Model exploration in gravitational-wave astronomy with the maximum population likelihood

Ethan Payne[1,2,3,4,*] and Eric Thrane[3,4,†]

[1]*Department of Physics, California Institute of Technology, Pasadena, California 91125, USA*
[2]*LIGO Laboratory, California Institute of Technology, Pasadena, California 91125, USA*
[3]*School of Physics and Astronomy, Monash University, Melbourne, Victoria 3800, Australia*
[4]*OzGrav: The ARC Centre of Excellence for Gravitational-Wave Discovery, Clayton, Victoria 3800, Australia*

(Received 20 October 2022; accepted 21 February 2023; published 7 April 2023)

Hierarchical Bayesian inference is an essential tool for studying the population properties of compact binaries with gravitational waves. The basic premise is to infer the unknown prior distribution of binary black hole and/or neutron star parameters such component masses, spin vectors, and redshift. These distributions shed light on the fate of massive stars, how and where binaries are assembled, and the evolution of the Universe over cosmic time. Hierarchical analyses model the binary black hole population using a prior distribution conditioned on hyperparameters, which are inferred from the data. However, a misspecified model can lead to faulty astrophysical inferences. In this paper we answer the question: given some data, which prior distribution—from the set of all possible prior distributions—produces the largest possible population likelihood? This distribution (which is not a true prior) is ⵠ (pronounced “pi stroke”), and the associated *maximum population likelihood* is Ł (pronounced “L stroke”). The structure of ⵠ is a linear superposition of delta functions, a result which follows from Carathéodory’s theorem. We show how ⵠ and Ł can be used for model exploration/criticism. We apply this Ł formalism to study the population of binary black hole mergers observed in the LIGO-Virgo-KAGRA Collaboration’s third gravitational-wave transient catalog. Based on our results, we discuss possible improvements for gravitational-wave population models.

DOI: 10.1103/PhysRevResearch.5.023013

I. MOTIVATION

Bayesian inference has become a mainstay of modern scientific data analysis as a means of analyzing signals in noisy observations. This procedure determines the posterior distributions for parameters given one or more model. In order to study the *population properties* of a set of uncertain observations, a hierarchical Bayesian framework can be employed. The basic idea is to model the population using a conditional prior $\pi(\theta|\Lambda, M)$, which describes, for example, the distribution of black hole masses $\{m_1, m_2\} \in \theta$ given some hyperparameters Λ, which determine the shape of the prior distribution. Here, M denotes the choice of model. One then carries out Bayesian inference using a “population likelihood”

$$\mathcal{L}(d|\Lambda, M) = \prod_i^N \frac{1}{\xi(\Lambda)} \int d\theta_i \, \mathcal{L}(d_i|\theta_i)\pi(\theta_i|\Lambda, M), \qquad (1)$$

where $\mathcal{L}(d_i|\theta_i)$ is the likelihood for data associated with event i given parameters θ_i, and $\xi(\Lambda)$ is the detected fraction for a choice of hyperparameters. Meanwhile, N is the total number of observations. For an overview of hierarchical modeling in gravitational-wave astronomy including selection effects, see Refs. [1–3].

The LIGO-Virgo-KAGRA (LVK) Collaboration’s third gravitational-wave transient catalog (GWTC-3) [4] contains the cumulative set of observations of $N = 69$ confident binary black-hole mergers [5] detected by the LVK [6–8]. Additional detection candidates have been put forward by independent groups [9–13]. Hierarchical inference is employed to study the population properties these merging binary black holes; see, e.g., Refs. [14–32]. These analyses have revealed a number of exciting results, such as the surprising excess rate of mergers with a primary black hole mass of $\sim 35 M_\odot$ [15], and the evolution of the binary merger rate with redshift [16], to name just two.

However, Bayesian inference has its limitations. One can use Eq. (1) in order to infer the distribution of binary black hole parameters, *given some model*; and one can compare the marginal likelihoods of two models to see which one better describes the data. However, Bayesian inference does not tell us if any of the models we are using are suitable descriptions of the data. While all models for the distribution of binary black hole parameters are likely to be imperfect, some may be adequate for describing our current dataset [33]. When a model fails to capture some salient feature of the data, it is said to be “misspecified” [34,35]. Some effort has been made to assess the suitability of gravitational-wave models, both qualitatively and quantitatively; see, e.g., [15,16,34,36].

[*] epayne@caltech.edu
[†] eric.thrane@monash.edu

However, the idea of "model criticism"—testing the suitability of Bayesian models—is still being developed within the context of gravitational-wave astronomy and beyond.

Hierarchical Bayesian inference studies often depend upon parametric models. Modelers design parametrizations in order to capture the key features of the astrophysical distributions. However, one must still worry about "unknown unknowns": features which do not occur to the modeler to add. For example, recent studies [15,16,37,38] find that a sub-population of binary black holes merge with spin vectors that are misaligned with respect to the orbital angular momentum axis. However, the degree to which the spins are misaligned might be model dependent. In Refs. [15,16,37], the inferred minimum spin tilt is confidently $\gtrsim 90^\circ$. In contrast, Refs. [17,28,38] argue this signature could be due to a lack of flexibility in LVK models to account for a subpopulation of black holes with negligible spin magnitude, finding support for misalignment at smaller minimum tilt angles. The inferred population distribution of spin misalignment has important consequences for understanding the formation channels of binary black-hole channels. This debate highlights how astrophysical inferences can be affected by model design.

In order to help alleviate some of the issues arising from model misspecification in Bayesian inference, we present a framework for assessing the suitability of a model. This framework is built around the concept of the *maximum population likelihood* $\pounds$ (pronounced "L stroke"): the largest possible value of $\mathcal{L}(d|\Lambda)$ in Eq. (1), maximized over all possible choices of population model $\pi(\theta|\Lambda)$ *independent of the choice of parametrization*. The "prior" distribution, which yields this maximum is $\not\pi(\theta)$ (pronounced "pi stroke"). It is not a true prior because it is determined by the data. The theory behind the maximization of population likelihoods has been studied previously in optimization and statistics literature [39–44]. This work is underpinned by Carathéodory's theorem [45] and the mathematics of convex hulls [43]. However, its application to observational science has been somewhat limited as far as we can tell.

The $\pounds$ framework is useful for several reasons. First, the numerical value of $\pounds$ is an upper bound on the population likelihood. We can compare the maximum likelihood for a specific model,

$$\mathcal{L}_{\max}(M) = \max_{\Lambda\sim p(\Lambda|d)} \mathcal{L}(d|\Lambda, M), \tag{2}$$

to $\pounds$. Often in Bayesian model selection, the Bayesian evidence values ($\mathcal{Z}_i$) of two hypotheses can be used to determine the extent to which one model is preferred over the other. A typical threshold chosen to rule out one model in favor of another is that $\ln(\mathcal{Z}_1/\mathcal{Z}_2) > 8$ [46]. In a similar vein, if $\ln(\pounds/\mathcal{L}_{\max}(M)) \lesssim 8$, we can be sure the model M is not badly misspecified since there is no second model M' that can be written down with that will yield a statistically significant improvement. We emphasize that a model which does not satisfy this condition is not necessarily misspecified.

Second, the $\pounds$ framework can be used to quantitatively assess if a model M is misspecified. By generating synthetic data from M, one can generate the expected distribution of $(\pounds, \mathcal{L}_{\max}(M))$. In this paper, we show how one can compare the observed values of $(\pounds, \mathcal{L}_{\max}(M))$ to the expected distribution in order to determine the extent to which M is misspecified, and the *way* in which it is misspecified.

Third, the $\pounds$ framework can be used for "model exploration," providing clues of *where* in parameter space unmodeled features might be lurking. By comparing $\not\pi(\theta)$ with the prior from our phenomenological model $\pi(\theta|M)$, one can see if the phenomenological model is capturing key structure present in $\not\pi$ and use the comparison to design new models to test on forthcoming datasets.

The remainder of this paper is organized as follows. In Sec. II, we introduce the $\pounds$ formalism, illustrating key features with a simple toy model. In Sec. III, we show how the formalism can be used for model criticism. In Sec. IV, we apply the formalism to study the population properties of merging binary black holes observed by the LVK. Our concluding remarks are presented in Sec. V.

II. THE MAXIMUM POPULATION LIKELIHOOD $\pounds$

A. Preliminaries

We begin with a brief review of Bayesian hierarchical inference with a parametric model. Our starting point is the population likelihood [copied here from Eq. (1):

$$\mathcal{L}(d|\Lambda, M) = \prod_i^N \frac{1}{\xi(\Lambda)} \int d\theta_i\, \mathcal{L}(d_i|\theta_i)\pi(\theta_i|\Lambda, M). \tag{3}$$

Here, $\mathcal{L}(d_i|\theta_i)$ is the likelihood of event-i data d_i given parameters θ_i. The quantity $\pi(\theta_i|\Lambda, M)$ is a conditional prior for θ_i given hyperparameters for some population model M, which describes the shape of the prior distribution. The term $\xi(\Lambda)$ accounts for selection effects; for example, high-mass systems are typically easier to detect than low-mass systems. It is the detectable fraction of the population given the model given hyperparameters Λ:

$$\xi(\Lambda) = \int d\theta\, p_{\text{det}}(\theta)\pi(\theta|\Lambda, M). \tag{4}$$

Here, $p_{\text{det}}(\theta)$ is the detection probability of an observation with parameters θ.

B. The maximum population likelihood $\pounds$

The maximum population likelihood $\pounds$ is obtained by taking Eq. (3) and maximizing over all possible prior distributions $\pi(\theta)$. Thus, $\pounds$ is an upper bound (or supremum) on the set of likelihoods from all possible choices of models for $\pi(\theta)$ such that

$$\pounds \equiv \mathcal{L}(d|\not M) \geqslant \mathcal{L}(d|\Lambda, M), \tag{5}$$

for all models M. The "prior" distribution that yields $\pounds$ is denoted

$$\not\pi(\theta) \tag{6}$$

(pronounced "pi stroke"). It is not a true prior because the distribution which maximizes the population likelihood in Eq. (3) depends on the data. One should therefore refer to $\not\pi$ as a pseudoprior. The associated model is denoted $\not M$ (pronounced "M stroke"). Combining this notation into a single equation, we have

$$\pounds \equiv \prod_{i=1}^{N} \frac{1}{\xi(\not{M})} \int d\theta_i \, \mathcal{L}(d_i|\theta_i) \not\pi(\theta_i). \tag{7}$$

C. Calculating $\not\pi$: special cases

Having introduced the concept of $\pounds$ and $\not\pi$, the natural next question is, given data d, how does one calculate these quantities? Before answering this question, we study three special cases where we can work out $\not\pi$ from intuition. This discussion will help sharpen our instincts for the more general solution that follows. Readers looking to skip to the answer may wish to skip this subsection.

1. A single measurement

For the first case, we consider a single measurement ($N = 1$) with a unimodal likelihood function $\mathcal{L}(d|\theta)$, which is maximal when the parameter θ is equal to the maximum likelihood value $\widehat{\theta}$. For the sake of simplicity, we ignore selection effects so that $\xi(\not{M}) = 1$. In this case, $\pounds$ in Eq. (7) is clearly maximized if the prior support is entirely concentrated at $\widehat{\theta}$. Thus, $\not\pi$ is a delta function,

$$\not\pi(\theta) = \delta(\theta - \widehat{\theta}), \tag{8}$$

which yields

$$\begin{aligned} \pounds &= \int d\theta \, \mathcal{L}(d|\theta) \, \delta(\theta - \widehat{\theta}) \\ &= \mathcal{L}(d|\widehat{\theta}). \end{aligned} \tag{9}$$

This result is intuitive: the prior that maximizes the population likelihood is the one that concentrates all its support at the maximum-likelihood value of θ.

2. N signals in the high-SNR limit

For the second case, we consider a scenario in which the data consists of N observations carried out in the limit of high signal-to-noise ratio. In this limit, the likelihood of the data for each measurement d_i given some parameter θ approaches a delta function,

$$\mathcal{L}(d_i|\theta_i) = \delta(\theta_i - \widehat{\theta}_i), \tag{10}$$

located at the maximum-likelihood value $\widehat{\theta}_i$. We assume that each measurement is distinct so that no two maximum-likelihood values $\widehat{\theta}_i$ are exactly the same. Again, for the sake of simplicity, we ignore selection effects so that $\xi(\not{M}) = 1$, though, the argument here holds even if we relax this assumption. Equation (7) becomes

$$\pounds = \prod_{i=1}^{N} \int d\theta_i \, \delta(\theta_i - \widehat{\theta}_i) \not\pi(\theta_i). \tag{11}$$

The population likelihood is maximized when $\not\pi$ is a sum of delta functions peaking at the set of $\{\widehat{\theta}_i\}$:

$$\not\pi(\theta) = \sum_{k=1}^{N} w_k \, \delta(\theta - \widehat{\theta}_k) \tag{12}$$

$$w_k = 1/N. \tag{13}$$

This solution for $\not\pi$ ensures that there is maximal prior support at every likelihood peak. Obviously, the population likelihood is not maximized if any prior probability density is wasted to values of θ where all the likelihood functions are zero. Choosing an equal weight for each delta function $w_i = 1/N$ produces the largest possible population likelihood [47].

We illustrate this case in Fig. 1(a) using high-SNR, toy-model data drawn from a mean-zero, unit-variance Gaussian distribution. In the top panel, we plot the set of $N = 10$ maximum likelihood points $\{\widehat{\theta}_i\}$ and the position of the delta functions (blue). In the lower panel, we "plot" the $\not\pi(\theta)$ for these ten data points. We put the word "plot" in quotation marks because, technically, we are not plotting $\not\pi(\theta)$, which goes to infinity, but rather we are plotting the weights w_k [Eq. (13)], which allows us to see the relative weight given to each delta function, something that will prove useful below. Throughout the paper, when we refer to plots of $\not\pi(\theta)$, it should be understood that we are actually plotting *representations* of $\not\pi(\theta)$ using the weights w_k. Finally, note that each peak in the distribution of $\not\pi(\theta)$ matches up with one of the maximum likelihood points in the upper panel.

3. N identical measurements

For the third case, we consider a set of N observations. This time, we do not assume the high-SNR limit, but we assume that every measurement has the same maximum-likelihood value of $\widehat{\theta}$. This case is highly contrived—one does not typically work with multiple identical measurements—but the example is nonetheless helpful for illustrative purposes. In this case, the integral in Eq. (7) is maximized when the prior support is entirely concentrated at $\widehat{\theta}$ (where all of the likelihood functions peak), so that $\not\pi$ is a single delta function:

$$\not\pi(\theta) = \delta(\theta - \widehat{\theta}), \tag{14}$$

while

$$\pounds = \prod_{i=1}^{N} \mathcal{L}(d_i|\widehat{\theta}). \tag{15}$$

This scenario is demonstrated in Fig. 1(b). The top panel shows the set of $N = 10$ maximum-likelihood points $\{\widehat{\theta}_i\}$, all with the same value. The horizontal lines represent the error bars for each measurement, which we draw from a uniform distribution on the interval (0.01,1). In the lower panel, we plot $\not\pi(\theta)$ for these ten data points. This time, since every measurement is identical, $\not\pi(\theta)$ is a single delta function peaking at $\theta = 0$.

From these three examples, we observe a pattern: in each case, $\not\pi(\theta)$ can be written as a weighted sum of delta functions. Indeed, it has been proven that this is in fact the case [39–44]. We refer readers interested in an explanation of the delta function structure of $\not\pi$ to the Appendix, where we summarize the key concepts surrounding the proof outlined in Ref. [43] using the mathematics of convex hulls. We do not reproduce the proof in its entirety, but rather we use visualisations to explain how it works with $N = 2$ observations, before providing a qualitative explanation for how it generalizes to arbitrary values of N. We explore this general structure and the consequences thereof in the next subsection.

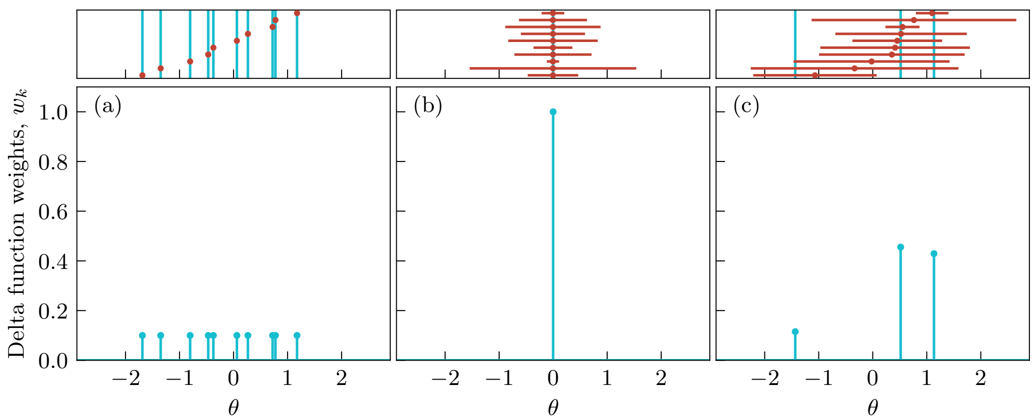

FIG. 1. Examples of the distribution $\not{\pi}(\theta)$ described in Secs. II B–II D. Each column represents a different dataset. The top-panel dots show the set of $N = 10$ maximum-likelihood estimates $\{\widehat{\theta}_i\}$. The top-panel horizontal lines represent error bars (in the first column they are too small to see), and the vertical lines (blue) indicate the inferred delta function locations. The bottom panels show the distribution of $\not{\pi}(\theta)$ associated with each data set. The left-hand column (a) represents data in the high-SNR limit so that the likelihood functions for each measurement approach delta functions (this is why the error bars are not visible). In this case, $\not{\pi}(\theta)$ consists of N delta functions, each associated with one of the maximum likelihood points $\widehat{\theta}_i$. In the middle column (b), we are no longer in the high-SNR limit, but the maximum likelihood points are all assumed to be identical with $\widehat{\theta}_i = 0$. In this case, $\not{\pi}(\theta)$ consists of one delta function peaking at $\theta = 0$. In the right-hand column (c), the data are not in the high-SNR limit, and each $\widehat{\theta}_i$ is random. In this case, $\not{\pi}(\theta)$ consists of $n = 3$ delta functions, each with different heights.

D. The general form of $\not{\pi}$

We proceed with the knowledge that Eq. (7) is true in general, regardless of the form of the likelihood $\mathcal{L}(d|\theta)$ and the selection effect term $p_{\rm det}(\theta)$. *For any set of observations, $\not{\pi}(\theta)$ is always of the form*

$$\not{\pi}(\theta) = \sum_{k=1}^{n} w_k\, \delta(\theta - \theta_k), \tag{16}$$

where w_k are weights which sum to unity,

$$\sum_{k=1}^{n} w_k = 1. \tag{17}$$

The number of delta functions is always less than or equal to the number of measurements, and the solution is unique in all but the most pathological of cases (e.g., multimodal distributions with regions of equivalent maximum likelihoods) so that

$$n \leqslant N. \tag{18}$$

The ratio

$$\mathcal{I} \equiv n/N \tag{19}$$

is a measure of the "informativeness" of the data. It compares the typical likelihood width to the scatter in the astrophysical distribution. In the high-SNR limit, $\mathcal{I} = 1$, since a delta function is required for every data point [see Fig. 1(a)]. The other limiting case is $\mathcal{I} = 1/N$, which happens when the likelihood for each measurement completely overlaps [see Fig. 1(b)].

Using this insight into the structure of $\not{\pi}(\theta)$, we now consider a variation on the toy-model problems discussed in the earlier subsections. In particular, we consider finite-SNR data drawn from our Gaussian, toy-model distribution. Using Eqs. (16) and (17) as an ansatz, we calculate $\not{\pi}(\theta)$ for $N = 10$ random data points. The maximum likelihood values $\widehat{\theta}_i$ are drawn from a mean-zero, unit-variance Gaussian and the error bars are drawn from a uniform distribution on the interval $(0.01, 1)$. The results of this calculation are shown in Fig. 1(c). The top panel shows the data, represented by the maximum-likelihood values $\{\widehat{\theta}_i\}$, which are arranged from bottom to top in increasing order. The horizontal lines show the uncertainty for each measurement and the vertical blue lines indicate the positions of the delta functions. In the bottom panel, we show $\not{\pi}(\theta)$ for this dataset. It consists of just $n = 3$ delta functions of varying heights ($\mathcal{I} = 0.3$). The exact weights, locations, and number of delta functions are not obvious; we obtain them numerically by maximizing Eq. (16) subject to Eq. (17) using the "combined" method described below in Sec. II E. Comparing the red data points with error bars to the turquoise representation of $\not{\pi}(\theta)$, one can see that every data point can be plausibly associated with at least one of the delta functions.

Given the form of $\not{\pi}(\theta)$ described by Eq. (16), we can write down a general expression for $\not{\mathcal{L}}$:

$$\not{\mathcal{L}} = \prod_{i=1}^{N} \frac{1}{\xi(\not{M})} \sum_{k=1}^{n} w_k\, \mathcal{L}(d_i|\theta_k), \tag{20}$$

where

$$\xi(\not{M}) = \sum_{k=1}^{n} w_k\, p_{\rm det}(\theta_k). \tag{21}$$

Given Eqs. (20) and (21), the problem of calculating $\not{\mathcal{L}}, \not{\pi}$ reduces to the problem of simply finding the locations and weights of n delta functions. In Sec. II E, we explore three different approaches to this problem.

E. Computing $\not{\pi}$

In this subsection, we consider three techniques that can be applied to compute $\not{\mathcal{L}}, \not{\pi}$: optimization, iterative grid, and stochastic methods. We show that a combined approach, which uses a grid-based approach to guess a solution that is subsequently refined through optimization, performs the best out of the algorithms we tried. Meanwhile, the stochastic approach allows us to illustrate the existence of the delta function structure proven in Ref. [43], but with minimal assumptions.

1. Optimization

The first approach we consider is to use an optimization algorithm subject to the constraint in Eq. (17) [48]. We use SCIPY's `trust-constr` optimization implementation [49,50]. We find this approach fails to find the correct global maximum of Eq. (20) once the number of peaks n becomes large. However, this issue can be resolved if a sufficiently close guess to the true shape of $\not{\pi}(\theta)$ can be made. Fortunately, the iterative-grid approach can be used to supply this initial guess.

2. Iterative grid

The second approach we consider is to iteratively place delta functions on a fixed grid. There are two steps: the greedy addition of many delta functions, and the removal of no-longer-useful delta functions. In the first step, we first attempt to place a delta function with a fixed height at each grid point and evaluate Eq. (20) (with appropriate normalization of the distribution). We determine which of all possible delta function additions produces the highest population likelihood. We then vary the height of this delta function between zero and twice the initial height in order to obtain an updated guess for $\not{\pi}(\theta)$. The addition of delta functions is repeated, reducing the initial height by a factor at each iteration. After many iterations, we then attempt to remove no-longer-useful delta functions to further increase the population likelihood. We repeat this procedure five times, iteratively adding 30 delta functions with varying heights at each iteration. After these iterations, $\pounds$ is usually well-converged for the problems we are studying. In some iterations, this procedure adds support to preexisting delta functions. This is how the approach "corrects" under-supported delta functions.

This method has a significant advantage over generic constrained optimization techniques as the procedure does not require the optimization of individual parameters governing the delta functions through the $\{\theta_k, w_k\}$ space. However, we find that this method is improved by pairing it with optimization. The most accurate optimization of the maximum population likelihood and structure of the distribution occurs when we utilize grid-based approximation to inform the starting location and weights for the constrained optimization. This allows for the grid-based approximation to find the region of parameter space where $\pounds$ is nearly maximal. The constrained optimization then purifies the delta function structure and slightly increases the maximum population likelihood. The *combined* method is used for all the maximum population likelihood computations in Sec. IV.

3. Stochastic construction

Our final approach is to stochastically generate samples for $\not{\pi}(\theta)$, which are accepted/rejected depending on whether the new samples increases the population likelihood. This is a form of importance sampling in which an arbitrary "proposal distribution" is used to generate proposal samples. When a proposal sample is generated, we add it to a list of previously accepted points and evaluate $\pounds$ as a Monte Carlo integral,

$$\pounds = \prod_{i=1}^{N} \frac{1}{\xi(\not{M})} \langle \mathcal{L}(d_i|\theta_i) \rangle_{\theta_i \sim \not{\pi}(\theta_i)}, \tag{22}$$

where

$$\xi(\not{M}) = \langle p_{\rm det}(\theta) \rangle_{\theta \sim \not{\pi}(\theta)}. \tag{23}$$

Here, the angle brackets indicate averaging over the samples. If the addition of the new sample increases $\pounds$, we retain the sample in the list of samples from $\not{\pi}$. As the process is repeated, the set of samples produces an ever-improving representation of $\not{\pi}$.

This method can be extended to employ an additional burn-in phase and/or a thinning phase to ensure more rapid convergence by removing unfavorable samples that sometimes get accepted early on before the distribution is well-converged. While this approach converges more slowly than the other two methods, *it does not employ any assumptions about the structure of the distribution*. Thus, this method can be used to validate the structure put forward in Eqs. (20) and (21), that $\not{\pi}(\theta)$ is a sum of delta functions.

4. Numerical study

We demonstrate each method using our Gaussian, toy-model distribution described in the last subsection: true maximum likelihood values $\widehat{\theta_i}$ drawn from a zero-mean, unit-variance Gaussian with error bars drawn from a uniform distribution on the interval (0.01, 1). The observed maximum likelihood values are then shifted from the true value by an offset generated from each individual observation's uncertainty. The results of this demonstration are compiled in Fig. 2. The three panels of Fig. 2 represent tests performed with $N = 10$, 100, and 1000 observations. In each panel, the black curve represents the true distribution $\pi(\theta)$. The colored spikes illustrate different numerical solutions for $\not{\pi}(\theta)$: cyan is the "combined" approach, which uses the iterative grid to obtain an initial guess that is subsequently refined using the optimization method. Meanwhile, orange represents the iterative grid approach by itself. For the grid-based approach we run 30 iterations of adding peaks with variable but decreasing weights, before repeating this process an additional ten times. Finally, gray represents the stochastic approach. For the stochastic method, we generate 3000 samples with 1000 samples for burn-in.

We see that the combined approach better estimates $\pounds$ relative to the other techniques considered [51]. We observe that, as N increases, $\not{\pi}(\theta)$ increasingly resembles the true Gaussian distribution $\pi(\theta)$ (shown in Fig. 2 as a black curve). To illustrate this more clearly, we take the inferred delta function locations from the $N = 1000$ "combined" result in Fig. 2(c) and compute the weighted histogram. This result is directly compared to the true distribution in Fig. 3, from which we see that indeed the inferred distribution is (albeit slowly) approaching the true distribution. *We conjecture that, in general,* $\not{\pi}(\theta)$ *approaches the true distribution in the infinite-data limit:*

$$\lim_{N\to\infty} \not{\pi}(\theta) \to \pi_{\rm true}(\theta). \tag{24}$$

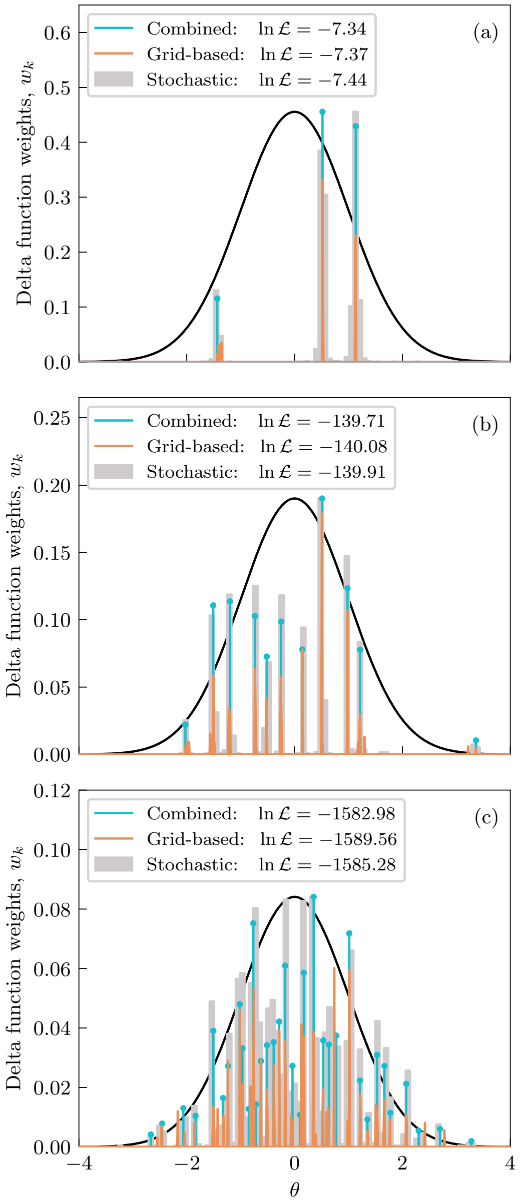

FIG. 2. Demonstration of different methods for calculating $\not\pi$, $\pounds$. Each panel shows the results for a different number of measurements with (a) $N = 10$, (b) $N = 100$, and (c) $N = 1000$. The black distribution is the true distribution $\pi(\theta)$ used to generate the data. The colored spikes show the reconstructed distribution $\not\pi(\theta)$ as determined by different methods. Cyan is for the "combined" technique, which uses the iterative grid to obtain a first guess that is refined with the optimization method. Meanwhile, orange is for the grid-based technique by itself and gray is for the stochastic method.

5. *Computational challenges*

Before continuing, we discuss two computational challenges. First, we note that the examples illustrative above are all one-dimensional. The discussion above generalizes to $\geqslant 2$ dimensions; $\not\pi(\theta)$ is still a sum of delta functions in $\geqslant 2$ dimensions. However, it becomes increasingly challenging to determine the location and height of these peaks in higher dimensions. Furthermore, by increasing the dimensionality of the problem, constructing continuous representations of the individual-event likelihoods and the detection probability $p_{\rm det}(\theta)$ becomes increasing difficult. Recent developments in using Gaussian mixture models to produce continuous representations of these distributions might alleviate these concerns [52,53]. Second, even if we stay in one dimension, the computational cost of calculating $\not\pi$, $\pounds$ grows with N [54].

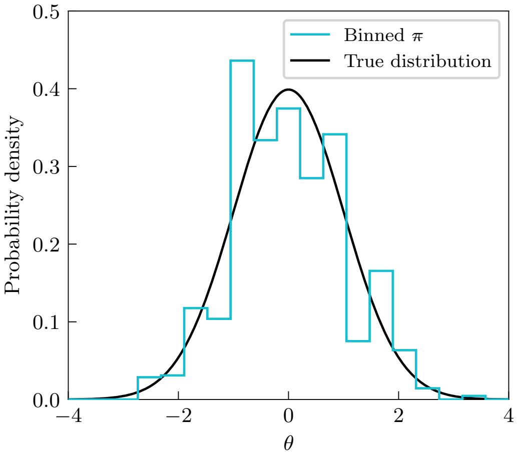

FIG. 3. Comparison between a binned representation of $\not\pi$ as computed for the toy model data set with $N = 1000$ observations and the true underlying population distribution. This representation more clearly shows that $\not\pi$ is approaching the true distribution in the limit of many observations.

III. MODEL CRITICISM WITH $\pounds$

In this section, we show how the $\pounds$ formalism can be used to determine if a model M is an adequate description of data. The first step is to generate synthetic datasets based on the posterior distribution for the model hyperparameters $p(\Lambda|d)$. For each data set, we calculate the maximum population likelihood $\pounds$ [Eq. (7)] as well as the maximum likelihood for M, which we denote

$$\mathcal{L}_{\rm max}(M) = \max_{\Lambda\sim p(\Lambda|d)} \mathcal{L}(d|\Lambda, M), \tag{25}$$

where $\mathcal{L}(d|\Lambda, M)$ is the population likelihood defined in Eq. (3). In this way we can estimate

$$p(\pounds, \mathcal{L}_{\rm max}(M)), \tag{26}$$

the joint distribution for $\pounds$ and $\mathcal{L}_{\rm max}(M)$ given model M. By comparing the *measured* values of $(\pounds, \mathcal{L}_{\rm max}(M))$ to this distribution of *expected* values, one can see if the dataset is typical of what one would expect given M. If the measured values of $(\pounds, \mathcal{L}_{\rm max}(M))$ are atypical, one can conclude that M is misspecified. Moreover, one may determine the nature of the misspecification by noting the location of the observed value of $(\pounds, \mathcal{L}_{\rm max}(M))$ relative to the typical values of $(\pounds, \mathcal{L}_{\rm max}(M))$. This is best illustrated with an example.

In our example, we imagine that an observer measures $N = 100$ values of some parameter θ. Their model M for the distribution of θ consists of a Gaussian distribution with mean $\mu = 0$ and width $\sigma = 1$:

$$\pi(\theta|M) \sim \mathcal{N}(\mu = 0, \sigma = 1). \tag{27}$$

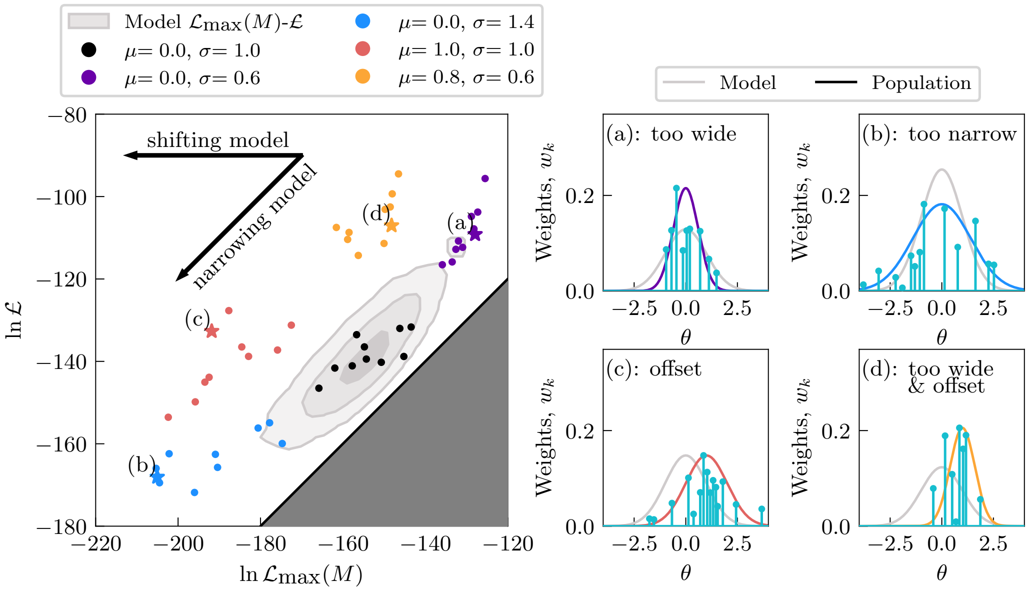

FIG. 4. An illustration of model criticism with the $\pounds$ formalism. In the left-hand panel, we plot ($\pounds$, $\mathcal{L}_{\max}(M)$) for five different underlying populations (each with ten different realizations), analyzing a toy-model with a mean of $\mu = 0$ and standard deviation $\sigma = 1$. Each population is represented by a different color. The gray contours show the one-, two-, and three-sigma credible intervals for the expected distribution of $p(\pounds, \mathcal{L}_{\max}(M))$ from the toy model. By comparing the measured values of ($\pounds$, $\mathcal{L}_{\max}(M)$) from an observed population to the expected distribution from our choice of model, one may determine if the dataset is typical of what one would expect given the model. If the measured values of ($\pounds$, $\mathcal{L}_{\max}(M)$) fall outside these intervals, one may conclude that the toy model is misspecified (does not accurately model the data). Moreover, the location of a point on this plot relative to the expected distribution conveys information about the way in which a model is misspecified. The right-hand panel shows the toy model (grey), the true population distribution for the starred and labeled data point (a)–(d), and the respective $\not\pi$ for the observed data (turquoise). This demonstrates that shifts away from the expected distribution (left-hand panel; grey) in ($\pounds$, $\mathcal{L}_{\max}(M)$) can be visually identifiable to the reconstruction of $\not\pi$.

However, their model may be misspecified so that θ is not really distributed according to M. We consider five "possible worlds" [55], one in which the observer's model is correctly specified and four in which it is not. Each world is assigned a color:

Black: model is correctly specified ($\mu = 0, \sigma = 1$).

Purple: model is too wide because the true distribution is ($\mu = 0, \sigma = 0.6$).

Blue: model is too narrow because the true distribution is ($\mu = 0, \sigma = 1.4$).

Salmon: model is shifted to one side because the true distribution is ($\mu = 1, \sigma = 1$).

Yellow: model is too wide *and* shifted to one side because the true distribution is ($\mu = 0.8, \sigma = 0.6$).

We create ten mock datasets for each of the five possible worlds (black, purple, blue, salmon, and yellow) and 5000 mock datasets from the model M (grey contours). For each dataset, we compute ($\pounds$, $\mathcal{L}_{\max}(M)$), always using model M [Eq. (27)] even if the data are generated according to, say, the blue-world distribution. This is because we are studying the case where our observer might apply a misspecified model.

The results are shown in Fig. 4. The vertical axis is $\ln \pounds$ while the horizontal axis is $\ln \mathcal{L}_{\max}(M)$. The dark-grey region in the bottom-right corner is forbidden since $\pounds \geqslant \mathcal{L}_{\max}(M)$ by construction. The grey contours show the one-, two-, and three-sigma contours for the expected distribution from the model. Only the black world datasets are consistent with the expected distribution, as the model is correctly specified in the black world. The colored dots, meanwhile, show ten random realizations of ($\ln \pounds$, $\ln \mathcal{L}_{\max}(M)$) in colored worlds where the model is misspecified in various ways. This is fundamentally different from a typical Bayesian inference plot where the data are fixed and the model is varied. Here, the model is fixed to M (Eq. 27), and we consider different datasets, which may or may not be misspecified depending on the world of our observer.

When the model M is sufficiently misspecified with respect to the true distribution, it becomes unlikely for our observer to obtain values of ($\pounds$, $\mathcal{L}_{\max}(M)$) that reside within the expected three-sigma interval, a sign of misspecification. Interestingly, the different colored dots cluster in different regions. For example, in the world where the model M is too broad (purple), the dots cluster above right of the gray contours. In the world where the model M is shifted away from the true peak (salmon), the dots cluster to the left of the gray contours. By studying *where* one's observed values of ($\pounds$, $\mathcal{L}_{\max}(M)$) fall on this diagram, one can gain some insight into the way in which one's model is misspecified. This example focuses on relatively simple forms of misspecification involving the mean and variance. Other forms of misspecification (e.g., involving skewness and kurtosis) are, of course possible as well. Given all the ways that a model can be misspecified, the "shifting model" and "narrowing model" arrows on Fig. 4 should be taken as rule-of-thumb signposts.

In practice, it is computationally challenging to create plots like Fig. 4 for population studies in gravitational-wave astronomy. While it is easy to create mock datasets, it is time consuming to calculate individual-event likelihoods for one dataset, let alone thousands. There may be workarounds. We discuss this possibility in greater detail below.

IV. APPLICATION TO GRAVITATIONAL-WAVE ASTRONOMY

In this section, we apply the $\not{\mathcal{L}}$ formalism to results from gravitational-wave astronomy to stress test models for the population of merging binary black holes. We analyze data from the second gravitational-wave transient catalog (GWTC-3) [4,56], which includes 69 confidently detected binary black hole mergers with false alarm rates $<1\ \mathrm{yr}^{-1}$. To ensure similarity to the GWTC-3 LVK population analysis [16,57], we utilize the same individual-event posterior samples, constructed from equally weighted samples generated from effective-one-body (SEOBNRv3 [58,59], SEOBNRv4PHM [60,61]) and phenomenological (IMRPhenomPv2 [62], IMRPhenomXPHM [63]) waveform results (see [16] for more details). To construct the lower-dimensional individual-event likelihoods, we utilize the same samples while marginalizing over all other "nuisance" parameters. For these nuisance parameters, we chose the distributions associated with the *maximum a posteriori* hyperparameters from the LVK's GWTC-3 population analysis with the Power Law + Peak–Default–Power Law model [16].

We divide out the sampling prior to convert the one-dimensional posterior to a likelihood. The likelihood normalization is computed using the Bayesian evidence of each event. The normalization is not important for the calculation of $\not{\pi}$, but it affects the misspecification tests demonstrated in Sec. IV C. We calculate the hyperparmeter distributions and $\mathcal{L}_{\max}(M)$ using GWPopulation [64], which employs Bilby [65,66] and Dynesty [67]. We utilize the combined injection set from Ref. [68] to compute the estimated detectable fraction of binary black-hole mergers over the first three observing runs.

A. Model inspiration through visual inspection

One straightforward application of the $\not{\mathcal{L}}$ formalism is to visually compare the reconstructed population distribution (obtained using a phenomenological model) with $\not{\pi}(\theta)$. By comparing these two distributions, it is possible to see which features in the phenomenological model reconstruction are due to prior assumptions, which features are due to real trends in the data, and which features might be missing from the phenomenological model. Formally, we compare $\not{\pi}(\theta)$ to the population predictive distribution (PPD)

$$\mathrm{PPD}(\theta|d,M)=\int d\Lambda\, p(\Lambda|d)\pi(\theta|\Lambda,M), \qquad (28)$$

which describes the astrophysical distribution of θ given a phenomenological model M with hyperparameters Λ.

In Fig. 5, we present $\not{\pi}(\theta)$ with the PPDs from the LVK analysis of GWTC-3 [16,57] for source-frame primary mass m_1 (top), the effective inspiral spin parameter χ_{eff} (middle), and redshift z (bottom). Each row contains two subpanels; the small upper panel shows the maximum-likelihood estimate for each gravitational-wave event and the 90% confidence interval while the larger lower panel compares $\not{\pi}$ with the PPD. The PPD is plotted as a thick band to show the 90% credibility region at each value of θ.

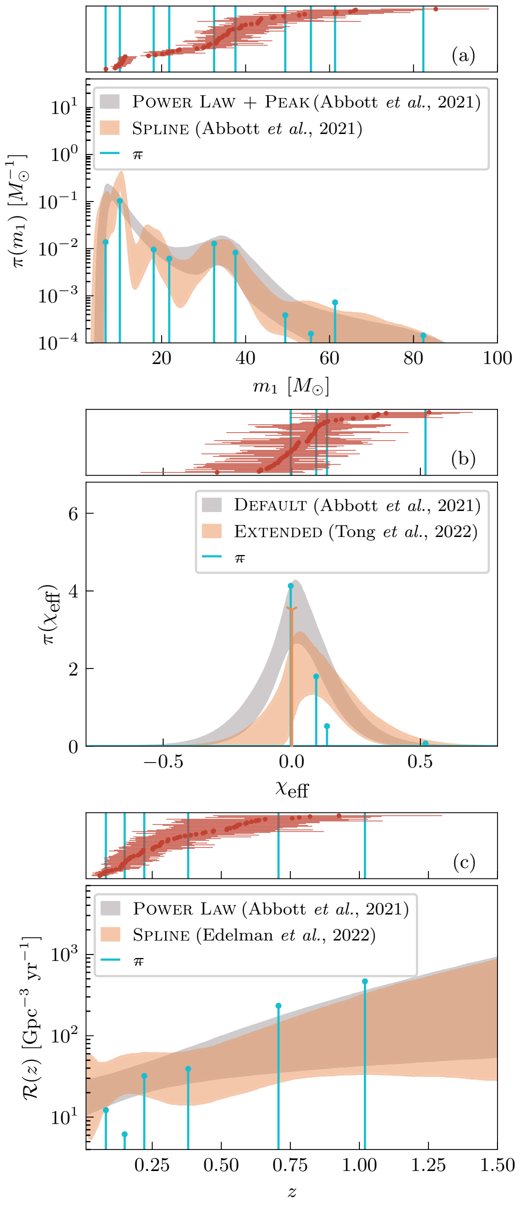

FIG. 5. Population predictive distributions (90% credibility) and $\not{\pi}$ for (a) the primary black-hole mass (m_1), (b) effective inspiral parameter (χ_{eff}), and (c) redshift (z) distributions. For the redshift, we divide by the evolution of the comoving volume and time delay as a function of redshift to plot the merger rate, $\mathcal{R}(z)$. Comparison of the different models with $\not{\pi}$ highlights which features are present in the data and which are due to assumptions in the model.

We first turn our attention to the primary mass distribution in the top row. There are $M=10$ delta function peaks, implying an informativeness of $\mathcal{I}=0.15$ [see Eq. (19)]. This result is computed in 169.3 seconds. The gray band is the Power Law + Peak model from [19] while the orange band is a (more flexible) semiparametric power-law-spline model denoted Spline from [30]. The agreement between $\not{\pi}$ and the two PPDs is striking, with cyan spikes closely matching several of the features in both models including the turnover at low masses near $\approx 12M_\odot$ and the bump at $30M_\odot$. Furthermore, we see that $\not\pi$ also recovers some of the finer detail features found only by the SPLINE model. In particular, the shift in the low-mass peak and the dips in posterior support at $\sim 16M_\odot$ and $\sim 25M_\odot$ are present in the structure of $\not\pi$. Based on our visual inspection, it appears that current models are capturing much if not all of the structure present in $\not\pi$.

Turning our attention to the middle row, we study the distribution of effective inspiral spin parameter [69],

$$\chi_{\text{eff}} \equiv \frac{\chi_1 \cos\theta_1 + q\chi_2 \cos\theta_2}{1+q}, \tag{29}$$

which measures the mass-weighted black hole spin projected along the orbital angular momentum [70]. This time, only $n = 4$ delta function spikes are required to fit the data ($\mathcal{I} = 0.06$), showing how much harder it is to measure χ_{eff} than m_1. Computing $\not\pi(\chi_{\text{eff}})$ requires 71.3 seconds. The quicker computation time is likely a result of the lower number of delta functions required. In gray, we plot the PPD for the DEFAULT model from Refs. [15,16], which draws on work from Refs. [20,71]. In orange we plot the PPD for the EXTENDED model from Refs. [28,38], which only analyze 68 binary black-hole events in the population due to data quality concerns regarding one event [72]. To plot the EXTENDED model results, which incorporate a delta function at $\chi_{\text{eff}} = 0$, we plot the 90% interval for the delta function height, δ, multiplied by the same scale factor as $\not\pi$. The continuous contribution to the EXTENDED model is then scaled by the ratio of the PPD evaluated at only the nonzero χ_{eff} $\not\pi$ delta functions to the previously computed scaling.

The data-driven $\not\pi$ includes a delta function at $\chi_{\text{eff}} \approx 0$ and three smaller peaks in the $\chi_{\text{eff}} > 0$ region, but no peaks with $\chi_{\text{eff}} < 0$. The lack of support for $\chi_{\text{eff}} < 0$ is in contrast to Refs. [15,16], which find support for a subpopulation of binary black holes with $\chi_{\text{eff}} < 0$. The strong delta function at $\chi_{\text{eff}} = 0$ lends support to the argument put forward in Refs. [17,27,28] that the data can be well modeled with a sub-population of "nonspinning" $\chi_{\text{eff}} = 0$ binaries, even if there is not strong statistical support for the existence of such a peak [37,38,73]. However, our visual comparison suggests that the EXTENDED model may overpredict the abundance of binaries with $\chi_{\text{eff}} \approx 0.3$. Moreover, we note that the distribution of $\chi_{\text{eff}} = 0$ appears to also be consistent with a smooth, one-sided distribution, maximal at $\chi_{\text{eff}} = 0$, and slowly decaying at larger positive values of $\chi_{\text{eff}} = 0$; that is, a single population.

Turning our attention to the bottom row of Fig. 5, we consider the case of redshift. For this parameter, $n = 6$ ($\mathcal{I} = 0.09$), and takes 116 seconds to compute. Here we plot the merger rate as a function of redshift, $\mathcal{R}(z)$, by dividing the posterior predictive distribution by the PPD by the evolution of the comoving volume and time delay with respect to redshift. The merger rate is more commonly utilized for interpreting the redshift evolution. The $\not\pi$ distribution fits a decrease in the merger rate at a redshift of $z \sim 0.13$. While we caution that $\not\pi$ is purely data informed, and such a feature might diminish with additional observations, the POWER LAW model utilized in Refs. [15,16] lacks the flexibility to resolve such a feature. Comparing our results to Ref. [31], we observe that $\not\pi$ is qualitatively different from the "nonparametric" model [74] used in that paper. Our best guess is that the reconstruction from Ref. [31] is reasonable, and that the different features in $\not\pi$ are due to noise fluctuations, though it is possible that the smooth spline structure imposed by the [31] model is misspecified or that the prior on "knot location" is somehow subtly influencing the fit. As more gravitational-wave observations are made, finer structure may emerge in the redshift evolution of the binary merger rate. These differences between the parametric reconstructions and $\not\pi$ might present the first hints of such structure. We suggest that future redshift models include additional flexibility to study the possibility of a deficit of mergers in the nearby Universe.

By using the iterative "grid-based" method (without further constrained optimization), we also demonstrate the computation of a two-dimensional $\not\pi$ distribution. In particular, we study the joint distribution of mass ratio q and effective spin inspiral parameter χ_{eff}. Recent studies have explored the possibility of astrophysical correlations between q and χ_{eff} [16,21,75], finding an anticorrelation, i.e., more unequal mass systems typically possess a effective spin inspiral parameter. The presence of an anticorrelation in the q-χ_{eff} distribution has implications for the formation environments of binary black holes. Ref. [76], for example, propose that such an anticorrelation could be due to assembly of binary black holes in active galactic nuclei.

In Fig. 6 we plot $\not\pi(q, \chi_{\text{eff}})$ as eight colored pixels. It is easier to digest this $\not\pi$ plot than the superposition of single-event, 90% credible intervals for all 69 events (gray). In order to compare $\not\pi$ to recent models, we plot the 90% contours of *maximum a posteriori* distribution estimates for the DEFAULT model in Ref. [16] which assumes no correlation (black curve), the CORRELATED model from Ref. [21] (blue curve) and the COPULA model from Ref. [75]. From visual examination of $\not\pi$, it is clear that the anticorrelation identified in Ref. [21] is based on actual features in the data: the pixels corresponding to the delta functions $\not\pi$ are consistent with anticorrelation between (q, χ_{eff}). However, $\not\pi$ is also consistent with they hypothesis that there are separate subpopulations located at different regions in the q-χ_{eff} space (an instance of Simpson's reversal [77]).

B. Upper bounds on population model likelihoods

In Table I we report the difference in natural-log likelihood comparing the various population models to the maximum population likelihood $\not{\mathcal{L}}$:

$$\ln \not{\mathcal{B}} \equiv \ln \not{\mathcal{L}} - \ln \mathcal{L}_{\max}(M). \tag{30}$$

The $\ln \not{\mathcal{B}}$ values in Table I measure the fit of population models relative to the best possible fit. Motivated by the typical threshold for model selection in terms of Bayes factors [46], a value of $\ln \not{\mathcal{B}} \lesssim 8$ indicates that the population model is very close to the maximum population likelihood [1], which would imply that the fit cannot be dramatically improved. A large value of $\ln \not{\mathcal{B}}$ by itself does not imply that a model is "wrong" or unsuitable to describe the data, but it does quantify the extent to which an alternative model can in principle improve over the current offerings.

Returning to Table I, the POWER LAW + PEAK model for m_1 shows the most potential room for improvement. This

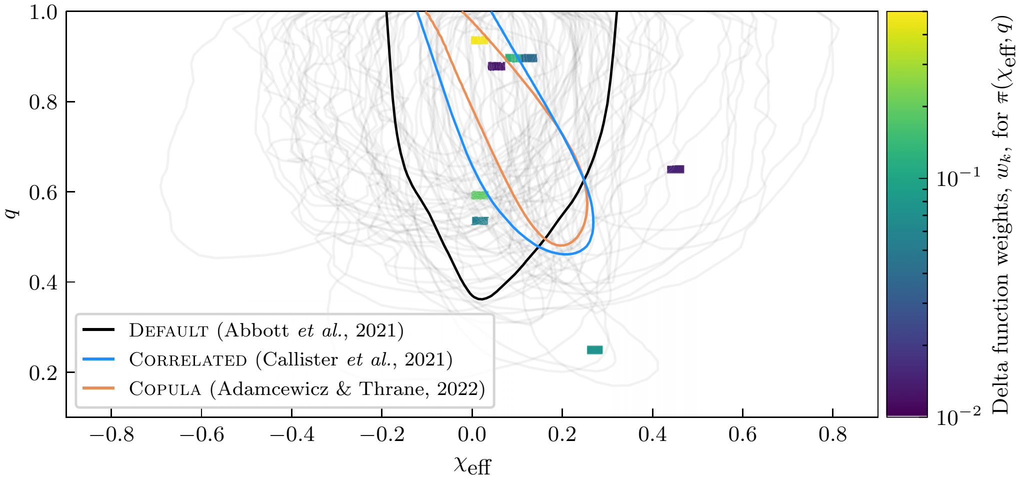

FIG. 6. The joint distribution $\not\pi(q, \chi_{\rm eff})$ represented by eight colored pixels. The pixel color is related to the delta-function weight. The purely data-derived $\not\pi$ can be compared to the 90% contours of *maximum a posteriori* distribution estimates for three specific models. The black curve shows the reconstructed population given the DEFAULT model from Ref. [16] (which does not allow for correlation) while the blue and orange curves show the reconstructed population given by the CORRELATED model from Ref. [21] and the COPULA model from Ref. [75], respectively. The grey contours correspond to the 90% credible intervals of the 69 events in GWTC-3 [4,16].

may be due to structure identified using the SPLINE model, which is missing from the less flexible POWER LAW + PEAK. However, the m_1 measurements are also the most informative in Table I (with the largest value of $\mathcal{I}$). With more information, it is probably easier to concoct an *a posteriori* model with a large population likelihood that explains various features in the distribution of m_1 through overfitting. The DEFAULT and EXTENDED spin models both exhibit $\ln \mathcal{B} < 8$, which implies that neither model can be unequivocally ruled out, though the EXTENDED model provides a somewhat better fit with a natural-log likelihood difference of 4.17. We also note that the $\chi_{\rm eff}$ and z observations are noticeably less informative, and simultaneously the associated values of $\mathcal{L}_{\rm max}(M)$ are closer to $\pounds$. This might indicate that, while there are features present in $\not\pi$ that are present in the data, they are not statistically significant.

C. Model criticism in gravitational-wave astronomy

It would be interesting to make a version of the left-hand panel of Fig. 4 using the population models from gravitational-wave astronomy discussed in the previous subsection. Unfortunately, this is quite computationally difficult. First, we would need to run single-event parameter estimation of $N \approx 69$ events drawn from a random realization of the population fit to the observed gravitational-wave events. This needs to be repeated $\mathcal{O}(1000)$ times to produce the refined contours as those shown in the toy-model example (Fig. 4). However, as an initial demonstration, we generate three simulated catalogs of 69 events using three draws from the POWER LAW + PEAK–DEFAULT–POWER LAW hyperposterior informed by observations from GWTC-3 [16]. These simulated observations were produced with injections of the IMRPHENOMXPHM [63] waveform into simulated Gaussian noise colored by the power spectral density from the first half of the third LVK observing run.

TABLE I. The performance of different population models relative to $\not\mathcal{M}$. The quantity $\mathcal{B}$ [Eq. (30)] is a measure of the population likelihood of each model relative the maximum possible population likelihood $\pounds$. The "informativeness" $\mathcal{I}$ [Eq. (19)] is a measure of the information available about the distribution of each parameter.

Parameter	$\mathcal{I}$	Model	$\ln \mathcal{B}$
m_1	0.15	POWER LAW + PEAK	14.89
		SPLINE	6.66
$\chi_{\rm eff}$	0.06	DEFAULT	7.70
		EXTENDED	3.53
z	0.09	POWER LAW	8.93
		SPLINE	6.59

We then run Bayesian hierarchical inference to determine the posterior predictive distributions from the parameterized model. Using the posterior predictive distributions, following the calculation undertaken for the collection of real gravitational-wave observations, we produce the one-dimensional marginal likelihoods which are then used to compute $\pounds$ and $\mathcal{L}_{\rm max}(M)$. Unlike in Fig. 4, where enough simulated catalogs are produced to construct an expected distribution in the $(\pounds, \mathcal{L}_{\rm max}(M))$ plane, here we are required to model and fit the distribution. We employ Bayesian inference and a simple multivariate Gaussian distribution model to estimate the structure in the expected $(\pounds, \mathcal{L}_{\rm max}(M))$ distribution. We use a Wishart prior on the covariance matrix [78]. We use the posterior predictive distribution of fitted Gaussian distributions to estimate whether the models utilized in Ref. [16] are inadequate for the observations.

The results are shown in Fig. 7 for the primary black-hole mass, effective inspiral parameter, and redshift. The blue dots correspond to the three simulated gravitational-wave catalogs, whereas the black star corresponds to the observed values from GWTC-3. The gray ellipses are 3σ intervals for $(\pounds, \mathcal{L}_{\rm max}(M))$, each associated with a different realisation of our Gaussian fit. (The large amount of scatter is due to the

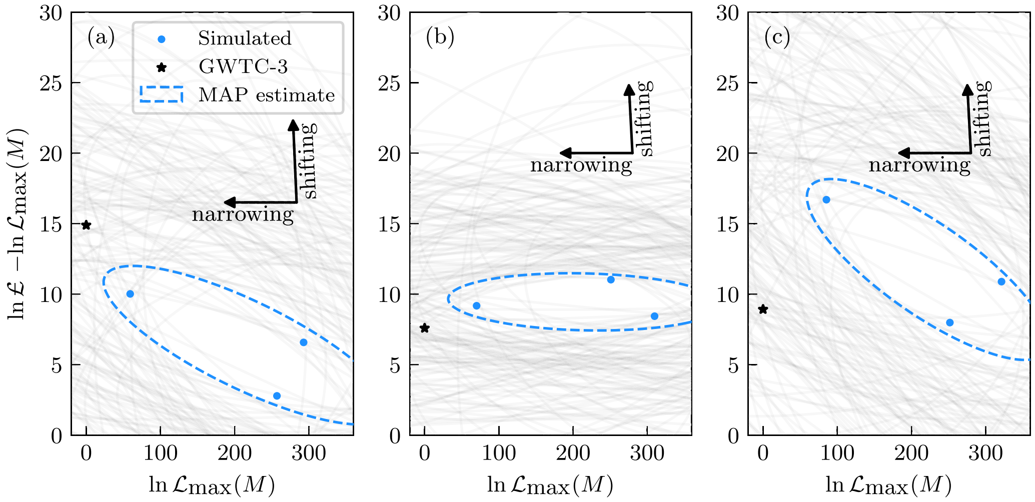

FIG. 7. Demonstration of the $(\pounds, \mathcal{L}_{\max}(M))$ model misspecification test for three parametrized models used in Ref. [16]: (a) the POWER LAW + PEAK model for the primary black-hole mass distribution, (b) the DEFAULT for the χ_{eff} distribution, and (c) the POWER LAW redshift distribution. Due to the limited number of simulated gravitational-wave catalogs, we model the expected distribution $p(\pounds, \mathcal{L}_{\max}(M))$ as a multivariate Gaussian distribution and infer the possible mean and covariance matrix from the three simulated values (blue). The grey ellipses correspond to the 3σ confidence intervals for 100 different realizations of the possible distribution. The dashed blue ellipses correspond to the *maximum a posteriori* (MAP) predictive distributions. The inferred values of $(\pounds, \mathcal{L}_{\max}(M))$ from the 69 events in GWTC-3 are shown by the black stars. The likelihoods are normalized by the maximum likelihood inferred from the GWTC-3 model. From the inferred ellipses, we can conclude that there is a possibility that some or all models used are inadequate for the observations. Further studies with larger simulated catalogs are required to truly determine whether these models are misspecified.

fact that we are attempting to fit a Gaussian to just three points.) The dashed blue curve corresponds to the *maximum a posteriori* (MAP) estimate. The value of $\mathcal{L}_{\max}(M)$ has been normalized to the value found for GWTC-3. The inferred points in $(\pounds, \mathcal{L}_{\max}(M))$ for GWTC-3 typically reside beyond the 3σ confidence interval, which we use as our criteria for misspecification.

We calculate a p value for each panel, which quantifies the probability of observing the GWTC-3 values for $(\pounds, \mathcal{L}_{\max}(M))$ given our fit; small p values are indicative of misspecification. For the POWER LAW + PEAK primary black-hole mass model is misspecified we find $p = 47\%$, for the DEFAULT χ_{eff} model we find $p = 44\%$, and for the redshift POWER LAW model we find $p = 10\%$. None of the models we consider are clearly ruled out as misspecified, as the sensitivity of this test is somewhat weakened by the small number of simulated catalogs. It would not surprise us if a more aggressive followup study $\mathcal{O}(1000)$ simulations identified one or more models as more obviously misspecified.

One important caveat to these results is that the overall normalization of the likelihood depends on the computation of the individual observation Bayesian evidences. With stark differences between the analyses made in Refs. [4,16], it is difficult to accurately emulate the correct overall normalization of the likelihood. This globally impacts in the scale of $\mathcal{L}_{\max}(M)$ for the simulated catalog, potentially shifting the distributions closer or further from the inferred GWTC-3 result. In addition, the robustness of the evidence computed within Ref. [4] is not guaranteed (see, e.g., Ref. [37]).

There are a number of solutions to address the computational cost of this analysis. While probably not realistic in the near future, it may be possible to represent the likelihood functions of simulated events using a Fisher matrix approximation, which would speed up the calculation significantly. However, verifying that this approximation produces adequately estimates for $\pounds, \mathcal{L}_{\max}(M)$ could remain a challenge. Another possibility worthy of investigation is the idea that the distribution of $\pounds, \mathcal{L}_{\max}(M)$ might have some quasi-universal properties. If it can be shown that a large class of problems produce a similarly shaped distribution of $\pounds, \mathcal{L}_{\max}(M)$, perhaps a relatively small number of simulations can be used to work out the shape of $\boldsymbol{p}(\pounds, \mathcal{L}_{\max}(M))$. We leave this for future work. Perhaps most promising are efforts to speed up inference with various machine learning schemes; see, e.g., Ref. [79]. As these tools become more reliable, it may become possible to estimate $(\pounds, \mathcal{L}_{\max}(M))$ in a matter of seconds, which would in turn enable precision tests of misspecification.

V. CONCLUSION

The $\pounds$ formalism provides a useful lens through which to view population studies in gravitational-wave astronomy. It provides an upper bound on the Bayesian evidence for population models, $\pounds$. The associated pseudo-prior distribution $\not\pi$ is a sum of delta functions. The $\not\pi$ distribution can be used to see which features in a reconstructed distribution are model dependent and which are genuinely present in the data. The $\not\pi$ distribution can also draw attention to features in the data that are not fit by current models, providing a tool for the design of new models. Finally, the $\pounds$ formalism can be used to determine if a model is misspecified, by comparing the values of $(\pounds, \mathcal{L}_{\max}(M))$ to the expected distribution of these quantities given the model M. This comparison can be made quantitatively with a p value. And, by comparing the measured values of $(\pounds, \mathcal{L}_{\max}(M))$ to the distribution expected given the model, it is possible to see the way in which the model is misspecified. Constructing a distribution of £, $\mathcal{L}_{\max}(M)$ may be computationally prohibitive in gravitational-wave astronomy, and future work is required to investigate simplifying assumptions that might bring down the cost.

While we have introduced the £ formalism within the context of gravitational-wave astronomy, the framework is general, and we expect it can be applied to a broad range of problems in astronomy and beyond where one seeks to infer the distribution of parameters θ with potentially unreliable hierarchical models.

ACKNOWLEDGMENTS

We thank Katerina Chatziioannou, Colm Talbot, Isaac Legred, Isobel Romero-Shaw, and Paul Lasky for insightful discussions about the £ formalism. We thank Rory Smith, Gael Martin, David Frazier, and Andy Casey for input on early discussions regarding using £ for model misspecification tests. We are grateful to Jacob Golomb for discussions focused on computing a continuous representation of the detection probability for gravitational-wave astronomy. We are indebted to Bernard Whiting for important discussions regarding the convex hull formulation of population distributions. We thank Tom Callister for comments on an early version of the manuscript. This material is based upon work supported by NSF's LIGO Laboratory which is a major facility fully funded by the National Science Foundation. This research has made use of data, software and/or web tools obtained from the Gravitational Wave Open Science Center [80], a service of LIGO Laboratory, the LIGO Scientific Collaboration and the Virgo Collaboration. Virgo is funded by the French Centre National de Recherche Scientifique (CNRS), the Italian Istituto Nazionale della Fisica Nucleare (INFN) and the Dutch Nikhef, with contributions by Polish and Hungarian institutes. The authors are grateful for computational resources provided by the LIGO Laboratory and supported by National Science Foundation Grants No. PHY-0757058 and No. PHY-0823459. This paper carries LIGO Document No. P2200309. E.T. is supported through Australian Research Council (ARC) Centre of Excellence Grants No. CE170100004 and No. ARC DP230103088.

APPENDIX: OUTLINE OF $\not\pi$ STRUCTURE PROOF

1. Overview

In this Appendix we outline the basic ideas underpinning the proof from Ref. [43] by Lindsay that $\not\pi$ consists of a sum of $\leqslant N$ delta functions:

$$\not\pi(\theta) = \sum_{k=1}^{n} w_k\,\delta(\theta - \theta_k). \tag{A1}$$

Our aim is to provide readers with a qualitative understanding. To this end, we consider a simple example of $N = 2$ measurements, each characterized by a Gaussian likelihood functions. Our example measurements are depicted in the right-hand column of Fig. 8, which shows two single-event likelihoods (one in purple, the other in red), both conditioned on some parameter θ. In each row of Fig. 8, we vary the separation of these two single-event likelihood functions relative to their width: far apart in the top row, becoming closer together in the two subsequent rows. We show below how $\not\pi$ consists of either one or two delta functions, depending on this relative separation and explain how this generalizes to $N > 2$.

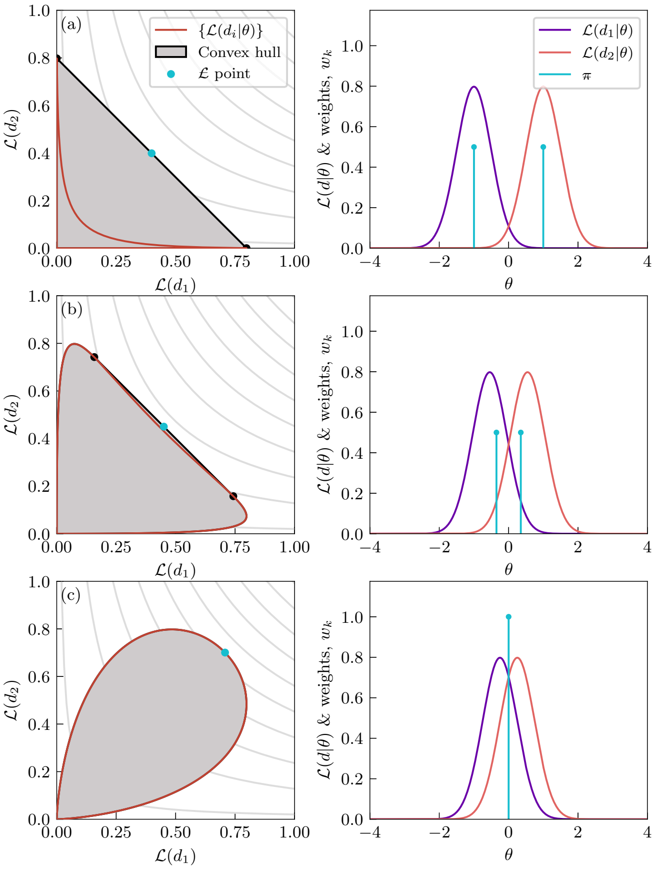

FIG. 8. Visual illustrations of the proof in Ref. [43]. The left-hand column panels show the atomic likelihood vectors (red), the convex hull produced from the red curve (grey with black outline), and the cyan point on the convex-hull boundary with the maximum population likelihood £. The black points correspond to the points from the set of atomic likelihood vectors which generate the maximum population likelihood. The right-hand column panels show three examples of $N = 2$ single-event likelihood functions (purple and red). The distribution of $\not\pi$ is indicated with one or more cyan spikes. These spikes correspond to the £ solution (cyan dot) in the corresponding left-hand panel. In (a), the two single-event likelihoods are mostly disjoint and so two delta functions are required to maximize the population likelihood (cf. Fig. 1 in Ref. [43]). As the two single-event likelihoods begin to overlap further, these two delta functions move closer together as shown in (b). Moving the single-event likelihoods closer still, the set of atomic likelihood vectors becomes the boundary of the convex hull, at which point only one delta function is required to maximize the likelihood as shown in (c).

Lindsay's proof relies on the mathematics of *convex hulls*, geometric shapes which can be defined in arbitrarily high dimensions. If one draws a line between any two points on a convex hull, all the points on that line are also part of the hull. (The gray shaded regions in the left-hand column of Fig. 8 are all examples of convex hulls.) Convex hulls are often used in optimization problems with constraints where the optimal solution occurs on the boundary of the hull, which is determined by the constraints. In Lindsay's proof, the relevant constraint equation is the unitarity of the $\not{\pi}(\theta)$:

$$\int d\theta \, \not{\pi}(\theta) = 1. \tag{A2}$$

The unitarity constraint means that the form of $\not{\pi}(\theta)$ that maximizes the population likelihood exists on the boundary of a complex hull.

2. A geometric picture

For the sake of simplicity, we ignore the impact of the selection function [81]. We represent the observations using what Lindsay refers to as an *atomic likelihood vector*,

$$\boldsymbol{L}(\widehat{\theta}) \equiv \{\mathcal{L}(d_1|\widehat{\theta}), \mathcal{L}(d_2|\widehat{\theta}), \ldots, \mathcal{L}(d_N|\widehat{\theta})\}. \tag{A3}$$

Each element of this vector is a single-event likelihood marginalized over a delta-function prior peaking at $\widehat{\theta}$:

$$\mathcal{L}(d_i|\widehat{\theta}) = \int d\theta_i \, \mathcal{L}(d_i|\theta_i)\, \delta(\theta - \widehat{\theta}). \tag{A4}$$

This allows us to represent the problem in an abstract N-dimensional likelihood space. The left-hand column of Fig. 8 provides a visualization of such a two-dimensional atomic likelihood vector space. Scanning over all possible values of $\widehat{\theta}$ traces out the red curve in the atomic likelihood vector space, which represents all possible values of the atomic likelihood vector $\boldsymbol{L}(\theta)$. By varying $\widehat{\theta}$, we can make an individual element of the atomic likelihood vector large, but doing so may make other elements of the vector small, as we see in the top row with widely separated single-event likelihood functions.

The weighted sum of atomic likelihood vectors,

$$\boldsymbol{L}(\vec{w}) = \sum_k w_k \, \boldsymbol{L}(\widehat{\theta}_k), \tag{A5}$$

yields a vector of likelihoods with elements

$$\mathcal{L}(d_i|\vec{w}) = \sum_k w_k \, \mathcal{L}(d_i|\widehat{\theta}_k), \tag{A6}$$

corresponding to the marginal likelihood given a prior of delta functions,

$$\pi(\theta) = \sum_k w_k \, \delta(\theta - \widehat{\theta}_k), \tag{A7}$$

where

$$\sum_k w_k = 1. \tag{A8}$$

This means we can construct more general *marginal likelihood vectors* with a linear combination of atomic vectors. Furthermore, in the continuum limit, *any* prior can be used to *marginalize* over the atomic likelihood vectors. Elements of the marginal likelihood vector in the continuum limit take the form

$$\mathcal{L}(d_i|M) = \int d\widehat{\theta}_i \, \mathcal{L}(d_i|\widehat{\theta}_i)\, \pi(\widehat{\theta}_i|M). \tag{A9}$$

Let us consider again the $N = 2$ example illustrated in Fig. 8. If we pick any two points on the red curve, each corresponding to some value of $\widehat{\theta}$, which we denote A and B, we can define two basis vectors: $\hat{e}_A$ and $\hat{e}_B$. The linear combinations of these two basis vectors forms a line connecting A and B. All of the points along this line represent likelihood vectors constructed from $N = 2$ delta functions. By connecting together every possible pair of points on the red atomic likelihood points, we map out the gray region: the convex hull. Every possible marginal likelihood vector (for *any* choice of prior) is part of the hull. That is, the set of all possible summations is the convex hull and is a representation of all possible probability distributions in the likelihood space. This result is profound: our original problem is reduced from an infinite set of possible population distributions to a closed region in an N-dimensional likelihood space. The construction of the convex hull is unique [43], except in pathological cases further discussed in Sec. 3 of this Appendix.

Now that we have studied the geometry of the atomic likelihood vector space, we ask the question, what point in our convex hull corresponds to the maximum population likelihood? The population likelihood can be written as a product of the marginal likelihood vector elements:

$$\mathcal{L}_{\rm pop}(\vec{d}|M) = \prod_{i=1}^{N} \mathcal{L}(d_i|M). \tag{A10}$$

In $N = 2$ dimensions, we can fix $\mathcal{L}_{\rm pop}(\vec{d})$ and identify hyperbolic curves of the form

$$\mathcal{L}(d_2) = \mathcal{L}(\vec{d})/\mathcal{L}(d_1), \tag{A11}$$

represented in the left-hand column of Fig. 8 by gray curves. All the points on one of these curves have the same population likelihood. If we jump up and to the right from one gray curve to another, the population likelihood increases. These constant-population-likelihood, hyperbolic curves do not depend on any population model. The population likelihood is then maximized by finding the point on the boundary of the hull tangent to the gray curve with the largest population likelihood (the topmost and rightmost gray curve). In general, the maximum population likelihood point lies on the boundary of the hull [43,82]. Our maximization problem can therefore be rewritten as a geometry problem.

We now turn our attention to the different rows of Fig. 8. In the top row, the two single-event likelihoods (right) are widely separated. The cyan dot on the left-hand plot shows the maximum population likelihood point on the surface of the hull. This is where the population likelihood has a value of £. It falls on a straight black surface of the hull, but not on the red atomic likelihood vector curve. This means that the cyan point is a linear combination of two atomic likelihood vectors, which are indicated by the two black points (cf. Fig. 1 in Ref. [43]). Thus, the maximum population likelihood solution consists of two delta functions, each corresponding to a different atomic vector. This linear combination of delta functions is shown in the right-hand panel with cyan spikes. Unsurprisingly, they coincide with the two single-event likelihood function peaks.

Moving down to the second row, the single-event likelihood functions (right) are now closer together. The shape of the hull changes accordingly (left). The hull boundary point that maximizes the population likelihood still does not fall on the red curve of atomic vectors. Again, it is a linear combination of two black points. However, since the shape of the hull has changed, the black points have moved relative to the top row. The corresponding delta function spikes (right) therefore shift toward $\theta = 0$ and no longer correspond to the maximum likelihood points of the single-event likelihoods.

In the bottom row, the single-event likelihood functions (right) are closer still. The hull (left) has now changed shape so that the cyan point marking the maximum population likelihood falls on the red curve denoting the set of atomic vectors (left). This means that the likelihood can be maximized with a single delta function at $\theta = 0$ (right). In each case (and almost all scenarios; see Sec. 3 in this Appendix) the convex hull is unique, and so the cyan point of maximum population likelihood is unique as well. In all but the most pathological cases, Carathéodory's theorem [45,83] states that all points on the boundary of a convex hull can be constructed by, at most, N points that were used to initially construct the hull (in our problem these are the atomic likelihood vectors). The relative weight of each delta function corresponds to the position along the boundary of the hull [43]. Thus, the population prior corresponding to the maximum population likelihood is a construction of a finite set of, at most, N delta functions.

The transition from two delta functions to one delta function occurs when the red curve passes through the black one (when the set of atomic likelihood vectors becomes convex). During this transition, the cyan point changes from residing on a straight line connecting two atomic vectors to residing on a single atomic vector point. This picture generalizes to higher dimensions. Solutions with three delta functions (which can only exist when $N \geqslant 3$) reside on two-dimensional planes. Solutions with four delta functions (which can only exist when $N \geqslant 4$) reside on three-dimensional hyperplanes, and so on.

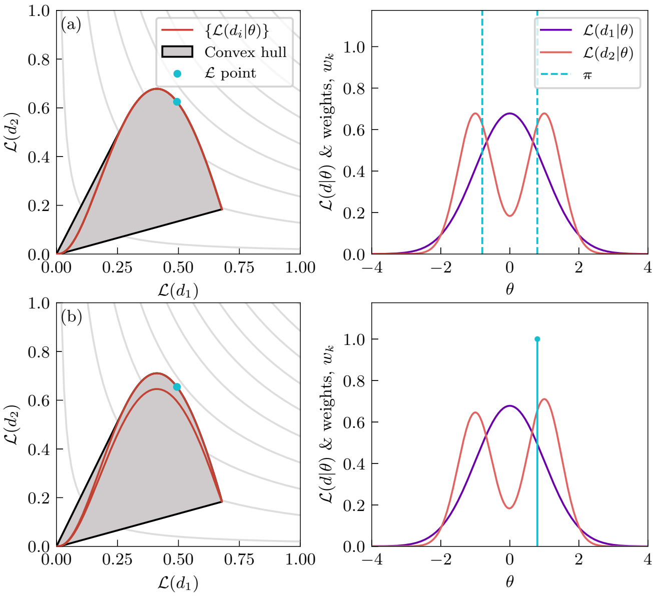

FIG. 9. Demonstration of a pathological failure of the uniqueness of $\bar{\pi}$. This occurs when multiple distributions map to exactly the same point on the convex hull. In (a), a perfectly symmetric, bimodal single-event likelihood has two delta functions with produce the same population likelihood. Therefore, any combination of the two is a valid $\bar{\pi}$. However, such perfectly symmetric multi-modal distributions do not typically occur in gravitational-wave data analysis. We see here we can break this degeneracy by only slightly breaking the symmetry, shown in (b).

3. Pathological cases

While we see that the maximum population likelihood almost always corresponds to a finite, unique set of N or fewer delta functions, there are pathological cases (not likely to come up in real-world data analysis) where this is not the case. Such cases stem from the maximum population likelihood point not being unique. So while the maximum population likelihood point is still found, multiple distributions can map to the same point in likelihood space. This requires artificial degeneracies in the measurements. In Fig. 9, we demonstrate one such example with two likelihood functions perfectly symmetric about $\theta = 0$ and one of which is bimodal. In the likelihood space, the $\pounds$ point corresponds to two possible positions of the delta function. However, unlike in Fig. 8(a) where the two possible delta function positions are separated, here they correspond to same point in likelihood space. Therefore, any normalized combination of the two delta functions produces the maximum population likelihood. This is emphasized by the dashed blue lines in the right column of Fig. 9(a), indicating that any combination of the two delta functions here is a permissible solution. However, we emphasize that this pathology arises from an artificial degeneracy, which is immediately broken if the likelihood functions are not precisely symmetric as demonstrated in Fig. 9(b). Other, even more pathological, situations can be constructed where infinitely many atomic likelihood vectors reside at the maximum population likelihood point, allowing for arbitrarily structured $\bar{\pi}$ distributions. However, all such situations require regions of perfectly uniform likelihood functions, which we do not expect in realistic observations, at least not in gravitational-wave astronomy.

[1] E. Thrane and C. Talbot, An introduction to Bayesian inference in gravitational-wave astronomy: parameter estimation, model selection, and hierarchical models, Publ. Astron. Soc. Aust. **36**, e010 (2019).

[2] S. Vitale, D. Gerosa, W. M. Farr, and S. R. Taylor, Inferring the properties of a population of compact binaries in presence of selection effects, in *Handbook of Gravitational Wave Astronomy*, edited by C. Bambi, S. Katsanevas, and K. D. Kokkotas (Springer, Singapore, 2022), pp. 1–60.

[3] I. Mandel, W. M. Farr, and J. R. Gair, Extracting distribution parameters from multiple uncertain observations with selection biases, Mon. Not. R. Astron. Soc. **486**, 1086 (2019).

[4] R. Abbott *et al.*, GWTC-3: Compact binary coalescences observed by LIGO and Virgo during the second part of the third observing run, arXiv:2111.03606.

[5] We adopt the threshold utilized in [16] of a false-alarm rate $<1\ \mathrm{yr}^{-1}$.

[6] J. Aasi *et al.*, Advanced LIGO, Class. Quantum Grav. **32**, 115012 (2015).

[7] F. Acernese *et al.*, Advanced Virgo: a second-generation interferometric gravitational wave detector, Class. Quantum Grav. **32**, 024001 (2015).

[8] T. Akutsu *et al.*, Overview of KAGRA: Detector design and construction history, Prog. Theor. Exp. Phys. **2021**, 05A101 (2021).

[9] S. Olsen, T. Venumadhav, J. Mushkin, J. Roulet, B. Zackay, and M. Zaldarriaga, New binary black hole mergers in the LIGO-Virgo O3a data, Phys. Rev. D **106**, 043009 (2022).

[10] A. H. Nitz, C. D. Capano, S. Kumar, Y. Wang, S. Kastha, M. Schäfer, R. Dhurkunde, and M. Cabero, 3-OGC: Catalog of gravitational waves from compact-binary mergers, Astrophys. J. **922**, 76 (2021).

[11] B. Zackay, L. Dai, T. Venumadhav, J. Roulet, and M. Zaldarriaga, Detecting gravitational waves with disparate detector responses: Two new binary black hole mergers, Phys. Rev. D **104**, 063030 (2021).

[12] T. Venumadhav, B. Zackay, J. Roulet, L. Dai, and M. Zaldarriaga, New binary black hole mergers in the second observing run of advanced LIGO and Advanced Virgo, Phys. Rev. D **101**, 083030 (2020).

[13] B. Zackay, T. Venumadhav, L. Dai, J. Roulet, and M. Zaldarriaga, Highly spinning and aligned binary black hole merger in the advanced LIGO first observing run, Phys. Rev. D **100**, 023007 (2019).

[14] R. Abbott *et al.*, Binary black hole population properties inferred from the first and second observing runs of advanced LIGO and Advanced Virgo, Astrophys. J. **882**, L24 (2019).

[15] R. Abbott Jr. *et al.*, Population properties of compact objects from the second LIGO-Virgo gravitational-wave transient catalog, Astrophys. J. Lett. **913**, L7 (2021).

[16] R. Abbott *et al.*, The population of merging compact binaries inferred using gravitational waves through GWTC-3, arXiv:2111.03634.

[17] J. Roulet, H. S. Chia, S. Olsen, L. Dai, T. Venumadhav, B. Zackay, and M. Zaldarriaga, Distribution of effective spins and masses of binary black holes from the LIGO and Virgo O1–O3a observing runs, Phys. Rev. D **104**, 083010 (2021).

[18] W. M. Farr, S. Stevenson, M. Coleman Miller, I. Mandel, B. Farr, and A. Vecchio, Distinguishing spin-aligned and isotropic black hole populations with gravitational waves, Nature (London) **548**, 426 (2017).

[19] C. Talbot and E. Thrane, Measuring the binary black hole mass spectrum with an astrophysically motivated parameterization, Astrophys. J. **856**, 173 (2018).

[20] C. Talbot and E. Thrane, Determining the population properties of spinning black holes, Phys. Rev. D **96**, 023012 (2017).

[21] T. A. Callister, C. Haster, K. K. Y. Ng, S. Vitale, and W. M. Farr, Who ordered that? Unequal-mass binary black hole mergers have larger effective spins, Astrophys. J. Lett. **922**, L5 (2021).

[22] M. Fishbach, C. Kimball, and V. Kalogera, Limits on hierarchical black hole mergers from the most negative χ_{eff} systems, Astrophys. J. Lett. **935**, L26 (2022).

[23] S. Biscoveanu, M. Isi, S. Vitale, and V. Varma, New Spin on LIGO-Virgo Binary Black Holes, Phys. Rev. Lett. **126**, 171103 (2021).

[24] S. Biscoveanu, T. A. Callister, C. Haster, K. K. Y. Ng, S. Vitale, and W. M. Farr, The binary black hole spin distribution likely broadens with redshift, Astrophys. J. Lett. **932**, L19 (2022).

[25] S. Vitale, R. Lynch, R. Sturani, and P. Graff, Use of gravitational waves to probe the formation channels of compact binaries, Class. Quantum Grav. **34**, 03LT01 (2017).

[26] S. Stevenson, C. P. L. Berry, and I. Mandel, Hierarchical analysis of gravitational-wave measurements of binary black hole spin–orbit misalignments, Mon. Not. R. Astron. Soc. **471**, 2801 (2017).

[27] S. Miller, T. A. Callister, and W. M. Farr, The low effective spin of binary black holes and implications for individual gravitational-wave events, Astrophys. J. **895**, 128 (2020).

[28] S. Galaudage, C. Talbot, T. Nagar, D. Jain, E. Thrane, and I. Mandel, Building better spin models for merging binary black holes: Evidence for nonspinning and rapidly spinning nearly aligned subpopulations, Astrophys. J. Lett. **921**, L15 (2021).

[29] M. Fishbach, D. E. Holz, and W. M. Farr, Does the black hole merger rate evolve with redshift? Astrophys. J. **863**, L41 (2018).

[30] B. Edelman, Z. Doctor, J. Godfrey, and B. Farr, Ain't no mountain high enough: Semiparametric modeling of LIGO–Virgo's binary black hole mass distribution, Astrophys. J. **924**, 101 (2022).

[31] B. Edelman, B. Farr, and Z. Doctor, Cover your basis: Comprehensive data-driven characterization of the binary black hole population, Astrophys. J. **946**, 16 (2023).

[32] J. Golomb and C. Talbot, Searching for structure in the binary black hole spin distribution, arXiv:2210.12287.

[33] Here, we paraphrase the aphorism attributed to statistician George Box: "all models are wrong, but some are useful."

[34] I. M. Romero-Shaw, E. Thrane, and P. D. Lasky, When models fail: an introduction to posterior predictive checks and model misspecification in gravitational-wave astronomy, Publ. Astron. Soc. Aust. **39**, e025 (2022).

[35] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian Data Analysis*, 3rd ed. (Taylor & Francis, London, 2013).

[36] R. Essick, A. Farah, S. Galaudage, C. Talbot, M. Fishbach, E. Thrane, and D. E. Holz, Probing Extremal gravitational-wave events with coarse-grained likelihoods, Astrophys. J. **926**, 34 (2022).

[37] T. A. Callister, S. J. Miller, K. Chatziioannou, and W. M. Farr, No evidence that the majority of black holes in binaries have zero spin, Astrophys. J. Lett. **937**, L13 (2022).

[38] H. Tong, S. Galaudage, and E. Thrane, Population properties of spinning black holes using the gravitational-wave transient catalog 3, Phys. Rev. D **106**, 103019 (2022).

[39] J. Kiefer and J. Wolfowitz, Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters, Ann. Math. Stat. **27**, 887 (1956).

[40] L. Simar, Maximum likelihood estimation of a compound poisson process, Ann. Stat. **4**, 1200 (1976).

[41] N. Laird, Nonparametric maximum likelihood estimation of a mixing distribution, J. Am. Stat. Assoc. **73**, 805 (1978).

[42] D. Bohning, Convergence of Simar's algorithm for finding the maximum likelihood estimate of a compound Poisson process, Ann. Stat. **10**, 1006 (1982).

[43] B. G. Lindsay, The geometry of mixture likelihoods: A general theory, Ann. Stat. **11**, 86 (1983).

[44] W. Jiang and C. H. Zhang, General maximum likelihood empirical Bayes estimation of normal means, Ann. Stat. **37**, 1647 (2009).

[45] C. Carathéodory, Über den variabilitätsbereich der fourier'schen konstanten von positiven harmonischen funktionen, Rendiconti del Circolo Matematico di Palermo (1884-1940) **32**, 193 (1911).

[46] H. Jeffreys, *Theory of Probability*, 3rd ed. (Oxford University Press, 1961).

[47] This is a well-known result known as the empirical distribution function [41].

[48] In theory, the constraint condition does not need to be enforced during the analysis. The normalization appears in the selection function term and $\not\pi$. However, since any multiple of the weights (without normalization) would produce an identical likelihood, many numerical optimization methods can falter at these likelihood "plateaus." Therefore, we enforce the constraint to ensure a more robust analysis.

[49] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright *et al.*, SciPy 1.0: Fundamental algorithms for scientific computing in Python, Nat. Methods **17**, 261 (2020).

[50] A. R. Conn, N. I. M. Gould, and P. L. Toint, *Trust Region Methods* (Society for Industrial and Applied Mathematics, Philadelphia, 2000).

[51] A method is "better" if it yields a larger value of $\not\mathcal{L}$ than another approach.

[52] C. Talbot and E. Thrane, Flexible and accurate evaluation of gravitational-wave Malmquist bias with machine learning, Astrophys. J. **927**, 76 (2022).

[53] J. Golomb and C. Talbot, Hierarchical inference of binary neutron star mass distribution and equation of state with gravitational waves, Astrophys. J. **926**, 79 (2022).

[54] For the results in Fig. 2, the computation time of the "combined" approach was the following: 10 observations required only 5.3 seconds, 100 observations required 65 seconds, and 1000 observations required 2780 seconds. Generally, more data tends to require more delta functions (each with a location and a height), meaning the computational difficulty grows with N.

[55] We borrow the language of "possible worlds" from the philosopher David Lewis, who invokes them in his account of counterfactuals and necessity [84].

[56] LIGO Scientific Collaboration and Virgo Collaboration and KAGRA Collaboration, GWTC-3: Compact binary coalescences observed by LIGO and Virgo during the second part of the third observing run parameter estimation, data release, 2021 (unpublished).

[57] LIGO Scientific Collaboration and Virgo Collaboration and KAGRA Collaboration, The population of merging compact binaries inferred using gravitational waves through GWTC-3, data release, 2021 (unpublished).

[58] Y. Pan, A. Buonanno, A. Taracchini, L. E. Kidder, A. H. Mroué, H. P. Pfeiffer, M. A. Scheel, and B. Szilágyi, Inspiral-merger-ringdown waveforms of spinning, precessing black-hole binaries in the effective-one-body formalism, Phys. Rev. D **89**, 084006 (2014).

[59] A. Taracchini *et al.*, Effective-one-body model for black-hole binaries with generic mass ratios and spins, Phys. Rev. D **89**, 061502 (2014).

[60] A. Bohé *et al.*, Improved effective-one-body model of spinning, nonprecessing binary black holes for the era of gravitational-wave astrophysics with advanced detectors, Phys. Rev. D **95**, 044028 (2017).

[61] S. Ossokine *et al.*, Multipolar effective-one-body waveforms for precessing binary black holes: Construction and validation, Phys. Rev. D **102**, 044055 (2020).

[62] M. Hannam, P. Schmidt, A. Bohé, L. Haegel, S. Husa, F. Ohme, G. Pratten, and M. Pürrer, Simple Model of Complete Precessing Black-Hole-Binary Gravitational Waveforms, Phys. Rev. Lett. **113**, 151101 (2014).

[63] G. Pratten *et al.*, Computationally efficient models for the dominant and subdominant harmonic modes of precessing binary black holes, Phys. Rev. D **103**, 104056 (2021).

[64] C. Talbot, R. J. E. Smith, E. Thrane, and G. B. Poole, Parallelized inference for gravitational-wave astronomy, Phys. Rev. D **100**, 043030 (2019).

[65] G. Ashton *et al.*, BILBY: A user-friendly Bayesian inference library for gravitational-wave astronomy, Astrophys. J. Suppl. Series **241**, 27 (2019).

[66] I. M. Romero-Shaw *et al.*, Bayesian inference for compact binary coalescences with BILBY: Validation and application to the first LIGO–Virgo gravitational-wave transient catalogue, Mon. Not. R. Astron. Soc. **499**, 3295 (2020).

[67] J. S. Speagle, DYNESTY: a dynamic nested sampling package for estimating Bayesian posteriors and evidences, Mon. Not. R. Astron. Soc. **493**, 3132 (2020).

[68] LIGO Scientific Collaboration and Virgo Collaboration and KAGRA Collaboration, GWTC-3: Compact binary coalescences observed by LIGO and Virgo during the second part of the third observing run O1+O2+O3 search sensitivity estimates, 2021 (unpublished).

[69] T. Damour, Coalescence of two spinning black holes: An effective one-body approach, Phys. Rev. D **64**, 124013 (2001).

[70] In Eq. (29), $q \equiv m_2/m_1$ is the mass ratio, $\chi_{1,2}$ are the dimensionless black hole spins, and $\theta_{1,2}$ are the spin vector tilt angles relative to the orbital angular momentum.

[71] D. Wysocki, J. Lange, and R. OShaughnessy, Reconstructing phenomenological distributions of compact binaries via gravitational wave observations, Phys. Rev. D **100**, 043012 (2019).

[72] E. Payne, S. Hourihane, J. Golomb, R. Udall, D. Davis, and K. Chatziioannou, The curious case of GW200129: interplay between spin-precession inference and data-quality issues, Phys. Rev. D **106**, 104017 (2022).

[73] M. Mould, D. Gerosa, F. S. Broekgaarden, and N. Steinle, Which black hole formed first? Mass-ratio reversal in massive binary stars from gravitational-wave data, Mon. Not. R. Astron. Soc. **517**, 2738 (2022).

[74] Reference [31]'s spline model is probably better described as "ultraparametrized."

[75] C. Adamcewicz and E. Thrane, Do unequal-mass binary black hole systems have larger χ_{eff}? probing correlations with copulas in gravitational-wave astronomy, Mon. Not. R. Astron. Soc. **517**, 3928 (2022).

[76] B. McKernan, K. E. S. Ford, T. Callister, W. M. Farr, R. O'Shaughnessy, R. Smith, E. Thrane, and A. Vajpeyi,

LIGO–Virgo correlations between mass ratio and effective inspiral spin: testing the active galactic nuclei channel, Mon. Not. R. Astron. Soc. **514**, 3886 (2022).

[77] E. H. Simpson, The interpretation of interaction in contingency tables, J. R. Stat. Soc. **13**, 238 (1951).

[78] Y. Chung, A. Gelman, S. Rabe-Hesketh, J. Liu, and V. Dorie, Weakly informative prior for point estimation of covariance matrices in hierarchical models, J. Educ. Behav. Stat. **40**, 136 (2015).

[79] M. Dax, S. R. Green, J. Gair, M. Pürrer, J. Wildberger, J. H. Macke, A. Buonanno, and B. Schölkopf, Neural importance sampling for rapid and reliable gravitational-wave inference, arXiv:2210.05686.

[80] https://www.gw-openscience.org.

[81] The selection function term, $p_{\rm det}(\theta)$, can be absorbed into the prior to determine $\underline{\pi}$ on the observed population before correcting the detection probability afterwards.

[82] S. Silvey, *Optimal Design: An Introduction to the Theory for Parameter Estimation* (Springer, Berlin, 1980), Vol. 1.

[83] A. W. Roberts and D. E. Varberg, *Convex Functions* (Academic, New York, 1973).

[84] D. K. Lewis, *Counterfactuals* (Blackwell, Cambridge, MA, 1973).