $\{\theta\}_S$ as was discussed in the text, because unless the likelihood is Gaussian, the numerical integration required for marginalization usually increases the computational cost and complexity. This relation to the profiling/marginalization similarities can nevertheless be exploited to motivate an ideal spacing of $\{\theta\}_S$ values. Out of the well-known objective priors, the Jeffreys prior [8] is known to produce a prior that would be uniform in the parametrization in which the likelihood is Gaussian, if such a parametrization exists. Since profiling and marginalization with a uniform prior over a Gaussian likelihood produce equivalent results up to a constant offset, the Jeffreys prior can be considered a good candidate for choosing the $\{\theta\}_S$ values at which to generate pseudoexperiments. For example, in the *CP* violation analysis that was discussed in the earlier section, it would be more suitable to choose a uniform spacing of parameter values not in δ_{CP} but in $\sin\delta_{CP}$ with equal probabilities for the sign of $\cos\delta_{CP}$,

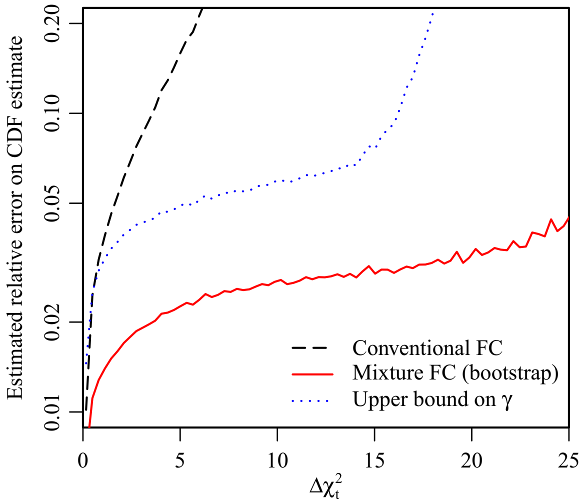


FIG. 7. Estimated relative errors on the CDF estimator $\hat{P}(y|\theta_t)$ with target $\theta_t = -\pi/2$. For the conventional-FC method, the standard error from the binomial distribution is shown (black dashed line), where the more precise CDF estimate from the mixture-FC method was used in computing these errors. For the mixture-FC method, the bootstrap error estimate (red solid line) is well below the theoretical upper limit of Eq. (31) calculated assuming $\epsilon = 0.3$ and $\Delta\chi^2_{\max} = 32$ (blue dotted line).

since the dominant constraint is due to the total number of events $N \sim \mathrm{Poisson}(\lambda = A + B \sin\delta_{CP})$ for some constants A and B, resulting in an approximately Gaussian likelihood over $\sin\delta_{CP}$.

C. Nuisance parameters

Because the significant error reduction in the mixture-FC method exploits the specific relation of the $\Delta\chi^2(\theta_t|x)$ statistic to the distribution that generates the pseudoexperiments, one cannot assume all features to directly translate to an analysis with nuisance parameters or "systematic" parameters as they are often called in physics. Especially for the commonly used methods of profile-FC [9] or posterior Highland-Cousins methods [10], where the space of nuisance parameters from which to generate the pseudoexperiments is significantly reduced based on constraints by the observed experimental data, it is possible to have situations where the straightforward application of the mixture-FC method does not yield the exponential reduction of errors on the estimated critical values given by Eq. (31). One should therefore not rely on these to estimate the number of required pseudoexperiments.

In a relatively general setting, when the target distribution is directly a part of the mixture distribution (so $C = 1$), one can show that, even in the worst case, the variance on the critical values only increases very slightly compared to the conventional method, by a factor $1/(1 - P(y))$ (see Appendix C). This factor is negligible considering that for high CL we have $P(y) \ll 1$. The weights are bounded from above by a similar limit, which is important for well-defined importance sampling behavior. The naive application of the mixture-FC method to Feldman-Cousins confidence intervals is therefore still worth a try. In fact, certain situations may yield near-exponential reduction of errors as in the case without nuisance parameters, but due the lack of theoretical guarantees it is suggested to carefully study the distribution of weights and the reliability of bootstrap error estimates in this situation.

For concreteness, we give one explicit method of dealing with nuisance parameters that is easy to implement, even if it may not yield the most optimal performance possible. To allow the straightforward extension, we specifically choose the profile-FC method to define the true values of nuisance parameters η_s for the parameter of interest value θ_s at which the pseudoexperiments are generated. In this method, one chooses

$$\eta_s = \hat{\hat{\eta}}(\theta_s|x_{\mathrm{obs}}) \coloneqq \arg\max_{\eta} L(\theta_s, \eta|x_{\mathrm{obs}}), \tag{36}$$

where x_{obs} is the actually observed data. The ensemble of pseudoexperiments $\{x\}_s$ is then defined by sampling from $p(x|\theta_s, \eta_s)$. The actual confidence intervals would be constructed, for example, using the profile log-likelihood, that is,

$$\Delta\chi^2_p(\theta|x) \coloneqq \Delta\chi^2(\theta, \hat{\hat{\eta}}(\theta|x)|x), \tag{37}$$

as test statistic. In other words, the critical values $\Delta\chi^2_{c,s}$ would simply be given by replacing $\Delta\chi^2(\theta_s|x)$ by $\Delta\chi^2_p(\theta_s|x)$ in Eq. (1). The extension to the mixture-FC method is simply given by reusing all generated pseudoexperiments for each target parameter space point θ_t (and associated η_t) with the weight

$$\begin{aligned} w(x|\theta_t,\eta_t) &\coloneqq \frac{p(x|\theta_t,\eta_t)}{\frac{1}{S}\sum_{s=1}^{S} p(x|\theta_s,\eta_s)} \\ &= \frac{1}{\frac{1}{S}\sum_{s=1}^{S} \exp\left[-\frac{1}{2}\{\Delta\chi^2(\theta_s,\eta_s|x) - \Delta\chi^2(\theta_t,\eta_t|x)\}\right]}. \end{aligned} \tag{38}$$

The important difference to the case without nuisance parameters is that the weights depend not on the value of the test statistic $\Delta\chi^2_p(\theta|x)$ used to construct the confidence intervals, but rather $\Delta\chi^2(\theta, \hat{\hat{\eta}}(\theta|x_{\mathrm{obs}})|x)$. The two differ in the dataset used to define the value of nuisance parameters at which the likelihood ratio is evaluated: $\hat{\hat{\eta}}(\theta|x)$ vs $\hat{\hat{\eta}}(\theta|x_{\mathrm{obs}})$. This subtle difference breaks the guarantee of exponential reduction of errors, in addition to requiring separate calculation of this quantity. Fortunately, the calculation of the weights is trivial because it does not require any minimization: one only needs to know the likelihood ratio between predefined points in the parameter space. Also, it is essential that η_t is defined using the *θ-dependent* profile best-fit value $\hat{\hat{\eta}}(\theta_t|x_{\mathrm{obs}})$, rather than using the global best-fit value $\hat{\eta}(x_{\mathrm{obs}}) = \mathrm{argmax}_\eta \max_\theta L(\theta, \eta | x_{\mathrm{obs}})$, since the latter would break the typical relationship between shifts in θ and η that exists, for example, in the Gaussian limit and helps recovering the sampling efficiency lost through the addition of nuisance parameters.

Because one cannot guarantee an exponential reduction of errors in a setting with nuisance parameters, the ability to interpolate critical values will be more interesting in this setting. Here it is important that the pseudoexperiments generated between neighboring θ_s values (and suitable values of nuisance parameters) sufficiently overlap in the space of pseudoexperiments. Otherwise, the mismatch between pseudoexperiment generation and the statistical model behind the test statistic may quickly result in a large spread of weight values, which would make both the estimated critical values as well as their error estimates unreliable. This is because, with nuisance parameters, there are significantly more dimensions in which the pseudoexperiments can differ, even if they have similar values for the test statistic.

In one specific situation, however, all properties discussed in earlier sections are directly applicable despite the presence of nuisance parameters. This is when using the prior Highland-Cousins method in conjunction with a marginal-$\Delta\chi^2$ statistic, where it is essential to use the same prior distribution $\pi(\eta)$ for the nuisance parameters η in both cases. This is because here the effect of nuisance parameters is entirely absorbed by the probability model to generate the pseudoexperiments, in the sense of $p(x|\theta) = \int \mathrm{d}\eta\, \pi(\eta) p(x|\theta, \eta)$, such that as far as the mixture-FC method is concerned, no nuisance parameters exist.

More detailed discussions with examples and possible modifications to the sampling distributions for pseudoexperiments will be discussed in a separate publication.

D. Relation to similar techniques for statistical inference

Very similar importance sampling techniques have been used for the calculation of p values under a null hypothesis with a likelihood-ratio statistic. For example, Woodroofe [11] discusses the case with a continuous prior over the parameter of interest. In our notation,

$$p_{\mathrm{sample}}(x) = \int \mathrm{d}\theta\, \pi(\theta) p(x|\theta) \tag{39}$$

with only a lower bound on the weights

$$w(x|\theta_t) \coloneqq \frac{p(x|\theta_t)}{p_{\mathrm{sample}}(x)} \geq \frac{p(x|\theta_t)}{p(x|\hat{\theta}(x))} = \exp\left[-\frac{1}{2}\Delta\chi^2(\theta_t|x)\right] \tag{40}$$

given, rather than an upper bound, which would be essential for showing small errors on the estimated p values. An asymptotic formula for the weights using the saddle point method is also given.

Cranmer ([12], Sec. 5.6) describes a method developed in the search for the Higgs boson by the ATLAS experiment [13]. They point out the difficulty of performing the integral over the continuous prior in Woodroofe's method and instead use a set of discrete points $\{\theta\}_S$ including θ_t, as we used for the mixture-FC method (with $C = 1$). The choice of weight function, however, is different, in that a pseudoexperiment is used only if the

$$\begin{aligned} \omega_s(x|\theta_t) &\coloneqq \frac{p(x|\theta_t)}{p(x|\theta_s)} \\ &= \exp\left[-\frac{1}{2}(\Delta\chi^2(\theta_t|x) - \Delta\chi^2(\theta_s|x))\right] \end{aligned} \tag{41}$$

value at the parameter value θ_s from which the pseudoexperiment was sampled from is the smallest among all other values in $\{\theta\}_S$ [i.e., $\omega_s(x|\theta_t) = \min_{s'} \omega_{s'}(x|\theta_t)$] and discarded otherwise. If the pseudoexperiment is used, it is weighted by $\omega_s(x|\theta_t)$. Then, by combining the pseudoexperiments sampled from all $\{\theta\}_S$ values with their weights, the desired distribution $p(x|\theta_t)$ is attained with higher probability to sample pseudoexperiments of large $\Delta\chi^2(\theta_t|x)$. Since $\theta_t \in \{\theta\}_S$, this procedure ensures that $w(x|\theta_t) \leq 1$ for well-behaved weights.

One downside of this vetoing technique, as explained by Cranmer, is that the spacing of $\{\theta\}_S$ must not be too dense in order not to reduce the efficiency of the method with a high vetoing probability. The mixture-FC method does not have this problem because the weights are computed using the actual sampling probability, which is the sum of probabilities over $\{\theta\}_S$, and no vetoing is necessary. While the claimed benefit of the vetoing technique is its independence from the exact normalization of the sampling probability distribution—due to only using the probability *ratios* ω_s—the same is true for the choice of weights in the mixture-FC method, whose weights from Eq. (2) can be written as

$$w(x|\theta_s) = \frac{1}{\frac{1}{S}\sum_{s=1}^{S}[\omega_s(x|\theta_t)]^{-1}}. \tag{42}$$

For the problem of finding p values under a null hypothesis with a likelihood-ratio statistic, the relevant part of the mixture-FC can therefore be regarded a slight improvement to the method by Ref. [12]. Furthermore, we have explicitly shown that, under suitable conditions, which for a typical setup requires the absence of nuisance parameters, the variance on the estimated p values is reduced exponentially for large values of the test statistic.

A very different method for importance sampling pseudoexperiments with small p values uses nested sampling [14]. Here, rather than sampling many pseudoexperiments and finding the fraction with more extreme values of a test statistic, this fraction is computed by employing a separate algorithm that explores the sampling space of pseudoexperiments with a sequentially increasing threshold of the test statistic. This method is very generic and can be applied for any choice of test statistic that includes, for example, goodness-of-fit tests to which the mixture-FC method cannot be applied because of requiring a likelihood-ratio test statistic between predefined hypotheses. Of course, for the particular problem of goodness-of-fit tests, one is typically not interested in very small p values. On the other hand, for hypothesis testing or inference involving likelihood-ratio test statistics, where small p values are of interest, the mixture-FC method presented in this work is easier to implement, especially for analyses that already employ the Feldman-Cousins method, but also for any analysis that just compute p values, because the computation can be performed almost entirely on exiting infrastructure. Furthermore, because the mixture-FC method provides an explicit construction of the sampling distribution, significant improvements can be seen already from relatively “large” p values, such as 2σ and 3σ. However, because of the absence of theoretically guaranteed improvements when dealing with nuisance parameters, there may be problems, especially at very small p values of 5σ and beyond, where the presented method cannot provide numerically stable estimates of the p value, in which case the method using nested sampling should be considered due to its more general applicability.

Finally, we note some of the differences of computing p values to the FC confidence interval construction in the context of importance sampling. When computing p values, we are typically interested in the distribution of the test statistic under a *single* null hypothesis. In contrast, in the FC method, we need the test-statistic distribution for *all* plausible parameter values, which in practice is achieved by computing them for a finite set $\{\theta\}_S$ and interpolating in between. The FC construction therefore benefits from the ability to interpolate critical values with importance sampling, which is not always of interest in the computation of p values. In addition, the pseudoexperiments sampled from different parameter values as required for the construction of the mixture distribution are already available even in the conventional-FC method, making the transition to the mixture-FC method straightforward.

VI. SUMMARY

We presented a new method to compute critical values for Feldman-Cousins confidence intervals. The method is a simple extension of the conventional method in that the same sets of pseudoexperiments generated at different parameter values are simply combined with suitable weights. We showed that this results in a significant reduction of the errors on the critical values, with exponential reduction for high confidence level critical values, at almost no additional computational cost. The method was further shown to enable accurate interpolation of critical values between the parameter values at which the pseudoexperiments were generated. The theoretically calculated performance was confirmed using a simple example for the analysis of neutrino oscillations. While the exponential reduction of errors is currently only guaranteed for analyses without nuisance parameters, the general technique is applicable to any analysis making use of the Feldman-Cousins method.

The code used in this paper is publicly available on GitHub [15].

ACKNOWLEDGMENTS

We would like to thank Christophe Bronner and Louis Lyons for useful discussions and connecting us to Kyle Cranmer, whom we would like to thank for introducing Refs. [11,12]. This research was supported by JSPS KAKENHI Grant No. 19J22440.

APPENDIX A: RELATIONSHIP BETWEEN THE UNION OF ENSEMBLES AND THE MIXTURE DISTRIBUTION

The union of N samples from S different distributions $\{x\}_s$ distributed according to $p(x|\theta_s)$, which we will refer to as the “union distribution” $\{x\}_{\mathrm{mix}} = \cup_{s=1}^{S} \{x\}_s$, differs from samples from the actual mixture distribution $p(x) = \frac{1}{S}\sum_{s=1}^{S} p(x|\theta_s)$ in lacking the categorical variance of choosing from which distribution to sample from. Here we show that the estimators for the mean using samples from the union distribution are consistent with those from the mixture distribution and have smaller variance.

Given any function $f(x)$ that only depends on x and not the index of the distribution s from which the samples were obtained, we can define an estimator for its mean using independent samples $x_{si} \sim p(x|\theta_s)$ from the union distribution,

$$\hat{f} = \frac{1}{S}\sum_{s=1}^{S}\frac{1}{N}\sum_{i=1}^{N} f(x_{si}). \tag{A1}$$

Its mean is given by

$$\mathbb{E}[\hat{f}] = \frac{1}{S}\sum_{s=1}^{S}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[f(x_{si})] \tag{A2}$$

$$= \frac{1}{S}\sum_{s=1}^{S}\mathbb{E}_s[f] \tag{A3}$$

$$= \frac{1}{S}\sum_{s=1}^{S}\int \mathrm{d}x\, p(x|\theta_s) f(x) \tag{A4}$$

$$= \int \mathrm{d}x\, p(x) f(x), \tag{A5}$$

where in going to the second line we used that x_{si} is identically distributed for the same s, $\mathbb{E}_s[\cdot]$ is the expectation under $x \sim p(x|\theta_s)$, and in going to the last line we exchanged the order of the finite sum and the integral. Thus, $\hat{f}$ is an unbiased estimator for the mean of f under the mixture distribution. Its variance is given by

$$\mathrm{Var}[\hat{f}] = \frac{1}{S^2}\sum_{s,s'=1}^{S}\frac{1}{N^2}\sum_{i,i'=1}^{N}\mathrm{Var}[f(x_{si}), f(x_{s'i'})] \tag{A6}$$

$$= \frac{1}{S^2}\sum_{s=1}^{S}\frac{1}{N}\mathrm{Var}_s[f] \tag{A7}$$

$$= \frac{1}{S^2}\sum_{s=1}^{S}\frac{1}{N}(\mathbb{E}_s[f^2] - \mathbb{E}_s[f]^2) \tag{A8}$$

$$= \frac{1}{SN}\left(\frac{1}{S}\sum_{s=1}^{S}\mathbb{E}_s[f^2] - \left\{\frac{1}{S}\sum_{s=1}^{S}\mathbb{E}_s[f]\right\}\left\{\frac{1}{S}\sum_{s'=1}^{S}\mathbb{E}_{s'}[f]\right\} - \frac{1}{S}\sum_{s=1}^{S}\mathbb{E}_s[f]^2 + \left\{\frac{1}{S}\sum_{s=1}^{S}\mathbb{E}_s[f]\right\}\left\{\frac{1}{S}\sum_{s'=1}^{S}\mathbb{E}_{s'}[f]\right\}\right) \tag{A9}$$

$$= \frac{1}{SN}(\mathrm{Var}[f] - \mathrm{Var}[\mu_s]), \tag{A10}$$

where in going to the second line we used independence and that x_{si} is identically distributed for the same s, in going to the fourth line we inserted the terms with curly braces with opposite signs, and in the last line $\mathrm{Var}[f]$ is the variance under the mixture distribution, whereas for the last term $\mu_s := \int \mathrm{d}x p(x|\theta_s) f(x)$ with its variance taken with respect to s as the random variable with discrete uniform probability distribution $p_s = 1/S$. We see that compared to the sample mean of SN samples from the mixture distribution, which would have variance $\mathrm{Var}[f]/SN$, the estimator of the mean from the union distribution has smaller variance, with difference equal to the variance of μ_s under the discrete uniform distribution over s.

While the probability distribution of samples from the union distribution is not equal to that of the mixture distribution, thanks to $\hat{f}$ being an unbiased estimator of the mean of f under the mixture distribution, we can insert weight functions designed for the mixture distribution to reweight to other distributions: given $w(x|\theta_t) = p(x|\theta_t)/p(x)$ and any function $g(x)$, we can take $f(x) = w(x|\theta_t)g(x)$ such that its mean is equal to the mean of g under $p(x|\theta_t)$ following

$$\mathbb{E}[\hat{f}] = \mathbb{E}[f] = \mathbb{E}[wg] = \mathbb{E}_t[g], \tag{A11}$$

despite the denominator of w being the mixture distribution and not the exact probability distribution of the union distribution.

In summary, if we have N independent samples from each subdistribution, we are not only allowed to skip the additional sampling step of choosing from which subdistribution to sample from by just using the union of the ensembles, but in doing so obtain an estimator that has lower variance. Thus, the importance sampling weights can be defined using the explicit functional form of the mixture distribution, whose properties are easier to study analytically, while allowing numerically superior results thanks to the lower variance of estimators based on the union distribution.

APPENDIX B: PERFORMANCE OF WEIGHTED HISTOGRAMS

With pseudoexperiments x sampled from some distribution and reweighted to a target distribution $p(x|\theta_t)$ using the ratio w of the target and sampling densities, the density $p(Y(x)|\theta_t)$ of some test statistic $Y(x)$ can be estimated by

$$\hat{P}_b := \frac{1}{n_{\mathrm{exp}}}\sum_{i=1}^{n_{\mathrm{exp}}} w(x_i) I(y_b \leq Y(x_i) < y_{b+1}), \tag{B1}$$

corresponding to the weighted sum of all pseudoexperiments with $Y(x)$ falling into the bin $[y_b, y_{b+1})$. Given the probability

$$\pi_b := \mathbb{E}[I(y_b \leq Y(x) < y_{b+1})] \tag{B2}$$

of a pseudoexperiment to fall into bin b under the *sampling* distribution, and the expectation and variance of weight-powers *among* pseudoexperiments falling into the bin under the sampling distribution

$$\mathbb{E}_b[w^k] := \mathbb{E}[w(x)^k | y_b \leq Y(x) < y_{b+1}] \tag{B3}$$

$$= \frac{1}{\pi_b}\mathbb{E}[w(x)^k I(y_b \leq Y(x) < y_{b+1})], \tag{B4}$$

$$\mathrm{Var}_b[w] := \mathbb{E}_b[w^2] - \mathbb{E}_b[w]^2, \tag{B5}$$

the expectation and variance of the distribution estimator is found to be [using $I(\cdot)^2 = I(\cdot)$]

$$\mathbb{E}[\hat{P}_b] = \pi_b \mathbb{E}_b[w], \tag{B6}$$

$$\mathrm{Var}[\hat{P}_b] = \frac{1}{n_{\mathrm{exp}}}(\pi_b(1-\pi_b)\mathbb{E}_b[w]^2 + \pi_b \mathrm{Var}_b[w]), \tag{B7}$$

where in Eq. (B7) the first term is the usual binomial error due to the number of pseudoexperiments falling into the bin, and the second is the additional term due to the variance of weights among pseudoexperiments falling into the bin.

APPENDIX C: ANALYSIS OF CRITICAL VALUE VARIANCES FOR GENERIC MIXTURES

Let us denote the target distribution of pseudoexperiments at θ_t by $p_t(x)$. In a setting with nuisance parameters η with probability distribution $p(x|\theta,\eta)$, this could, for example, be $p(x|\theta_t,\hat{\hat{\eta}}(\theta_t|x_{\text{obs}}))$ for the profile-FC method or $\int \mathrm{d}\eta\, \pi(\eta|x_{\text{obs}},\theta_t)p(x|\theta_t,\eta)$ in the posterior Highland-Cousins method, with $\hat{\hat{\eta}}(\theta|x_{\text{obs}}) = \arg\min_\eta \chi^2(\theta,\eta|x_{\text{obs}})$ the profile best-fit values and $\pi(\eta|x_{\text{obs}},\theta)$ the posterior distribution for nuisance parameters conditioned by the target θ value for a fit to the observed data x_{obs}. The other pseudoexperiments are sampled from $p_a(x)$, whose distribution we do not explicitly specify here, but could, for example, be a mixture over different θ and η values. The mixture of N_t pseudoexperiments sampled from $p_t(x)$ and N_a pseudoexperiments sampled from $p_a(x)$ weighted by $w(x) = (N_t+N_a)p_t(x)/[N_t p_t(x) + N_a p_a(x)]$ can be evaluated analogously to the main text and using the estimators

$$\hat{P}(y) \coloneqq \frac{1}{N_t+N_a}\sum_{i=1}^{N_t+N_a} w(x_i) I(Y(x_i) \geq y), \tag{C1}$$

$$\hat{P}_{\text{conv}}(y) \coloneqq \frac{1}{N_t}\sum_{i=1}^{N_t} I(Y(x_i^{(t)}) \geq y), \tag{C2}$$

$$\mathbb{E}[\hat{P}(y)] = \mathbb{E}[\hat{P}_{\text{conv}}(y)] = P(y) \tag{C3}$$

yield a variance reduction of

$$\gamma = \frac{\mathrm{Var}[\hat{P}(y)]}{\mathrm{Var}[\hat{P}_{\text{conv}}(y)]} \tag{C4}$$

$$\leq \frac{\frac{1}{N_t}P(y) - \frac{1}{N_t+N_a}P(y)^2}{\frac{1}{N_t}P(y) - \frac{1}{N_t}P(y)^2} \tag{C5}$$

$$\leq \frac{1}{1-P(y)}, \tag{C6}$$

where

$$x_i \sim \frac{N_t p_t(x) + N_a p_a(x)}{N_t + N_a}, \tag{C7}$$

$$x_i^{(t)} \sim p_t(x). \tag{C8}$$

[1] S. S. Wilks, The large-sample distribution of the likelihood ratio for testing composite hypotheses, Ann. Math. Stat. **9**, 60 (1938).

[2] G. J. Feldman and R. D. Cousins, A unified approach to the classical statistical analysis of small signals, Phys. Rev. D **57**, 3873 (1998).

[3] L. Li, N. Nayak, J. Bian, and P. Baldi, Efficient neutrino oscillation parameter inference using Gaussian processes, Phys. Rev. D **101**, 012001 (2020).

[4] B. Efron, Bootstrap methods: Another look at the jackknife, Ann. Stat. **7**, 1 (1979).

[5] R. G. Miller, The jackknife–a review, Biometrika **61**, 1 (1974).

[6] K. Abe *et al.* (T2K Collaboration), The T2K experiment, Nucl. Instrum. Methods Phys. Res., Sect. A **659**, 106 (2011).

[7] A. M. Ferrenberg and R. H. Swendsen, Optimized Monte Carlo analysis, Phys. Rev. Lett. **63**, 1195 (1989).

[8] H. Jeffreys, An invariant form for the prior probability in estimation problems, Proc. R. Soc. A **186**, 453 (1946).

[9] M. A. Acero *et al.* (NOvA Collaboration), The profiled Feldman-Cousins technique for confidence interval construction in the presence of nuisance parameters, arXiv:2207.14353.

[10] R. D. Cousins and V. L. Highland, Incorporating systematic uncertainties into an upper limit, Nucl. Instrum. Methods Phys. Res., Sect. A **320**, 331 (1992).

[11] M. Woodroofe, *Importance sampling and error probabilities, Banff Discovery Workshop* (2010), https://www.birs.ca/workshops/2010/10w5068/files/woodroofe.pdf.

[12] K. Cranmer, Practical statistics for the LHC, in *2011 European School of High-Energy Physics* (2014), pp. 267–308, 10.5170/CERN-2014-003.267.

[13] G. Aad *et al.* (ATLAS Collaboration), Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC, Phys. Lett. B **716**, 1 (2012).

[14] A. Fowlie, S. Hoof, and W. Handley, Nested sampling for frequentist computation: Fast estimation of small p values, Phys. Rev. Lett. **128**, 021801 (2022).

[15] L. Berns, Lukasberns/mixture-FC-paper-code (on GitHub) (2024), 10.5281/zenodo.10980364.