Fast non-Markovian sampler for estimating gravitational-wave posteriors

Vaibhav Tiwari,[1] Charlie Hoy,[2] Stephen Fairhurst,[1] and Duncan MacLeod[1]

[1]*Gravity Exploration Institute, School of Physics and Astronomy, Cardiff University, Queens Buildings,*
The Parade Cardiff CF24 3AA, United Kingdom
[2]*University of Portsmouth, Portsmouth, PO1 3FX, United Kingdom*

(Received 13 March 2023; accepted 2 June 2023; published 6 July 2023)

This article introduces VARAHA, an open-source, fast, non-Markovian sampler for estimating gravitational-wave posteriors. VARAHA differs from existing nested sampling algorithms by gradually discarding regions of low likelihood, rather than gradually sampling regions of high likelihood. This alternative mindset enables VARAHA to freely draw samples from anywhere within the high-likelihood region of the parameter space, allowing for analyses to complete in significantly fewer cycles. This means that VARAHA can significantly reduce both the wall and CPU time of all analyses. VARAHA offers many benefits, particularly for gravitational-wave astronomy where Bayesian inference can take many days, if not weeks, to complete. For instance, VARAHA can be used to estimate accurate sky locations, astrophysical probabilities and source classifications within minutes, which is particularly useful for multimessenger follow-up of binary neutron star observations; VARAHA localizes GW170817 ~30 times faster than LALInference. Although only aligned-spin, dominant multipole waveform models can be used for gravitational-wave analyses, it has the potential to include additional physics. We envision VARAHA being used for gravitational-wave studies, particularly estimating parameters using expensive waveform models, analyzing subthreshold gravitational-wave candidates, generating simulated data for population studies, and rapid posterior estimation for binary neutron star mergers.

DOI: 10.1103/PhysRevD.108.023001

I. INTRODUCTION

Compact binary coalescences (CBCs)—binary black holes (BBHs), binary neutron stars (BNSs) and neutron star (NS)–black hole (BH) binaries—are likely the only gravitational wave (GW) sources observed by the network of ground-based GW observatories so far [1–4]; although other sources have also been suggested including, for example, cosmic strings [5] and vector boson-star mergers [6,7]. The well-understood GW signal morphology produced by CBCs, see, e.g., [8–16], and references therein, facilitates the estimation of binary parameters through parameter estimation (PE), see, e.g., [17,18]. These parameter estimates are needed to, e.g., infer the properties of the source population, see, e.g., [19–26], enhance our understanding of the equation of the state of neutron stars, or probe the cosmological history of the universe [27–29].

These estimated parameters are broadly categorized as (a) intrinsic parameters: parameters that are directly responsible for the orbital evolution of the binary, such as masses, spins, tidal parameters, eccentricity, periastron distance, etc., and (b) extrinsic parameters: parameters that are observer dependent, namely, luminosity distance, binary's orientation from the line of sight, sky location, coalescence phase and coalescence time of the GW signal. For binaries moving on a quasicircular orbit with spins aligned with the orbital angular momentum, the extrinsic parameters do not impact the orbital evolution of a binary and, consequently, can only impart an overall shift to the amplitude or phase evolution of a signal. However, for binaries moving on an eccentric orbit, or with spins misaligned from the orbital angular momentum, the GW signal morphology depends on the extrinsic parameters [30,31]. Although the number of intrinsic parameters may change depending on the physics of the problem, the number of extrinsic parameters remains fixed at 7 [32]. The basics of PE have been thoroughly discussed, and we cite some of the early works [33,34].

Multiple methods have been developed to perform PE on GW signals. The minute-scale analysis, BAYESTAR [35], focuses on the rapid localization of GW signals by estimating only the extrinsic parameters of the signal. It achieves this by keeping the intrinsic parameters fixed to the estimate provided by the GW search analysis that first identified the signal. The packages that perform PE of both the extrinsic and intrinsic (full) parameters through stochastic sampling methods include LALInference, which was previously the go-to analysis for GW PE [36], PyCBC Inference [37] and Bilby [38–41], which offer greater flexibility and modularity. These analyses often employ nested [42] or Markov-chain Monte Carlo (MCMC) sampling [43] to obtain estimates for the binaries parameters. The packages RapidPE and RIFT use a non-Markovian approach to create an embarrassingly parallel infrastructure and provide comparatively faster processing times. They can also provide the marginal likelihood for straightforward model selection [44–46] (although see Refs. [47,48] for other model selection algorithms). Alternatively, methods to approximately estimate the binaries parameters have also been developed [49], including recent advancements with utilizing machine-learning techniques [50–56].

Typically, stochastic sampling analyses that perform full PE can take hundreds or thousands of CPU hours of processing time [57,58]. This high computational requirement is not sustainable, as the detection rate of GW signals is expected to increase [59] due to the continued improvement in the sensitivity of the GW detectors. Fast, minute-scale PE is therefore crucial, especially for low-latency analyses where accurate skymaps and source classification probabilities are needed for timely follow-up by other multimessenger facilities.

Attempts at improving computation time have primarily focused on speeding up waveform generation and computation of the likelihood function [48,60–64], or by utilizing machine learning techniques [50–56]. However, a significant improvement in computation time can also be achieved by efficiently populating the parameter space. In this paper, we introduce VARAHA, an alternative sampling technique that iteratively discards regions of low likelihood, and converges to the region of the parameter space that contains high posterior probability density (i.e. the posterior mass). We achieve significant gains in speed by introducing (a) a non-Markovian method that performs a comparable number of computational operations, resulting in a similar number of effective samples, as nested sampling but in significantly fewer iterations, and (b) splitting one large-dimensional sampling problem into two small-dimensional problems, where it samples the extrinsic parameters first and uses the acquired information to also sample the intrinsic parameters. These advantages result in significantly reduced processing times arising from greatly improved process parallelization and array vectorization in the analysis.

VARAHA can perform GW PE in a matter of a few minutes. Currently, it is limited to using waveform models that (a) assume the spins are aligned with the orbital angular momentum, meaning that the binary does not precess [30], and (b) restrict attention to only the $\ell = 2$ gravitational-wave multipole, meaning that higher multipoles are neglected. Nevertheless, VARAHA can meaningfully be used to perform fast PE to localize and classify a source for electromagnetic follow-up, estimate parameters for a large number of subthreshold GW candidates, and generate PE for simulated populations. We intend to make future extensions that will extend its applicability for the estimation of in-plane spins, eccentricity, and tidal deformability. Future extensions will also include uncertainties arising from the calibration of detector data, and fast methods for waveform generation and matched filtering, which currently consumes a significant portion of the computation.

In Sec. II we describe the basics of the analysis and the factors responsible for the faster processing time. We describe its application to the parameter estimation of GWs in Sec. III. In Sec. IV, we present PE for the individual observations GW151226 [65] and GW170817 [66], as well as a population level validation using hundreds of simulated signals. We also discuss the computational requirements of VARAHA as well as its scalability with the number of CPUs.

II. METHOD

PE is the process of obtaining the probability distribution of parameters, $\boldsymbol{\theta}$, for a model, m, which is believed to describe the observed data, d. Given d, PE generates an estimate for the *posterior* probability density function (PDF), $p(\boldsymbol{\theta}|d, m)$, through Bayes' theorem,

$$p(\boldsymbol{\theta}|d, m) \propto \mathcal{L}(d|\boldsymbol{\theta}, m)\pi(\boldsymbol{\theta}|m). \tag{1}$$

The likelihood, $\mathcal{L}(d|\boldsymbol{\theta}, m)$, is a function of the observed data and model parameters, and $\pi(\boldsymbol{\theta}|m)$ is the prior probability for the model parameters. The posterior distribution can alternatively be described as a weighted prior, with weights given by

$$w(\boldsymbol{\theta}) = \frac{\mathcal{L}(d|\boldsymbol{\theta}, m)}{\mathcal{L}_{\max}}, \tag{2}$$

where $\mathcal{L}_{\max}$ is the maximum likelihood.[1]

Typically, obtaining a closed-form expression for the posterior probability across the parameter space $\boldsymbol{\theta}$ is not possible. This means that we are not able to trivially evaluate Eq. (1), even if a functional form for the prior distribution is given. It is therefore common to draw samples from the unknown posterior distribution through stochastic sampling techniques, such as nested sampling [42] or Markov-chain Monte Carlo [43]. For the case of nested sampling, a series of contours of increasing likelihood converge, through an iterative process, to the region of high likelihood. Practically, a nested sampling routine draws a series of *live points* and, at each cycle, it stores the live point with the lowest likelihood, and replaces it with a new point drawn randomly from the prior; the new point is accepted through the Matropolis-Hasting's algorithm [67] conditioned on an

[1] Ordinarily the weight is simply $\mathcal{L}(d|\boldsymbol{\theta}, m)$. We use a slightly modified definition to bind the maximum weight to be $w \leq 1$.

increased likelihood. A new contour that encases the current set of live points is generated, and eventually, the contours converge to the regions of the highest likelihood. The stored points, along with their weights, constitute the *samples* drawn from the posterior distribution.

A drawback of nested sampling is that a few thousand live points are evolved in series, meaning that sampling can take thousands of cycles to complete. In addition, while the recovered posterior distribution becomes more accurate as the number of live points is increased, the number of cycles needed to sample from the unknown posterior distribution also increases (the number of cycles scales linearly with the number of live points [68]). Dynamic nested sampling [69,70] enables the number of live points to change throughout the analysis. It has been shown that this can be optimized for PE analyses, allowing for a reduction in run-time by a factor of ∼70 for relatively simple cases. This significant improvement is possible when the region of high probability is contained in a small region of the prior volume (as is typically the case in high-dimensional problems) [69].

In this paper, we introduce an alternate sampling technique that has a major advantage over nested sampling: a significantly reduced wall and CPU time. Our sampler, VARAHA, achieves this by (a) drawing thousands of points from the *relevant regions in the parameter space that contain the posterior probability mass* and (b) intelligently defining a *likelihood threshold*, below which defines a region of the parameter space that can be safely ignored. This approach means that although VARAHA evaluates the likelihood a comparable number of times as nested sampling, it computes the likelihood for a large number of points at once, meaning that the computation can be efficiently vectorized and parallelized over multiple CPUs for enhanced performance. This is in contrast to computing the likelihood for a relatively low number of points at once, as is done in nested sampling. Since evaluating the likelihood is often the most computationally expensive element of PE, VARAHA is able to perform PE in a fraction of the wall and CPU time compared to other conventional samplers.

VARAHA iteratively discards regions of the parameter space that do not contribute to the posterior distribution and restricts attention to the remaining regions of high likelihood through a series of cycles. Subsequent cycles draw points from within only the regions of high likelihood identified in the previous cycle. Once the final volume has been found, the points contained within the final volume are returned, along with weights given by Eq. (2). We pictorially show VARAHA's algorithm in the top row of panels in Fig. 1.

VARAHA defines each volume to contain likelihoods greater than $\mathcal{L}_\star$. This volume is defined as

$$V(\mathcal{L}_\star) = \int \Theta[\mathcal{L}(d|\boldsymbol{\theta}, m) - \mathcal{L}_\star]\mathrm{d}\boldsymbol{\theta} \tag{3}$$

and the posterior probability mass contained within the volume is

$$P(\mathcal{L}_\star) = \int \Theta[\mathcal{L}(d|\boldsymbol{\theta}, m) - \mathcal{L}_\star]p(\boldsymbol{\theta}|d, m)\mathrm{d}\boldsymbol{\theta}. \tag{4}$$

Here $\Theta[x]$ is the Heaviside step function, $\Theta[x] = 1$ for $x \geq 0$ and 0 otherwise. As $\mathcal{L}_\star \to -\infty$, the probability tends to 1 and the volume becomes the full prior volume. However, by choosing $\mathcal{L}_\star$ for which $P(\mathcal{L}_\star)$ is close to, but slightly smaller than, unity the volume can be significantly smaller than the full prior volume. VARAHA defines $\mathcal{L}_\star$ such that the corresponding volume $V(\mathcal{L}_\star)$ contains a probability P_{thr}. This likelihood value is hereafter referred to as the likelihood threshold. By identifying this region and sampling only from it, VARAHA can very efficiently generate samples.

The challenge, then, is to efficiently find the volume that contains a probability P_{thr}, referred to as the *live volume*. Since it is generally not possible to evaluate Eq. (4) analytically to find the appropriate likelihood threshold, other than for simple toy examples, VARAHA computes it numerically and iteratively identifies both the appropriate threshold and the corresponding region of parameter space.

At the beginning of the first cycle, shown in the top left panel of Fig. 1, the live volume is simply the full prior volume. A large number of live points, $N_{\mathrm{live}} = N_{\mathrm{pts}}$, are randomly generated within the live volume (shown in green in the figure), and at each point, the likelihood is evaluated. We can use these points to perform a Monte Carlo integral of Eq. (4) to obtain the required likelihood threshold $\mathcal{L}_{\mathrm{thr}}$ that contains the desired P_{thr}. However, particularly for sharply peaked posteriors in large parameter spaces, it is likely that only a small number of points will contribute significantly to the Monte Carlo integral. Therefore, in addition, we also calculate the likelihood threshold $\mathcal{L}_{N_{\mathrm{min}}}$ such that a minimum number of points, N_{min}, lie above the threshold. If $\mathcal{L}_{\mathrm{thr}} < \mathcal{L}_{N_{\mathrm{min}}}$, we have successfully identified the region which contains P_{thr}. However, if $\mathcal{L}_{N_{\mathrm{min}}} < \mathcal{L}_{\mathrm{thr}}$, we have not sampled the posterior distribution sufficiently densely. In this case, we exclude the regions of parameter space with likelihood below $\mathcal{L}_{N_{\mathrm{min}}}$ and repeat the process.

Selecting N_{min} points with the largest likelihoods identifies a region of high likelihood in the parameter space, which is shown by the orange dots in the top row of panels in Fig. 1. The volume of this region is given by Eq. (3) and is numerically evaluated through the Monte Carlo integration,

$$V \approx \bar{V} \coloneqq V_0 \frac{N_{\mathrm{min}}}{N_{\mathrm{pts}}}, \tag{5}$$

where V_0 is the full prior volume. The uncertainty in the Monte Carlo integration is directly related to the Poisson fluctuations in N_{min}:

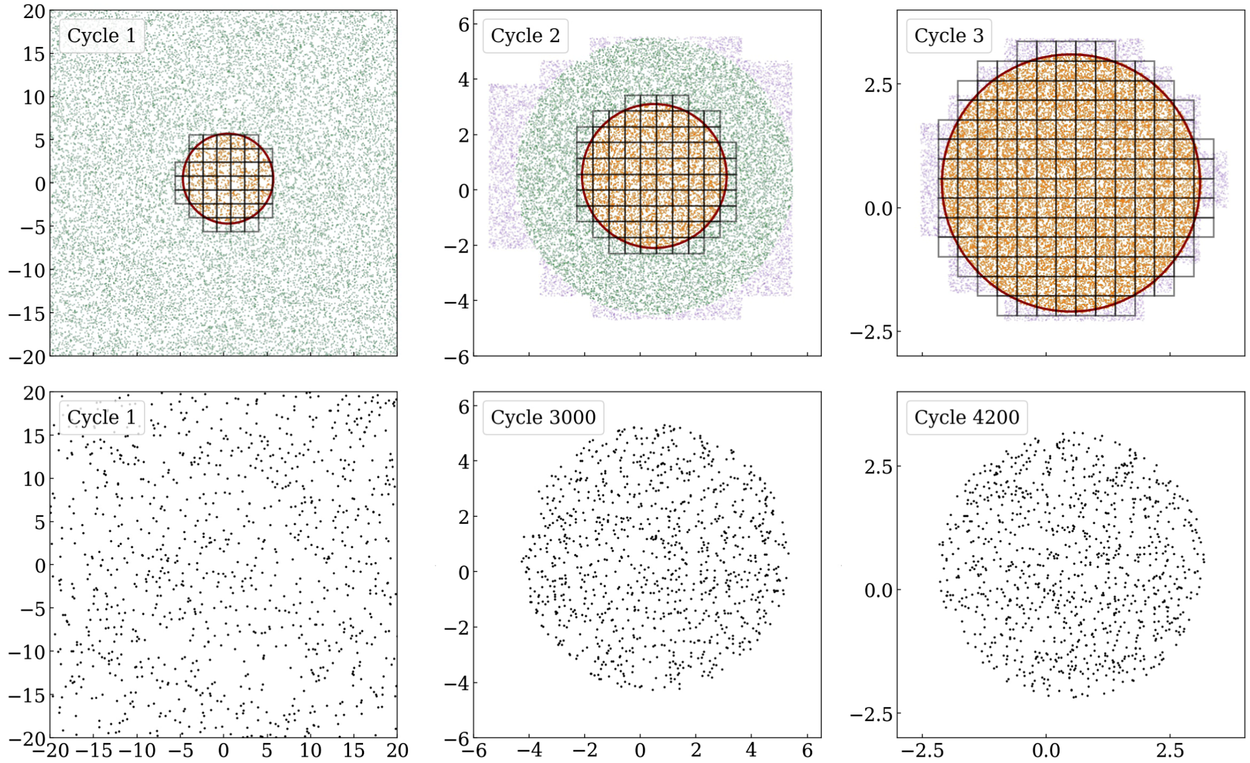

FIG. 1. Top row: a pictorial representation of VARAHA's sampling algorithm for a two-dimensional Gaussian likelihood distribution with mean [0.5, 0.5] and covariance [0.5, 0.5]. The left, middle and right panels show cycle 1, cycle 2 and cycle 3, respectively. The purple dots show the points randomly drawn from within the multidimensional grid and the green dots show the points that have a likelihood larger than the likelihood threshold from the previous cycle (step 1). The orange dots show the points with a likelihood larger than the likelihood threshold calculated at the current cycle (step 2). The red line shows the contour of fixed likelihood equal to the likelihood threshold, representing the live volume, and the black lines show the multidimensional grid that surrounds the live volume (step 5). Bottom row: a pictorial representation of a nested sampling algorithm for the same two-dimensional Gaussian likelihood distribution; for this case, we use the DYNESTY [70] nested sampler. The left, middle and right panels show cycle 1, cycle 3000 and cycle 4200 respectively, and the gray dots show the nested live points.

$$\delta V \approx \delta\bar{V} \coloneqq V_0 \frac{\sqrt{N_{\rm min}}}{N_{\rm pts}} = \frac{\bar{V}}{\sqrt{N_{\rm min}}}. \qquad (6)$$

The set of $N_{\rm min}$ points provide an estimate of the volume and discretely constitute the region of the parameter space enclosed by this volume. With $N_{\rm min}$ chosen to be a few thousand, the error in the estimated volume is the order of a few percent.

The procedure outlined above presents two situations:

(i) *The posterior mass enclosed by $\mathcal{L}_{N_{\rm min}}$ is larger than* $P_{\rm thr}$.—Starting from the full prior volume, the live volume reduces to the volume enclosed by $\mathcal{L}_{N_{\rm min}}$ (bounded by the red contour in Fig. 1). The fluctuation in $\mathcal{L}_{N_{\rm min}}$, due to sampling with a finite number of points, will have a negligible impact on the posterior mass contained by it. The largest fluctuation occurs for a uniform distribution and is equal to the Monte Carlo (MC) errors of a few percent in the volume. However, in general, the likelihood distribution is expected to decay from one or several maxima.

(ii) *The posterior mass enclosed by $\mathcal{L}_{N_{\rm min}}$ is smaller than* $P_{\rm thr}$.—This occurs when there are a large number of points with non-negligible weights, $n > N_{\rm min}$, and a threshold smaller than $\mathcal{L}_{N_{\rm min}}$ is required to enclose the posterior mass, $P_{\rm thr}$. To estimate n, VARAHA calculates the empirical distribution function (EDF), $\hat{F}(\mathcal{L}_\star)$, which simply gives the fraction of points with likelihood less than $\mathcal{L}_\star$.[2] The final likelihood threshold is then calculated as $\mathcal{L}_{\rm thr} \coloneqq \hat{F}^{-1}(P_{\rm thr})$.

VARAHA evolves the first situation (i) using multiple cycles of MC integration until it reaches the second situation (ii). Each cycle performs the integration using live volumes, which themselves decrease as cycles progress. A generalized form of Eq. (3) for this progression is given by

[2] We really wish to calculate the cumulative distribution function (CDF). However, since we only have discrete samples, the true CDF is unknown. We, therefore, approximate the CDF by computing the EDF and taking a conservative limit to ensure that we enclose *at least* the desired probability. A bound on the difference between the EDF and CDF is given, e.g., by using the Dvoretzky-Kiefer-Wolfowitz-Massart inequality [71].

$$V_i \approx \bar{V}_i \coloneqq \frac{N_{\min}}{N_{\text{live}}^{i-1}} \bar{V}_{i-1}. \tag{7}$$

For the second and consecutive cycles, VARAHA samples from the live volume and ignores the remaining parameter space. Since we only store samples contained within the volume and not the volume itself, the structure of the live volume is not known. VARAHA navigates this by (a) generating a multidimensional grid covering the full parameter space, and (b) selecting those hypercubes that contain at least one of the live points from the previous cycle. Once the relevant hypercubes have been identified, the live volume has been reconstructed. N_{pts} points are now uniformly scattered within the reconstructed live volume (the purple points in the top middle panel of Fig. 1). The likelihood of all the points is calculated and only those with a likelihood above the threshold are kept (the green points in the top middle panel of Fig. 1). Since the same number of points are now scattered within a much smaller volume, VARAHA is able to increase the number of live points contained within the live volume and thus setting the stage to perform the next MC integration.

It is important to choose an appropriate spacing for the multidimensional grid. If the grid spacing is too large, the reconstructed live volume is much larger than the value estimated in Eq. (5). If the grid spacing is too small, the reconstructed live volume will not include relevant regions of the parameter space that did not get sampled due to random fluctuations in the location of points. VARAHA constructs the multidimensional grid over the full parameter space, requiring that the volume of each hypercube in the grid is equal to the error in the estimated volume $\delta\bar{V}$. We motivate this choice as follows: the chosen hypercube volume is $\sqrt{N_{\min}}$ times larger than the average volume of $\bar{V}/N_{\min}$ approximately occupied by each live point. This choice ensures that the volume will have a maximum uncertainty of $\delta\bar{V}$ if this uncertainty arises due to Poisson fluctuation near a single live point. In this case, the multidimensional grid is expected to enclose most of V even though constructed using information gained from the MC volume $\bar{V}$. A uniform grid spacing in all dimensions, therefore, leads to the number of bins per dimension: $N_{\text{bins}} = (V_0/\delta\bar{V})^{(1/N_{\text{dim}})}$, where N_{dim} is the dimensionality of the parameter space.[3]

The iterative process described above gradually increases the value of the likelihood threshold and discards the uninteresting regions of the parameter space. However, a very small amount of posterior mass is also lost in the process. A new likelihood threshold approximately increases the discarded posterior mass by an amount,

$$1 - \frac{\sum_j w_j \Theta[\mathcal{L}_j - \mathcal{L}_{N_{\min}}]}{\sum_j w_j}, \tag{8}$$

where j identifies the samples that are enclosed by the live volume in the current cycle, and $\mathcal{L}_{N_{\min}}$ is the new likelihood threshold calculated for the next cycle. This corresponds to the posterior mass contained between two concentric circles in Fig. 1. In practice, when deciding if $\mathcal{L}_{N_{\min}} > \mathcal{L}_{\text{thr}}$, by calculating the EDF, we require the discarded posterior mass not to accumulate to more than $1 - P_{\text{thr}}$. This implicitly ensures that the posterior mass enclosed by the live volume is at least P_{thr}. VARAHA does not evolve the likelihood threshold any further if the discarded posterior mass becomes very close to $1 - P_{\text{thr}}$. Even though the likelihood threshold no longer evolves, the number of bins used to create the multidimensional grid continues to increase as the number of samples inside the live volume also increases.

Cycles are terminated once the desired accuracy is obtained. The stopping condition could be determined based on a fixed number of weighted samples or a fixed number of cycles. At the end of the analysis, VARAHA returns a set of weighted samples, where the weight of each sample is determined by Eq. (2). All the samples have a likelihood value greater than the final value of $\mathcal{L}_{\text{thr}}$. Samplers that employ MCMC methods typically return a set of unweighted samples, with sample weights equal to 1. It is possible to generate these samples from the weighted samples by performing rejection sampling on the weighted samples (see the Appendix for details). The final set of unweighted samples has a sample size that is approximately $\sum_i w_i$ [72]. However, since $n_{\text{eff}} \geq \sum_i w_i$ [73], with equality only if all the weights are equal to 1, the rejection sampling process leads to a reduction in the information contained in the samples.

A. Implementation

VARAHA implements this algorithm as follows:

(1) *Sprinkle points within the multidimensional grid.*—(a) N_{pts} points are uniformly drawn from the reconstructed live volume and the likelihood for all points is calculated. (b) Live points with a likelihood larger than the threshold from the previous cycle are then identified. If this is the first cycle, N_{pts} points are uniformly distributed within the entire prior volume and all points are kept, meaning that $N_{\text{pts}} = N_{\text{live}}$ (this is equivalent to setting the likelihood threshold from the *previous cycle* to negative infinity).

(2) *Calculate the likelihood threshold.*—The likelihood that accumulates no more than $1 - P_{\text{thr}}$ of the truncated posterior mass up to the current cycle, $\mathcal{L}_{\text{thr}}$, is calculated using the N_{live} live points that are enclosed by the live volume. A second threshold,

[3] In some problems, there is significantly more structure in some dimensions of parameter space than others. Then, it is desirable to employ a grid with different numbers of bins in each dimension. The challenge, however, is to derive requirements for the relative number of bins in each dimension.

$\mathcal{L}_{N_{\mathrm{min}}}$, which ensures that N_{min} points lie above the threshold, is also calculated. The final likelihood threshold is chosen to be the minimum of these two values.

(3) *Calculate the volume of the live volume*.—The volume and uncertainty in the live volume is calculated through MC integration; see Eq. (5).

(4) *Calculate effective sample size*.—All points from the current or previous cycles that cross the current likelihood threshold are stored as weighted samples. The number of effective samples is calculated using

$$n_{\mathrm{eff}} = \frac{(\sum_i w_i)^2}{\sum_i (w_i^2)}, \tag{9}$$

where w_i is the weight of each sample, defined in Eq. (2) [74].

(5) *Reconstruct the live volume*.—A multidimensional grid is created that spans the whole parameter space and hypercubes that register at least one live point are kept. The resulting hypercubes reconstruct the live volume.

(6) *Repeat*.—Steps 1 to 5 are repeated until a stopping criterion is reached. Example stopping criteria include terminating once a specific number of effective samples have been obtained or terminating after a specific number of cycles have elapsed. The final output is then a set of weighted samples, where the weights are given by the likelihood evaluated at the sample, normalized by the maximum likelihood; see Eq. (2).

B. Sampler settings

The number of samples drawn from the uniform distribution N_{pts}, the posterior mass required to be contained within the live volume, P_{thr}, and the minimum number of points to retain, N_{min}, are the sampler parameters that the user is free to specify. In our testing, we found that the choice of N_{pts} has only a small effect on the recovered posterior probability. Although reducing N_{pts} decreases the number of computations in each cycle, it increases the number of cycles needed to obtain the required effective number of samples. We find that choosing N_{pts} to be in the range 100,000–1,000,000 provides a good compromise between these. Similarly, N_{min} sets the minimum fractional error on the estimated MC volume. In our testing, we found that as the dimensionality of the problem increases, N_{min} should increase correspondingly. Depending on the complexity, N_{min} within the range of 1000–10,000 is adequate for most distributions with dimensionality between 2 and 8.

As VARAHA only keeps points that cross a specific likelihood threshold, defined through P_{thr}, we deliberately discard part of the parameter space with low likelihood during each cycle. This can have an impact on the recovered posterior distribution if P_{thr} is too small. In contrast, if the threshold is chosen to be too large, then we exclude a limited region of parameter space, which leads to a significant increase in analysis time for limited benefit. In our testing, we found that $P_{\mathrm{thr}} = 0.999$ is sufficient for most cases.

C. Comparison with nested sampling

In Fig. 1, we compare the convergence of VARAHA (top panels) and a nested sampling algorithm (DYNESTY[4] [70], bottom panels) for a simple two-dimensional Gaussian likelihood distribution. We used the static version of DYNESTY with default settings, but increased the number of live points to 1000, as this is similar to the number of live points used for gravitational-wave analyses, see, e.g., [39]. For this example, VARAHA used $N_{\mathrm{pts}} = 20{,}000$, $N_{\mathrm{min}} = 1000$ and $P_{\mathrm{thr}} = 0.999$.

As expected, VARAHA rapidly converges to the region of high likelihood (within three cycles), while DYNESTY requires significantly more cycles to constrain to a similar region of the parameter space (~4200 cycles). The significant reduction in the number of cycles of VARAHA, compared to DYNESTY, is primarily caused by the likelihood threshold: we see that in VARAHA's first cycle, we constrain the high likelihood region to within a circle of radius 5, while DYNESTY takes ~3000 cycles to constrain to a comparable region of the parameter space. VARAHA further constrains the high likelihood region to within a circle of radius 2.5 within three cycles, while DYNESTY takes ~4200 cycles to obtain a similar constraint. This culminates in a significantly reduced wall and CPU time. VARAHA also obtains a larger number of effective samples, with $n_{\mathrm{eff}} \sim 6000$ compared to $n_{\mathrm{eff}} \sim 4000$ for DYNESTY.

D. The evolution of the multidimensional grid

One of the key features of VARAHA is that it is able to draw points from within the live volume without having to know the structure. It achieves this by generating a multidimensional grid that covers the full parameter space and registering hypercubes that contain at least one point with likelihood above the previous likelihood threshold. To demonstrate this in practice, we analyze a complex two-dimensional distribution and explicitly show how the multidimensional grid is constructed and how it converges to the region of high likelihood. The chosen distribution has multiple disconnected regions of equal likelihood. While this example is only two dimensional, the disconnected peaks in likelihood can prove challenging to identify. Since this distribution is analytically known, it provides a good illustration of VARAHA. For this example, we used

[4] We use DYNESTY = 1.0.1, as this is the version in the International Gravitational-Wave Observatory Network (IGWN) Conda environment (https://computing.docs.ligo.org/conda/) at the time of writing.

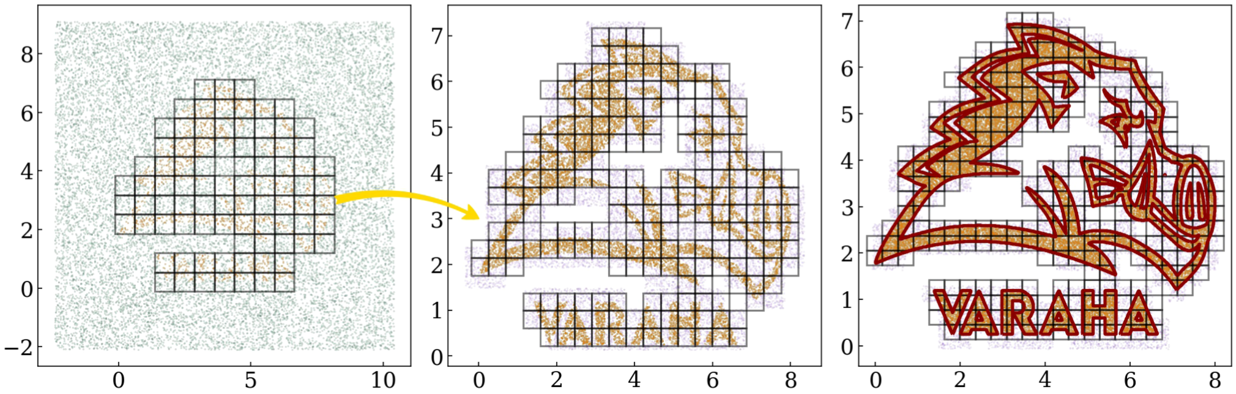

FIG. 2. Plot showing the evolution of the multidimensional grid that surrounds the live volume, when VARAHA samples from a complex two-dimensional distribution with a constant density. The purple dots show the points randomly drawn from within the multidimensional grid and the green dots show the points that have a likelihood larger than the likelihood threshold from the previous cycle (step 1). The orange dots show the points with a likelihood larger than the likelihood threshold calculated at the current cycle (step 2) and the black lines show the multidimensional grid that surrounds the live volume (step 5). The left, middle and right panels show to the first, third and fifth cycles, respectively. In the right panel, the red line shows the contour of fixed likelihood equal to the likelihood threshold, representing the final live volume.

$N_{\mathrm{pts}} = 20{,}000$, $N_{\mathrm{min}} = 1000$ and we terminated VARAHA once five cycles have elapsed.

In Fig. 2 we show the evolution of the multidimensional grid. We see that in the first cycle, VARAHA is able to identify the rough location of the high-likelihood region and construct a coarse multidimensional grid that surrounds the entire volume. As VARAHA progresses, we see that the number of hypercubes increases and the multidimensional grid converges to the complex two-dimensional distribution with gaps appearing between adjacent hypercubes where there is little to no probability support. Unlike in Fig. 1, we do not see any green dots beyond the first cycle. This is because VARAHA rapidly identifies the high-likelihood region (owing to a constant density throughout) in the first cycle and maintains a constant likelihood threshold for all subsequent cycles. This implies that all points that cross the previously defined likelihood threshold (green dots) are used as live points for the current cycle (orange dots).

E. Example: Bimodal multivariate Gaussian distribution

Next, we showcase the full VARAHA sampling algorithm. We chose to analyze a six-dimensional bimodal multivariate Gaussian distribution and compare results with existing nested and Markov-chain Monte Carlo samplers. We also explicitly show how the likelihood threshold evolves as the number of cycles increases. For this example, we terminated VARAHA once 20,000 effective samples were collected. During our sampling we used $P_{\mathrm{thr}} = 0.999$, $N_{\mathrm{pts}} = 400{,}000$ and $N_{\mathrm{min}} = 1000$.

We analyzed an asymmetric multivariate Gaussian distribution with each mode's mean and covariance randomly chosen: the mean of the first and second modes are randomly chosen between $[-1, 1]$ and $[-5, -3]$, respectively, and we use a covariance matrix that is obtained by applying an inverse Wishart distribution [75] to a diagonal matrix with elements randomly chosen between 1 and 2.

The output of VARAHA is shown in Table I. The prior volume spans from −20 to 20 in each dimension. As VARAHA converges to the region of high likelihood, the number of bins steadily increases over the cycles, reflecting the decreased uncertainty in the recovered volume. As expected, the likelihood threshold monotonically increases from −170 to −13.6 as the number of cycles increases. The likelihood threshold for the first three cycles is set by $\mathcal{L}_{N_{\mathrm{min}}}$ with subsequent cycles using $\mathcal{L}_{\mathrm{thr}}$. For the fourth cycle and beyond, the likelihood threshold remains fixed, which reflects the fact that the volume enclosing P_{thr} of the posterior mass has been found. The number of effective samples is relatively low in the first few cycles since the majority have $\mathcal{L}_i \ll \mathcal{L}_{\mathrm{max}}$, but increases steadily as the likelihood threshold increases. VARAHA obtains just over

TABLE I. Output from VARAHA showing the evolution of the number of bins in each dimension N_{bins}, the likelihood threshold in each cycle, and the number of effective samples n_{eff} from the multivariate Gaussian example. The likelihood threshold in each cycle is set by either $\mathcal{L}_{N_{\mathrm{min}}}$ or $\mathcal{L}_{\mathrm{thr}}$ depending on the situation (see text for details); for this case, the first three cycles are set by $\mathcal{L}_{N_{\mathrm{min}}}$ and subsequent cycles are set by $\mathcal{L}_{\mathrm{thr}}$. The sampled distribution is a bimodal multivariate Gaussian and VARAHA collected more than 20,000 effective samples across 13 cycles in 30 seconds on a single CPU thread.

Cycle	N_{bins}	Log-likelihood threshold	n_{eff}
1	4	−170.7	3
3	12	−26.9	6
4	19	−13.6	21
8	28	−13.6	2896
13	32	−13.6	22323

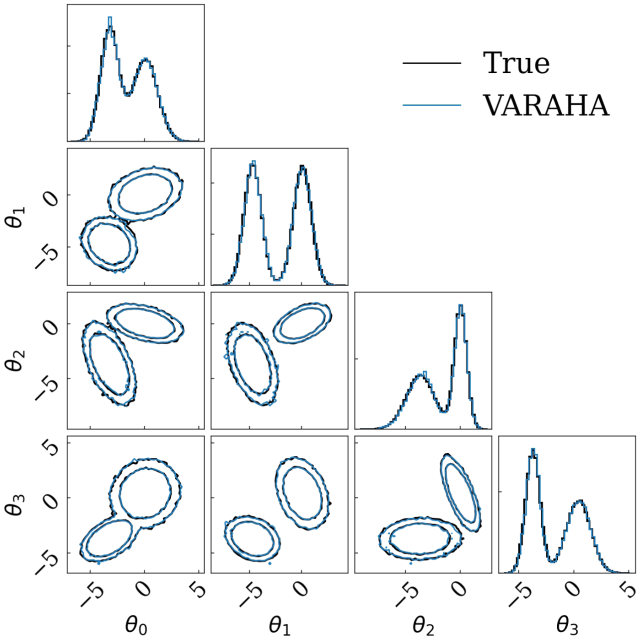

FIG. 3. Corner plot [76] showing the posterior PDF obtained by VARAHA and the true bimodal multivariate Gaussian distribution across four of the six dimensions. The two-dimensional contours show the 1.5σ and 2.5σ confidence levels and the individual histograms on the leading diagonal show the data marginalized to a particular dimension. VARAHA completes sampling in 30 seconds on a single CPU thread.

23,000 effective samples from a total of approximately 800,000 weighted samples. Of course, rejection sampling could be used to obtain a set of unweighted samples; however, this would result in a smaller final sample size. We see that a larger number of cycles is needed than in the previous example, which is a result of the higher dimensionality of the likelihood surface.

In Fig. 3, we plot the posterior distribution obtained with VARAHA and the known analytic distribution. We see that VARAHA recovers the true distribution to high accuracy, with the mean and widths of each mode correctly identified. For comparison, we also analyzed the same analytic distribution with two external samplers: DYNESTY [70], a nested sampler, and Bilby MCMC [41], a Markov-chain Monte Carlo sampler, both operated through the Bilby infrastructure [38].[5] In our testing, we found that VARAHA finished sampling in 30 seconds on a single CPU: at least 60× faster than either DYNESTY or Bilby MCMC. Although the run times for both DYNESTY and Bilby MCMC can vary significantly depending on the chosen settings, it is unlikely that either sampler can obtain a comparable number of posterior samples as VARAHA in 30 seconds that accurately recovers the mean and widths of each mode, when using a single CPU.

III. APPLICATION TO PARAMETER ESTIMATION OF GRAVITATIONAL WAVES

In this section, we demonstrate the application of VARAHA to the PE of gravitational wave signals. We focus only on gravitational waves originating from compact binary mergers and compare results to those obtained with LALInference, the software that has regularly been used since the first gravitational-wave detection in 2015 [17] (see also [36–40,44,45,50,51,56,77,78]). For this article, we restrict attention to quasicircular binaries with spins aligned with the orbital angular momentum, referred to as an aligned-spin binary, meaning that the binary does not precess [30]. In addition, we focus only on the leading (2,2) harmonic of the waveform, which is typically the most significant contribution to the observed GW signal [79]. In this case, the binary parameters can be cleanly decomposed into intrinsic parameters, which determine the properties of the binary, and extrinsic parameters, which determine the location and orientation of the binary relative to the earth. Throughout this work, we use the IMRPhenomD [80,81] gravitational-wave model to evaluate the likelihood since it is optimized for aligned-spin binary systems that contain only the leading (2,2) harmonic (see also [82–85]).

For an aligned-spin model, the intrinsic parameters primarily affect the amplitude and phase evolution of the gravitational wave, and the extrinsic parameters only affect the overall amplitude, phase and time of arrival of the signal at each of the detectors. Table II lists the extrinsic and

TABLE II. GW signal parameters sampled by VARAHA. We group extrinsic parameters, except coalescence time, into one variable $\boldsymbol{\Omega}$ (top section), and all the intrinsic parameters into $\boldsymbol{\theta}$ (bottom section). The component masses m_1 and m_2 are characterized by the mass ratio $q = m_2/m_1$ and chirp mass $\mathcal{M} = (m_1 m_2)^{(3/5)}/(m_1 + m_2)^{(1/5)}$. The masses are measured in the frame of the detector from the signal that has already suffered cosmological redshift.

Label	Description
α	Right ascension of the source
δ	Declination of the source
d_L	Luminosity distance of the source
ι	Inclination angle
ψ	Polarization angle
ϕ_c	Coalescence phase
t_c	Coalescence time in the reference detector
$\mathcal{M}$	Detector frame chirp mass
q	Mass ratio defined to be less than 1
χ_1	First aligned spin component
χ_2	Second aligned spin component

[5] For both DYNESTY and Bilby MCMC, we used the default settings provided by Bilby with some modifications to ensure reasonable convergence. Both samplers used a single CPU. For DYNESTY, we used 1000 live points and a nact (the number of autocorrelation times before accepting a point) of 5. For Bilby MCMC, we used four temperature chains and we specified that we would like to obtain 10,000 independent samples. VARAHA made five million likelihood calculations over thirteen vectorized cycles. In comparison, nested sampling made 2.5 million likelihood calculations over 25,000 iterations.

intrinsic parameters of the system estimated by VARAHA. VARAHA allows separable sampling of extrinsic and intrinsic parameters and breaks one large dimensional problem into two small ones. We note that other analyses, e.g., RapidPE/RIFT [44,45], employ a similar methodology as used by VARAHA. This framework is extensible to include higher harmonics in the gravitational-wave signal and additional signal parameters, such as in-plane spins, which lead to additional physical effects in the waveform, and can reduce biases in the recovered PE [4,86–92]. It will require accounting for the morphological dependence of a signal on the extrinsic parameters. This extension will be investigated in future work.

A. Factorizing the likelihood

For a network of gravitational-wave detectors (e.g. [93–95]), the probability that the observed detector data contains a gravitational wave signal from a coalescing binary and with parameters, $(t_c, \mathbf{\Omega}, \boldsymbol{\theta})$, is given by the Bayes formula as

$$p(t_c, \mathbf{\Omega}, \boldsymbol{\theta}|\vec{d}) = \frac{\mathcal{L}(\vec{d}|t_c, \mathbf{\Omega}, \boldsymbol{\theta})\pi(t_c, \mathbf{\Omega}, \boldsymbol{\theta})}{\{\vec{d}\}}, \qquad (10)$$

where $\vec{d} \equiv \{d_1, d_2, \ldots\}$ represents the strain data observed in each gravitational-wave detector, t_c is the coalescence time in the reference detector, $\mathbf{\Omega}$ denotes the remaining extrinsic parameters, and $\boldsymbol{\theta}$ are the intrinsic parameters. Under the assumption that the data in each detector is independent, stationary, Gaussian and containing a gravitational wave signal, a Gaussian likelihood, $\mathcal{L}(\vec{d}|t_c, \mathbf{\Omega}, \boldsymbol{\theta})$, is constructed as

$$\begin{aligned} \log(\mathcal{L}(\vec{d}|t_c, \mathbf{\Omega}, \boldsymbol{\theta})) &= c\sum_{i\in\mathrm{dets}} \langle d^i - h^i | d^i - h^i\rangle \\ &= c\sum_{i\in\mathrm{dets}} (\langle d^i|d^i\rangle - \langle d^i|h^i\rangle \\ &\quad - \langle h^i|d^i\rangle + \langle h^i|h^i\rangle), \end{aligned} \qquad (11)$$

where $c = (2\pi)^{-k/2}$ for a k dimensional distribution and $h^i \equiv h^i(t_c, \mathbf{\Omega}, \boldsymbol{\theta})$ is the expected GW signal in the ith detector. The term $\{\vec{d}\}$ is the marginal likelihood (or evidence),

$$\{\vec{d}\} = \int \mathcal{L}(\vec{d}|t_c, \mathbf{\Omega}, \boldsymbol{\theta})\pi(t_c, \mathbf{\Omega}, \boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}\mathrm{d}\mathbf{\Omega}\mathrm{d}t_c. \qquad (12)$$

In Eq. (11), the *complex* noise weighted inner product between two time-domain functions $\langle a(t)|b(t)\rangle$ is defined in the frequency domain as

$$\langle a(t)|b(t)\rangle = 4\int_{f_{\min}}^{f_{\max}} \frac{\tilde{a}(f)\tilde{b}(f)^\star}{S(f)}\mathrm{d}f, \qquad (13)$$

where $S(f)$ is the power spectrum of the detector noise [33], $\tilde{a}(f)$ and $\tilde{b}(f)$ are the Fourier transforms of $a(t)$ and $b(t)$, respectively, and star represents the complex conjugate. Equation (13) can be evaluated at an arbitrary time shift by using the convolution theorem and taking the inverse Fourier transform [96],

$$\langle a(t)|b(t+\Delta t)\rangle = 4\int_{f_{\min}}^{f_{\max}} \frac{\tilde{a}(f)\tilde{b}(f)^\star}{S(f)}\mathrm{e}^{2\pi i\Delta t}\mathrm{d}f. \qquad (14)$$

Consequently, the likelihood can be evaluated for a fixed set of intrinsic parameters, from a single inner product calculation.

Returning to Eq. (11), the first term, $\langle d_i|d_i\rangle$, is independent of the parameters $\mathbf{\Omega}$ and $\boldsymbol{\theta}$ and is therefore constant. Since it will be absorbed in the normalization for Eq. (10) we neglect it in what follows. The final term is the inner product of the expected signal in the ith detector with itself,

$$\langle h^i(t_c, \mathbf{\Omega}, \boldsymbol{\theta})|h^i(t_c, \mathbf{\Omega}, \boldsymbol{\theta})\rangle =: \varrho^i(\mathbf{\Omega}, \boldsymbol{\theta})^2. \qquad (15)$$

If $h(\boldsymbol{\theta})$ is the signal arriving from a binary with intrinsic parameters $\boldsymbol{\theta}$ at a unit distance, from overhead the detector and with the orbital plane facing the observer, the relative amplitude of a signal from a given distance d_L sky location and orientation is determined by the effective distance, given by

$$D^i_{\mathrm{eff}} = d_L\Big/\sqrt{\left[F^i_+{}^2\left(\frac{1+\cos^2\iota}{2}\right) + F^{\times 2}_i\cos^2\iota\right]}, \qquad (16)$$

where the detector response functions $F^i_{+,\times}$ depend upon the sky location, polarization and time of arrival of the source [97]. Thus, once

$$\varrho^i_o(\boldsymbol{\theta})^2 = \langle h(\boldsymbol{\theta})|h(\boldsymbol{\theta})\rangle \qquad (17)$$

has been calculated, it is straightforward to evaluate ϱ^i as

$$\varrho^i(\mathbf{\Omega}, \boldsymbol{\theta}) = \frac{\varrho^i_o(\boldsymbol{\theta})}{D^i_{\mathrm{eff}}}. \qquad (18)$$

Finally, we turn to the two middle terms in Eq. (11) which constitute the inner product of the data d^i with the expected signal. We have already seen that the variation of the signal amplitude can be encoded in the effective distance D^i_{eff}. Similarly, the phase of the signal observed in detector i is given by [98]

$$\phi^i_{\mathbf{\Omega}} = \phi^i - 2\phi_c, \quad \text{where } \phi^i = \tan^{-1}\left(\frac{F^i_\times}{F^i_+}\frac{2\cos\iota}{1+\cos^2\iota}\right), \qquad (19)$$

and the time of arrival is given by

$$t^i = t_c + \Delta t^i(\mathbf{\Omega}, t_c), \tag{20}$$

where Δt^i depends upon the location of the source relative to the detectors.

Since the amplitude and phase evolution of the signal is unchanged by the extrinsic parameters, we can calculate the inner product for any set of extrinsic parameters by rescaling and time shifting the inner product time series for the reference waveform $h(\boldsymbol{\theta})$. Thus, the inner product of the data d^i with the expected signal is given by

$$\begin{aligned}\langle d^i | h^i(t_c, \mathbf{\Omega}, \boldsymbol{\theta})\rangle &= \left\langle d^i \middle| \frac{h^i(\boldsymbol{\theta}, t^i)\exp(i\phi^i_{\mathbf{\Omega}})}{D^i_{\mathrm{eff}}} \right\rangle \\ &= \frac{\varrho^i_o(\boldsymbol{\theta})\rho^i(\boldsymbol{\theta}, t^i)\exp(i\phi^i_{\mathbf{\Omega}})}{D^i_{\mathrm{eff}}},\end{aligned} \tag{21}$$

where we have defined the signal to noise ratio (SNR) for the template with intrinsic parameters $\boldsymbol{\theta}$ as

$$\rho^i(\boldsymbol{\theta}, t_c) = \frac{\langle d_i | h(\boldsymbol{\theta}, t^i)\rangle}{\sqrt{\langle h(\boldsymbol{\theta}) | h(\boldsymbol{\theta})\rangle}}. \tag{22}$$

The third term in Eq. (11) is simply the complex conjugate of (21).

Combining these expressions, the resulting log-likelihood from Eq. (11) assumes the form

$$\begin{aligned}&\log(\mathcal{L}(\vec{d}|t_c, \mathbf{\Omega}, \boldsymbol{\theta})) \\ &= \sum_{i\in\mathrm{dets}}\left[-\frac{1}{2}\frac{\varrho^i_o(\boldsymbol{\theta})^2}{(D^i_{\mathrm{eff}})^2} + \frac{\varrho^i_o(\boldsymbol{\theta})|\rho^i(\boldsymbol{\theta}, t^i)|}{D^i_{\mathrm{eff}}}\cos(\Delta\phi^i)\right],\end{aligned} \tag{23}$$

where $\Delta\phi^i$ is the difference between the measured phase in detector i and the expected phase, given the parameters of the signal:

$$\Delta\phi^i = \arg(\rho^i(\boldsymbol{\theta}, t^i)) - \phi^i_{\mathbf{\Omega}}. \tag{24}$$

The log-likelihood is maximized for a given set of intrinsic parameters when $D^i_{\mathrm{eff}} = \varrho^i(\boldsymbol{\theta})/\rho^i(\boldsymbol{\theta}, t^i_c)$, and the cosine term in the last equation is unity. However, as the cosine term is solely dependent on the phase acquired due to the extrinsic parameters and $\rho_i(t^i_c)$ is dependent on the arrival times in different network detectors, the maximization puts a time-phase constraint [99].

The expression in Eq. (23) clearly separates the likelihood dependence on the intrinsic parameters $\boldsymbol{\theta}$ from the extrinsic parameters $\mathbf{\Omega}$ and the coalescence time t_c. In particular, the effective distance, D^i_{eff}, phase, $\phi^i_{\mathbf{\Omega}}$ and time of arrival t^i in each detector depends only upon the extrinsic parameters $\mathbf{\Omega}$ and the time of arrival t_c. The SNR time series associated with the fiducial waveform $h(\boldsymbol{\theta})$, and its overall normalization, is dependent only on the intrinsic parameters $\boldsymbol{\theta}$. However, the specific time at which to evaluate the SNR does depend upon the intrinsic parameters through Δt^i. In the following sections, we make repeated use of this splitting of the likelihood to independently estimate the intrinsic and extrinsic parameters. In particular, the most time-consuming step is the generation of simulated waveform and evaluation of the SNR time series. Thus, by writing the likelihood in the form of Eq. (23), it becomes clear that the entire extrinsic parameter space, for a fixed set of intrinsic parameters, can be explored with a single evaluation of the SNR time series.

B. Extrinsic parameters

VARAHA starts by first fixing the intrinsic parameters $\boldsymbol{\theta}$ to a reference waveform and samples the posterior distribution for the seven extrinsic parameters, $\mathbf{\Omega}$ and t_c. BAYESTAR [35], a rapid, non-Markovian sky localization algorithm commonly used by the LIGO, Virgo and KAGRA collaborations (see e.g. [66]), also samples the extrinsic parameters by fixing the intrinsic parameters to a reference waveform. In this subsection, we explain how VARAHA varies from BAYESTAR and demonstrate that it is able to compete with BAYESTAR's performance.

In the initial evaluation, it is natural to use the values for the intrinsic parameters, $\boldsymbol{\theta}_o$, which are reported by the search analysis that identified the signal [100–103]. The detector which has the largest value of $\rho^i(\boldsymbol{\theta}_o)$ in the network is chosen as the reference detector. We then perform PE over the seven-dimensional extrinsic parameter space using the method described in Sec. II. The first step involves scattering millions of points across the extrinsic parameter space. Since five dimensions are angles, it is natural to cover the full range of possible values:

$$\begin{aligned}\alpha &= [0, 2\pi] \\ \sin(\delta) &= [-1, 1] \\ \cos(\iota) &= [-1, 1] \\ \phi_c &= [0, 2\pi] \\ \psi &= [0, 2\pi].\end{aligned} \tag{25}$$

The initial choice of bounds for the coalescence time and luminosity distance requires more care. We wish to determine the narrowest ranges of t_c and d_L that will ensure we chose a range that encloses the posterior mass. If the initial choice is too narrow, then we risk missing part of the relevant parameter space and if it is too broad this will lead to unnecessary exploration of uninteresting regions of the parameter space which will increase the analysis time.

To fix the range of coalescence time, we restrict attention to the reference detector (the one with the largest SNR). We assume that the extrinsic parameters are chosen to maximize the likelihood contribution from the reference detector, i.e. that $D^i_{\mathrm{eff}} = \varrho^i(\boldsymbol{\theta_o})/|\rho^i(\boldsymbol{\theta_o}, t_c)|$ and that the phase $\Delta\phi_i = 0$. In that case, the likelihood contribution from the reference detector is $\frac{1}{2}|\rho^i(\boldsymbol{\theta_o}, t_c)|^2$. This is normally distributed in t_c. Furthermore, from the GW search result, we know the time t_o which gives the maximum SNR, $\rho^i(\boldsymbol{\theta_o}, t_o)$. As we vary the coalescence time t_c^i in the reference detector, the observed SNR will be reduced and, consequently, the maximum contribution of the reference detector to the likelihood will be reduced. The initial range of coalescence times is chosen so that the boundary is at least 4-σ, i.e. we require

$$\frac{1}{2}[|\rho^i(\boldsymbol{\theta_o}, t_o)|^2 - |\rho^i(\boldsymbol{\theta_o}, t_c)|^2] < 4^2/2. \tag{26}$$

So far we have restricted attention to a reference detector. Let us assume that there exists a set of extrinsic parameters which is a good fit to the data in all detectors, i.e. $D_{\rm eff}^i \approx \varrho^i(\boldsymbol{\theta_o})/|\rho^i(\boldsymbol{\theta_o}, t_o)|$ and $\Delta\phi^i \approx 0$. Then, as we vary the coalescence time t_o, *at best* we will find a set of extrinsic parameters that matches the data in all detectors other than the reference, while in the reference detector, the time is offset from the observed peak. Thus, the loss in likelihood in the reference detector gives an (approximate) lower limit on the loss in the network likelihood.

We restrict the initial range of allowed coalescence times that satisfies Eq. (26). Figure 4 plots an example for the bounds on $\rho(t_c)$ estimated for the data from the LIGO Hanford detector corresponding to the signal GW151226 [65]. The discrete values of $\rho(t_c)$ are obtained by taking a discrete inverse Fourier transform of the frequency domain inner product between data and waveform, sampled at 2048 Hz. A smooth function is obtained by interpolating using a cubic spline.

The initial range of distances is chosen to ensure that the chosen range encloses the posterior mass. We also require the bounds to be as small as possible to prevent VARAHA from sampling in regions of parameter space with low likelihood. Returning to Eq. (15), we see that the effective distance is always equal to or greater than the luminosity distance, with equality only for face on systems ($\cos\iota = \pm 1$), lying either directly above or below the detector ($\sqrt{{F_+^i}^2 + {F_\times^i}^2} = 1$). We choose the maximum distance to be 3 times the effective distance in the reference detector:

$$d_L^{\rm max} = 3D_{\rm eff}^i = 3\varrho^i(\boldsymbol{\theta_o})/|\rho^i(\boldsymbol{\theta_o}, t_o)| \tag{27}$$

as well as choosing a minimum distance of 0. At a distance, $d_L^{\rm max}$, the maximum possible likelihood in the reference detector occurs for a face-on, overhead system. In that case, the log-likelihood is reduced by an amount $\frac{2}{9}|\rho^i(\boldsymbol{\theta_o}, t_o)|^2$. At an SNR of 7 in the reference detector, which corresponds to a relatively weak signal, this leads to a reduction in the log-likelihood of 11. However, as with the discussion of the coalescence time, it is likely that there will also be a reduction in the likelihood in the other detectors, meaning this is a lower limit on the loss in the network likelihood. As often is the case, instead of the distance, we use a uniform prior on the volume,

$$p(d_L) \propto d_L^2. \tag{28}$$

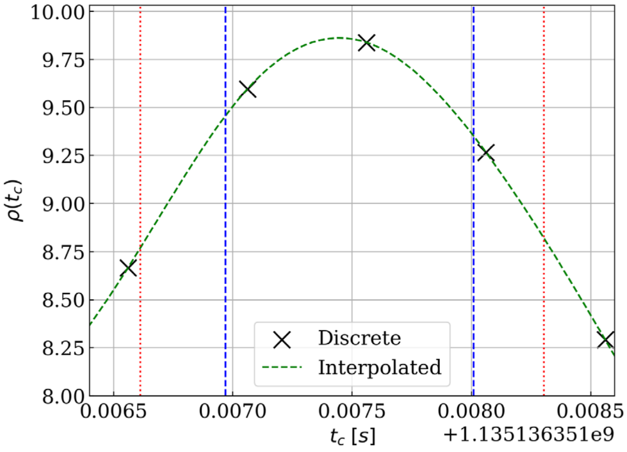

FIG. 4. The figure plots $\rho(t_c)$ for values of t_c around the observed GPS of the signal GW151226 [65] for the LIGO-Hanford detector. The dotted red lines set the bound on t_c as described in Eq. (26). The dashed blue lines show the 99.9% credible interval of the posterior distribution on t_c obtained from performing parameter estimation on the full set of extrinsic parameters.

We have verified the efficacy of our choices on the range of coalescence time and distance by performing parameter estimation runs on hundreds of simulated signals.

As an example, we estimate the extrinsic parameters for the signal GW151226 [65]. We scatter $N_{\rm pts} = 1{,}000{,}000$ points within the multidimensional grid for each cycle (step 1 in Sec. II), set $P_{\rm thr} = 0.9999$ when evaluating $\mathcal{L}_{\rm thr}$, and keep a minimum of $N_{\rm min} = 8000$ points at each cycle to evaluate $\mathcal{L}_{N_{\rm min}}$ (step 2 in Sec. II). We terminate sampling once eight cycles have elapsed. Table III shows the evolution of various quantities. The process collects ~10,000 effective samples from a total of ~160,000 weighted samples in eight cycles. As expected, for the first two cycles, the likelihood threshold is set by $\mathcal{L}_{N_{\rm min}}$ while for later cycles, where there is a higher probability of randomly scattering points within the high likelihood region, the threshold is set by $\mathcal{L}_{\rm thr}$. Unlike in previous examples, the likelihood threshold is positive and

TABLE III. Output from VARAHA showing the evolution of the number of bins in each dimension $N_{\rm bins}$, the likelihood threshold in each cycle (either $\mathcal{L}_{N_{\rm min}}$ or $\mathcal{L}_{\rm thr}$ depending on the situation, see text for details), and the number of effective samples $n_{\rm eff}$ when estimating the extrinsic parameters for GW151226.

Cycle	N_b	Log-likelihood threshold	$n_{\rm eff}$
1	4	56.8	42
2	6	69.4	257
4	8	72.1	2671
8	8	72.1	9751

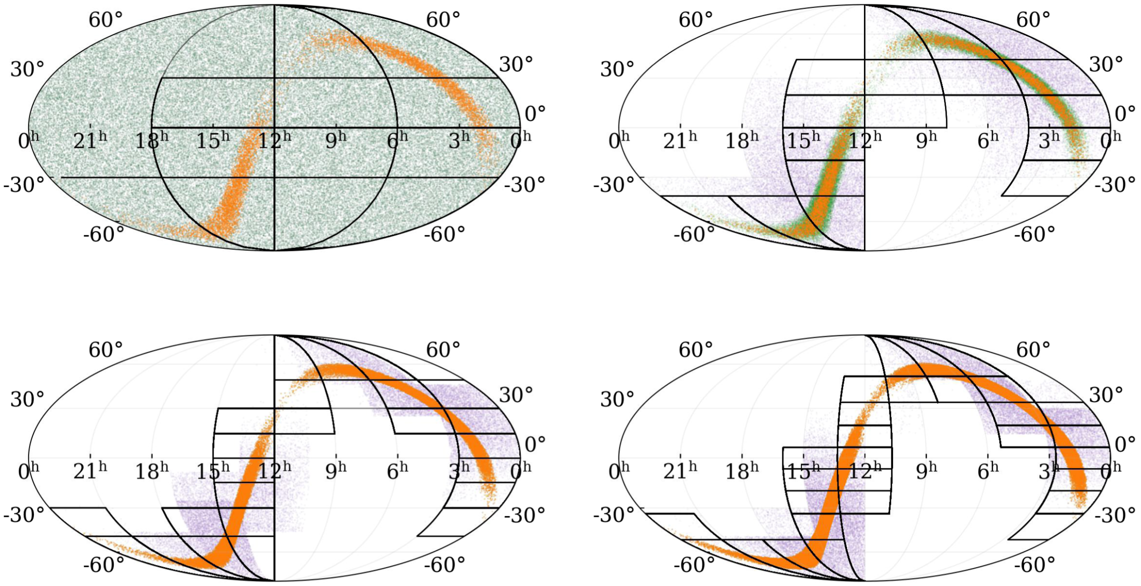

FIG. 5. Points obtained when sampling over the extrinsic parameters for the observation GW151226. The top left panel shows the first cycle, the top right shows the second cycle, the bottom left shows the fourth cycle and the bottom right shows the eighth cycle. The purple dots show the points randomly drawn from within the multidimensional grid, the green dots are all of the points that have a likelihood larger than the likelihood threshold from the previous cycle and the orange dots are the points with likelihoods larger than the likelihood threshold for the current cycle. The black lines show the multidimensional grid that surrounds the live volume. Eight cycles were completed in less than 45 s using one CPU thread and accumulated around 10,000 effective samples.

increases monotonically. This is because we have neglected the $\langle d|d\rangle$ term in Eq. (11). Figure 5 pictorially shows VARAHA converging to the most probable sky location of GW151226. We see that as the number of cycles increases, the live volume shrinks to the region of high likelihood. As a result, the number of bins in a multidimensional grid that encloses the live volume steadily increases with each cycle. This indicates that the error on the MC volume is steadily decreasing over time. For this example, VARAHA localizes GW151226 within 45 seconds on a single CPU thread. Assuming an approximately linear scaling with the number of CPU threads, we expect to localize most gravitational-wave signals in less than five seconds when parallelizing over ten CPU threads (see Sec. IV C for a discussion about CPU scaling). The exact run-time of BAYESTAR is unknown for this case, but we expect that BAYESTAR completed in ~2 minutes when running on a single CPU thread (based on Fig. 12 in [35]).

C. Intrinsic parameters

In order to obtain samples from the full posterior distribution, we need to vary both the intrinsic and extrinsic parameters in Eq. (23). However, as changing the extrinsic parameters only changes the overall amplitude, phase and time of arrival of the gravitational wave signal (since we are restricting to the leading multipole of a nonprecessing system), samples can be obtained by combining independent and separated sampling over the extrinsic and intrinsic parameters. In order to achieve this, we require two likelihoods: one which is conditioned only on the extrinsic parameters (introduced in Sec. III B), and another, marginalized likelihood, which is dependent only on the intrinsic parameters of the source. The marginalized likelihood that is dependent only on the intrinsic parameters is simply [44]

$$\mathcal{L}_{\mathrm{intr}}(\vec{d}|\boldsymbol{\theta}) = \int \mathrm{d}t_c \mathrm{d}\boldsymbol{\Omega} \mathcal{L}(\vec{d}|t_c, \boldsymbol{\Omega}, \boldsymbol{\theta}). \tag{29}$$

This procedure breaks one high-dimensional problem into two smaller-dimensional problems and has two significant benefits. First, the computational requirement of sampling decreases with decreased dimensionality [42] which is expected to reduce the overall cost. Second, an analysis sampling the full parameter space needs to generate a simulated gravitational waveform for each likelihood calculation, and compute the inner product of the waveform with the detector data. Often the computational cost to generate and filter these waveforms is high. By splitting the sampling, for each waveform generation and matched filtering operation, the likelihood over arbitrary values of extrinsic parameters can be calculated using inexpensive operations that change the overall amplitude, phase and time of the signal.

To evaluate the marginalized likelihood in Eq. (29), we integrate over the extrinsic parameter space for each draw of the intrinsic parameters $\boldsymbol{\theta}$. We identify the high-likelihood region of the intrinsic parameter space by following the method outlined in Sec. II, albeit with the likelihood calculation replaced by the more complex calculation of $\mathcal{L}_{\mathrm{intr}}$. To begin, we define the region of interest in the intrinsic parameter space. Here, our four parameters are the chirp mass $\mathcal{M}$, mass ratio q, and the spins of each black hole aligned (or antialigned) with the orbital angular momentum χ_1 and χ_2. We set the initial range of the intrinsic parameter space to be

$$\begin{aligned}\mathcal{M} &= [\mathcal{M}_0 - \Delta\mathcal{M}, \mathcal{M}_0 + \Delta\mathcal{M}]\\ q &= [0.05, 1.0]\\ \chi_1 &= [-0.9, 0.9]\\ \chi_2 &= [-0.9, 0.9].\end{aligned} \tag{30}$$

The ranges for mass ratio and spins are primarily driven by the region of validity of the IMRPhenomD waveform model [80,81], as well as the physical restriction that $q \leq 1, |\chi| \leq 1$. The central value of the chirp mass range, $\mathcal{M}_o$, is the chirp mass of the GW search template that identified the signal. The width $\Delta\mathcal{M}$ is chosen as

$$\Delta\mathcal{M} = \min\left(1.2\times 10^{-3}\left(\frac{10}{\rho_0}\right)\mathcal{M}_0^{8/3}, \mathcal{M}_o^{1.1}/20\right), \tag{31}$$

where ρ_0 is the reported SNR. The first term in Eq. (31) is motivated by the expected accuracy of measurements of the chirp mass for low-mass signals where the inspiral is the dominant part of the signal observed in the detectors [33]. The second term in Eq. (31) is taken from empirical uncertainties of chirp mass measurements from GWTC-3 [4] and is conservatively broad to ensure that the range is broader than the observed distribution. We again note that the mass values are in the frame of the detector, thus $\mathcal{M} = (1+z)\mathcal{M}_{\text{source}}$.

For each draw of intrinsic parameters $\boldsymbol{\theta}_j$, we marginalize the likelihood by integrating it over a fiducial parameter space for the extrinsic parameters, $\boldsymbol{\Omega}_o$. To generate $\boldsymbol{\Omega}_o$, we make use of the samples generated in the extrinsic parameter space associated with the intrinsic parameters identified by the search, $\boldsymbol{\theta}_o$. In general, we expect there to be a minimal correlation between the masses and spins and several of the extrinsic parameters. As discussed in [35,104], while changing the masses and spins will impact the measured coalescence time in each detector, the relative time delay will be only weakly impacted and, consequently, the inferred sky location will be largely independent of the intrinsic parameters. Similarly, the orientation of the binary, encoded in the inclination ι primarily depends upon the observed ratio of power in the two gravitational wave polarizations and this is unlikely to change significantly with mass or spin. Finally, although the intrinsic amplitude of a gravitational wave signal does vary linearly with mass, the chirp mass width is constrained to be at most a few percent of the measured value resulting in the inferred distance varying a few percent with changes in mass. Thus, the overall change in the volume that contains the high likelihood region in the extrinsic parameter space only varies a few percent with any change in the intrinsic parameters. To accommodate these fractional changes, we use $\boldsymbol{\theta}_o$ and specify a lower P_{thr} for the intrinsic parameter space than what is desired for the extrinsic parameters. This means that VARAHA defines each volume in the intrinsic parameter space to accommodate a slightly smaller probability than what is desired for the extrinsic parameters. For instance, if $P_{\text{thr}} = 0.999$ is specified for the extrinsic parameters, VARAHA uses a $P_{\text{thr}} = 0.995$ for the intrinsic parameters. This *ad hoc* choice results in the recovery of sane posteriors even at a population level. However, we will ascribe a more rigorous treatment of this problem in a future presentation.

We generate $\boldsymbol{\Omega}_o$ by retaining samples of $(\alpha, \delta, \iota, d_L)$ that cross the likelihood threshold from the extrinsic-only analysis and augment it with samples from the remaining three parameters: (ψ, ϕ_c, t_c) as defined later. For each draw in the intrinsic parameter space, we evaluate Eq. (29) by numerically integrating over $\boldsymbol{\Omega}_o$.[6] The observed gravitational wave phase and coalescence time will vary significantly with the masses and spins. Thus, we draw (ϕ_c, ψ) from the full range $(0, 2\pi)$. In addition, due to the degeneracy between the coalescence phase and the polarization angle, retaining samples of the latter from the extrinsic-only analysis or regenerating them does not impact the posterior on the intrinsic parameters. The appropriate range for the coalescence time t_c can be derived using the same method as for the initial point $\boldsymbol{\theta}_o$, as described around Eq. (26), although using the peak SNR in the reference detector for the intrinsic parameters $\boldsymbol{\theta}_j$. For each sample in $\boldsymbol{\Omega}_o^i$ we precalculate D_{eff}^i and $\phi_{\boldsymbol{\Omega}}^i$. Each draw of $\boldsymbol{\theta}_j$ requires waveform generation, matched filtering of the data and calculation of likelihood using these precalculated values. Finally, the (marginalized) likelihood is obtained by approximating the integral in Eq. (29) by writing

$$\mathcal{L}_{\text{intr}}(\vec{d}|\boldsymbol{\theta}_j) \approx \sum_i \sum_{k\in\text{dets}} \mathcal{L}(d_k|t_c, \boldsymbol{\Omega}_o^i, \boldsymbol{\theta}_j). \tag{32}$$

We continue sampling using different draws of $\boldsymbol{\theta}_j$ and follow our sampling procedure guided by their marginalized likelihood values.

Calculating the marginal likelihood is computationally expensive. An approximate but optimistic value can be calculated by maximizing the likelihood independently over the time of arrival and phase in each detector. This is done by ignoring the phase term in Eq. (29) and using the maximum SNR in each detector independently. By ignoring both of these factors, we obtain a likelihood which will always be equal to or greater than the true likelihood:

$$\begin{aligned}\mathcal{L}_{\text{intr}}(\vec{d}|\boldsymbol{\theta}_j) \leq& \sum_i \exp\left(\sum_{k\in\text{dets}}\left[-\frac{1}{2}\frac{\varrho_o^k(\boldsymbol{\theta}_j)^2}{(D_{\text{eff}}^i)^2} + \frac{\varrho_o^k(\boldsymbol{\theta}_j)}{D_{\text{eff}}^i}\right.\right.\\ &\left.\left.\times \max_t|\rho^k(\boldsymbol{\theta}_j, t)|\right]\right)\\ \leq& \exp\left(\sum_{k\in\text{dets}}\frac{1}{2}(\max_t|\rho^k(\boldsymbol{\theta}_j, t)|)^2\right).\end{aligned} \tag{33}$$

[6] Note that the prior remains uniform in this procedure. The samples corresponding to $\boldsymbol{\Omega}_o$ are uniformly distributed.

The last line in Eq. (33) further maximizes the likelihood on all the extrinsic parameters. The benefit is that the term in the second line can be calculated after matched filtering the data, in addition, for the term in the first line we only need to generate D^i_{eff} for each of the intrinsic samples. Both these calculations require significantly less computation than the full likelihood calculation. We thus marginalize only if both of these values are larger than the likelihood threshold.

The marginalized likelihood $\mathcal{L}_{\mathrm{intr}}(\vec{d}|\boldsymbol{\theta}_j)$ assigns a weight for each point in the intrinsic parameter space

$$w_j = \frac{\mathcal{L}_{\mathrm{intr}}(\vec{d}|\boldsymbol{\theta}_j)}{\mathcal{L}_{\max}(j)}, \tag{34}$$

where $\mathcal{L}_{\max}(j)$ is the maximum marginalized likelihood value among all the samples. In addition, we obtain a weight w^i_j for each sample $\boldsymbol{\Omega}^i_o$ corresponding to the intrinsic parameters $\boldsymbol{\theta}_j$, which is simply

$$w^j_i = \frac{\mathcal{L}(\vec{d}|\boldsymbol{\theta}_i, t_{ci}^{\,j}, \boldsymbol{\Omega}^j_i)}{\mathcal{L}_{\max}(i,j)}, \tag{35}$$

where $\mathcal{L}_{\max}(i,j)$ is the maximum likelihood of all the samples in the fiducial volume estimated for each sample of the intrinsic parameters.

Independent samples in the full eleven-dimensional parameter can be generated by performing rejection sampling on w^j_i. However, storing all these weights is challenging. We circumvent this problem by randomly keeping only one set of extrinsic parameters for each set of intrinsic parameters and discarding the rest. This choice is made by performing rejection sampling on the weights w^j_i to select a single $\boldsymbol{\Omega}^i_o$ and t^i_j for each $\boldsymbol{\theta}_j$. Thus for each point in the intrinsic parameter space that we sample, we store a single sample $(\boldsymbol{\theta}_j, \boldsymbol{\Omega}^i_o, t^i_j)$ and the weight, w_j, associated with the intrinsic parameter, $\boldsymbol{\theta}_j$.

We are now left with the sampling of intrinsic parameters and the associated marginalized likelihoods, in the successive cycles we identify live volumes across four dimensions $(\mathcal{M}, q, \chi_1, \chi_2)$. For each cycle, we evaluate the marginalized likelihood values and continue the cycles until we obtain the desired number of effective samples,

$$N^{\mathrm{eff}}_{\mathrm{int}} = \frac{(\sum w_j)^2}{\sum w_j^2}. \tag{36}$$

We continue sampling intrinsic parameters for GW151226. We scatter points within the multidimensional grid for each cycle (step 1 in Sec. II), set $P_{\mathrm{thr}} = 0.9999$ when evaluating $\mathcal{L}_{\mathrm{thr}}$, and keep a minimum of $N_{\min} = 1000$ points at each cycle to evaluate $\mathcal{L}_{N_{\min}}$ (step 2 in Sec. II).[7] We terminate sampling once eight cycles have been completed.

TABLE IV. Output from VARAHA showing the evolution of the number of bins in each dimension N_{bins}, the likelihood threshold in each cycle (either $\mathcal{L}_{N_{\min}}$ or $\mathcal{L}_{\mathrm{thr}}$ depending on the situation, see text for details), and the number of effective samples n_{eff}. Here we estimate the intrinsic parameters for the observation of GW151226.

Cycle	N_b	Log-likelihood threshold	n_{eff}
1	8	65.6	133
2	11	71.4	940
4	13	71.4	3420
8	15	71.4	9070

Table IV lists the number of bins in the multidimensional histogram, the likelihood threshold and the number of effective samples over the cycles. Like previous examples, the likelihood threshold increases initially before reaching a final value. Figure 6 shows the evolution of the live volume in the $\mathcal{M}$-q plane with the increase in cycle number.

D. Implementation

To summarize, VARAHA analyzes gravitational-wave signals as follows:

(1) *Obtain posterior on the extrinsic parameters*.—Obtain the posterior distribution for the extrinsic parameters by following the steps detailed in Sec. II A. The full gravitational-wave likelihood is used but we fix the intrinsic parameters to the values reported by the GW search pipelines.

(2) *Construct a fiducial volume for the extrinsic parameters*.—Retrieve luminosity distance, inclination angle, right ascension, and declination and augment with the remaining extrinsic parameters as defined in Sec. III C.

(3) *Obtain posterior on the intrinsic parameters*.—Marginalize the full likelihood on the fiducial volume calculated in step 2, and obtain the posterior distribution for the intrinsic parameters by following steps detailed in Sec. II A. We now use the marginalized likelihood given in Eq. (32) and a smaller P_{thr} than in step 1. The fiducial volume based on the extrinsic parameters remains fixed.

E. Processing time

The faster processing times of this analysis are due to the following reasons:

(1) A large dimensional problem has been broken into two small dimensional problems resulting in reduced computational requirement.

(2) The sampling method is entirely likelihood driven and swiftly converges to the relevant region of the parameter space.

(3) Taking inverse Fourier transform of Eq. (13) allows constraining bounds on t_c and vectorized estimation

[7] As we do not accurately compute the likelihood for all the intrinsic samples [see Eq. (33)] we sample until the number of live points increases by $N_{\min}$ and count N_{pts} accordingly.

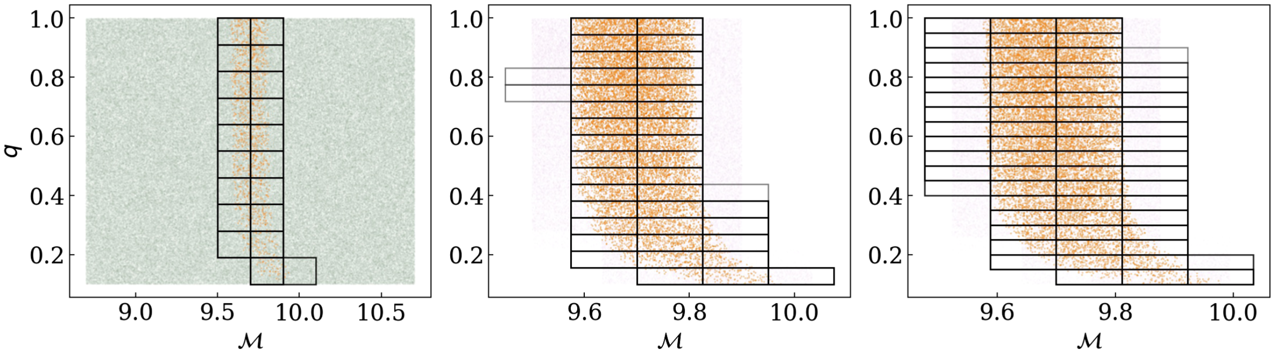

FIG. 6. The 2D projection of the evolution in the $\mathcal{M}$-q plane for the analysis sampling the intrinsic parameters for the observation GW151226. The left panel shows the first cycle, the middle panel shows the fourth cycle and the right panel shows the eighth cycle. The purple dots show the points randomly drawn from within the multidimensional grid, the green dots are all of the points that have a likelihood larger than the likelihood threshold from the previous cycle and the orange dots are the points with likelihoods larger than the current likelihood threshold. The black lines show the multidimensional grid that surrounds the live volume with likelihood equal to the likelihood threshold.

of likelihood values at thousands of samples in the fiducial set of extrinsic parameters.

(4) The waveform morphology of the inferred templates is expected to be similar. This analysis does not match filter if the phase accumulated in the detector's sensitivity band (∼20–2000 Hz) by the fiducial waveform and a template waveform differs by more than 30 radians (approximately five cycles).

(5) This analysis does not marginalize over extrinsic parameters if an approximate but optimistic estimate of marginalized likelihood given in Eq. (33) is smaller than the likelihood threshold.

(6) Analysis has been rigorously optimized and performs a judicious vectorized operation to save on computation times.

IV. RESULTS

A. Example: GW151226

Figure 7 compares the VARAHA's posterior with the posterior obtained using LALInference. Both analyses use IMRPHENOMD [80,81] for waveform generation and use almost equivalent priors on masses and spins. LALInference allows priors on the component masses, while VARAHA uses uniform priors on the chirp mass and mass ratio. We used a wide prior for the component masses in LALInference and then applied a chirp mass constraint to produce almost equivalent priors between the two algorithms. There is a good agreement in the two results; there are small differences in the marginalized one-dimensional posteriors, but they are consistent at the 90% confidence level.

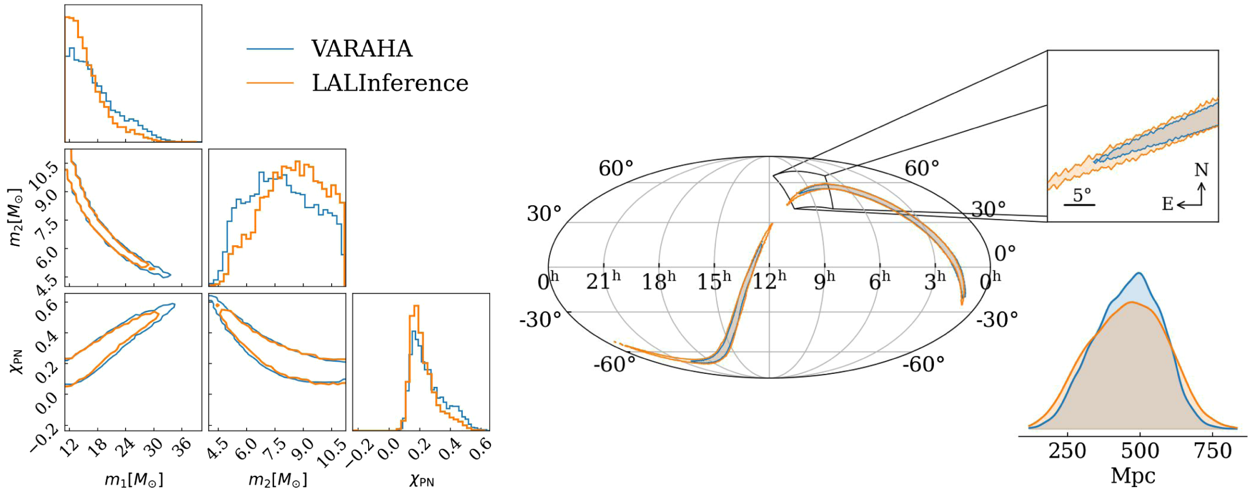

FIG. 7. Plot comparing the posterior distributions obtained from VARAHA (blue) and LALInference (orange) when analyzing GW151226. The left panel contains a corner plot for the primary and secondary masses as well as χ_{PN} [Eq. (5) in [105]]. The right panel shows the most probable sky location of GW151226 as well as the inferred distance. In both panels, contours enclose 90% probability mass.

This difference is likely a result of the slightly different priors assumed between the two codes. We also see good agreement in the recovered skymap and distance estimate, with any deviations likely a consequence of LALInference marginalizing over the calibration uncertainty while VARAHA does not yet have the functionality to do so. VARAHA obtained the posterior in less than one CPU hour. Based on the experience gained while running the two codes on different signals, we expect more than 2 orders of magnitude shorter computation times for VARAHA.

B. Example: GW170817

VARAHA can rapidly estimate the origin of the observed gravitational wave. This can have significant implications on electromagnetic follow-up programs for BNS observations. To demonstrate this, we analyze GW170817 using data from all three detectors and compare the estimated skymap with the location of the known host galaxy: NGC 4993 [66].

Figure 8 shows two skymaps: the skymap produced when VARAHA samples only the extrinsic parameters, and the skymap produced when VARAHA samples both the intrinsic and extrinsic parameters. We see that within two CPU minutes, VARAHA is able to localize GW170817 to within 49 square degrees (at 90% confidence) when sampling over only the extrinsic parameters. For this analysis, VARAHA used $N_{\rm pts} = 1,000,000$, $N_{\rm min} = 8000$, $P_{\rm thr} = 0.999$ and we stopped sampling once eight cycles were completed. The localization area was reduced to 17 square degrees (at 90% confidence) after 16 CPU hours when VARAHA samples over the intrinsic and extrinsic parameters. For this more detailed analysis, VARAHA used $N_{\rm min} = 1000$, $P_{\rm thr} = 0.995$ and we terminated sampling once eight cycles were completed. For comparison, BAYESTAR localizes GW170817 to within 31 square degrees (at 90% confidence) when analyzing only the extrinsic parameters, and LALInference localizes GW170817 to within 23 square degrees (at 90% confidence) [66] when analyzing both intrinsic and extrinsic parameters. Although the exact run times of BAYESTAR and LALInference are unknown for this case, we expect that BAYESTAR completed in ~2 CPU minutes (based on Fig. 12 in Ref. [35]) and LALInference completed in ~500 CPU hours (based on Ref. [58]).[8] Consequently, VARAHA matches the performance of BAYESTAR when sampling over only the extrinsic parameters, but, importantly, significantly improves upon LALInference when sampling over the intrinsic and extrinsic parameters. We note, however, that there has been recent work to reduce the run-time of LALInference by utilizing reduced order quadrature models [106], as well as using meshfree approximations [107]. Since the inclusion of intrinsic parameters is preferred, as it reduces the 90% contour for most cases, VARAHA may be pivotal for the rapid follow-up of binary neutron star observations.

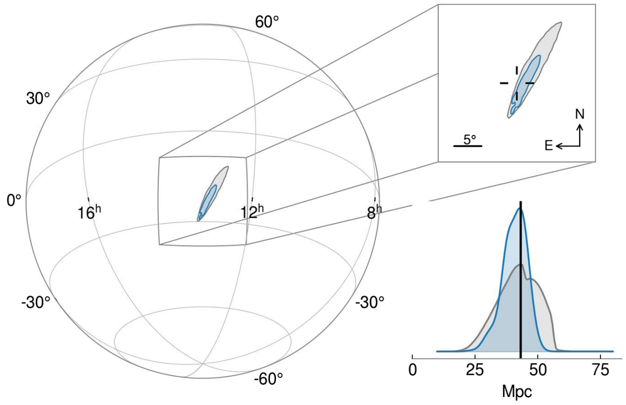

FIG. 8. The most probable sky location of GW170817 when VARAHA samples only the extrinsic parameters (gray) and intrinsic plus extrinsic parameters (blue). The reticle marks in the top-right inset show the position of NGC 4993. The bottom-right panel shows the posterior distribution for the luminosity distance and the black vertical line shows the distance to NGC 4993. The contours show the 90% confidence interval. The extrinsic-only analysis completed eight cycles in less than two minutes using one CPU thread. The intrinsic plus extrinsic analysis was completed in two hours using eight CPU threads.

C. Population level test

We evaluate the population level accuracy of VARAHA by performing a percentile-percentile (P-P) test [108]; this test involves performing hundreds of parameter estimation runs on synthetic signals embedded in simulated detector noise. The P-P test investigates if the measured interval of parameters at a credibility $f\%$ also encloses $f\%$ of true values among all the measurements. We perform parameter estimation on 500 simulated signals and show the P-P plot in Fig. 9. The parameters of the synthetic signals are drawn from VARAHA's prior, and we only analyze signals that cross a chosen SNR threshold. As described Sec. III D, we first estimate the extrinsic parameters. We do this by fixing the intrinsic parameters to the true values used when generating the synthetic signals. This is a reasonable choice as we do not expect the detector noise to bias the measurement, and the inferred population should average out to the true population. We then construct the fiducial volume for the extrinsic parameters and use it to estimate the marginalized likelihood for sampling the intrinsic parameters, as well as obtaining the 11-dimensional posterior on the full parameter space, as described in Sec. III C.

The distribution of injection parameters is listed in Eqs. (25) and (30). The luminosity distance is uniform in volume and chirp mass is uniformly distributed between

[8] The estimated run times of BAYESTAR and LALInference are based on results generated with, potentially, older CPUs than those used by VARAHA. Running on identical CPUs may decrease the expected run-time of BAYESTAR and LALInference. For the latest BAYESTAR run times, see https://lscsoft.docs.ligo.org/ligo.skymap/performance.html.

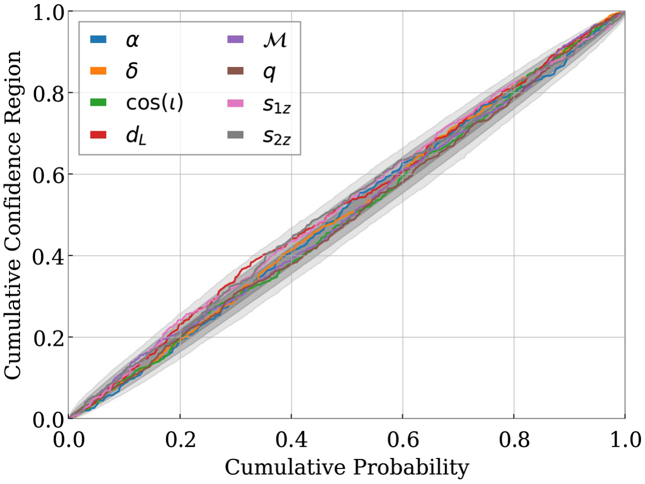

FIG. 9. Percentile-percentile (P-P) plot for 500 simulated injections. The 1-, 2- and 3-σ confidence intervals are indicated by the gray bands. For results to be unbiased the trails are required to be enclosed by the bands [108].

$10M_{\odot}$ and $20M_{\odot}$. Any injection that crosses a matched filter network SNR of 10 is selected for estimating the parameters. In this analysis, the network is comprised of advanced LIGO Livingston/Hanford and the advanced Virgo detector [109,110].

Most of the injections required eight seconds of simulated noise to accommodate the duration of simulated signals last in the detector's sensitivity band (~20–2000 Hz). Figure 10 shows the time taken by the analysis in performing PE. The median time needed was less than four minutes using ten threads. Almost all the PE runs were completed in less than sixteen minutes. We see that, in general, VARAHA takes longer to analyze binaries with more asymmetric masses, with the longest run-time of 40 minutes arising from a binary with mass ratio $q = 0.1$. Waveform generation and matched filtering consumed around 60% of the time and calculation of the reduced likelihood consumed around 30% of the time.

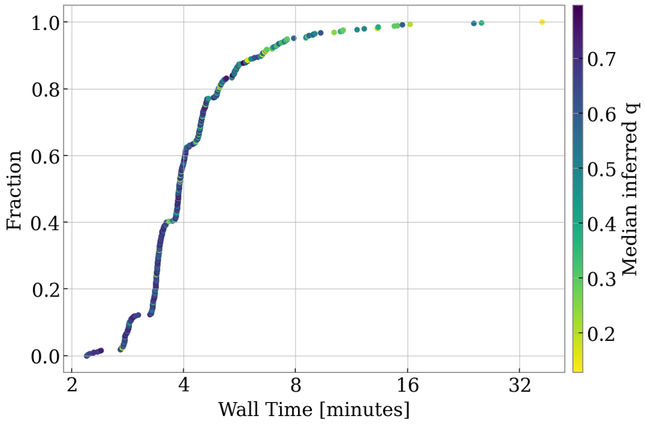

FIG. 10. Plot showing the wall time taken to perform PE on 500 synthetic signals embedded in simulated detector noise (see Fig. 9). Each of the 500 individual PE runs used ten CPU threads and each point is color coded by the median of the inferred mass ratio distribution. The median wall time is four minutes.

In Fig. 11, we show the scalability of the analysis with the number of CPU threads. We perform two additional sets of runs each using an expensive likelihood calculation and performed using one and 40 CPU threads, respectively. We make the likelihood calculation expensive, by including a one-tenth of a second delay in waveform generation, to reflect what we expect the scalability to be when VARAHA is extended to include additional physics (for example precession, higher order multipoles and eccentricity) since these waveform models are more expensive to generate than the simple aligned-spin case. The median time when using 40 CPU threads was 508 seconds while the median time when using one CPU thread was 16,910 seconds. Increasing the number of threads by a factor of 40 reduced the analysis time by a factor of 33. We expect this scaling to improve if the likelihood calculation is made more expensive.

V. DISCUSSION

In this article, we introduced a new sampling method that estimates Bayesian posteriors by identifying the volume that encloses the posterior mass and calculating the likelihood values inside the identified volume. This approach significantly increases our ability to parallelize the analysis

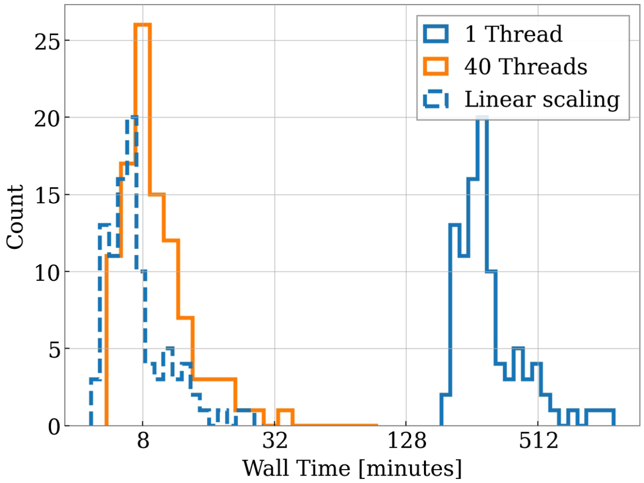

FIG. 11. Plot showing the wall time taken to perform PE on 100 synthetic signals embedded in simulated detector noise when the likelihood calculation is made expensive by including a delay in the waveform generation. The wall time taken when using a single CPU thread is shown in blue and 40 CPU threads in orange. Both analyses used identical settings. Assuming a linear CPU scaling from the one CPU thread analysis, the expected time taken to analyze 100 PE runs on 40 CPU threads is shown by the blue dashed histogram. In reality, increasing the number of CPU threads to 40 reduces the computation time by a factor of 33.

over multiple CPU threads and increases the efficacy to use vectorized likelihood calculations. In addition, we introduced the use of a likelihood threshold to judiciously populate the parts of the parameter space based on our prior understanding of the distribution at hand. Compared to nested sampling, which focuses directly on the estimation of probability mass, this approach is less robust in estimating higher-dimensional multimodal distributions.

Our sampling method is ideally suited for estimating parameters of the compact binaries from their GW signals. These parameters are inherently assumed to follow a unimodal likelihood distribution. We show that a large one-dimensional sampling problem can be broken into two small-dimensional sampling problems. Using vectorized likelihood calculations, we first sample the extrinsic parameters and subsequently obtain posteriors on the intrinsic parameters. We employ several approximations of the likelihood functions to draw boundaries in the parameter space and calculate likelihood values in the parts that meaningfully contribute to the posterior distribution. Tests indicate that our analysis can estimate parameters for most of the BBH signals in a few minutes.

VARAHA has the potential to include additional parameters when estimating parameters. The sky location is dominated by GW's time of arrival at the detectors of the network. The choice of intrinsic parameters has a weak impact, thus it is expected the samples of these parameters obtained using fiducial waveform can be used in calculating the marginal likelihood [44]. In comparison, the luminosity distance shows a greater dependence, specifically when higher harmonics are involved [111]. To accommodate this dependence luminosity distance samples obtained from fiducial waveform, using a relaxed value of likelihood threshold that encompasses a wider range, can cater to this dependence. The luminosity distance can also be numerically marginalized as it just divides the individual terms in Eq. (23). For the same reason, the inclination angle needs to be sampled along with the intrinsic parameters. Thus, including higher harmonics in the aligned spin model will increase the dimensionality from 4 to 5 when sampling using marginalized likelihood. Inclusion of in-plane spins will require including in-plane spins' magnitude and phase angle further increasing the dimensionality [112]. We have verified, the computational requirement for a six–eight dimensional distribution is comparable to what is needed from nested sampling. In addition, the PE also needs to account for inaccuracies introduced when calibrating the interferometric output [113]. Usually, this implies a significant increase in the dimensionality of the problem. However, being independent of GW signals, the inclusion of calibration errors only requires modulating the amplitude and phase of a template commensurate with the calibration error envelope and independent of the value of the intrinsic parameters. As VARAHA provides just a collection of samples with the corresponding weights, we anticipate this can be achieved by estimating the likelihood in the vicinity of live samples as guided by the calibration error envelopes. Alternatively, a Metropolis-Hastings algorithm can be constructed to calculate likelihoods in the vicinity of samples. Incorporating calibration errors should increase the computational requirement.

There is significant scope to decrease the computation time further. The most expensive component of the analysis is waveform generation and matched filtering. Both of these can be significantly reduced by using existing proposals [77,114,115]. We reconstruct the volume containing the probability mass using a structured multidimensional grid. It results in the reconstructed volume being much larger than what is estimated from the MC integral. A more efficient reconstruction can employ the use of unstructured grids reducing the number of cycles needed to obtain an effective sample size. Many of the calculations are done on the fly (time delay between detectors and antenna patterns) and can be precalculated to save computation time. The computation can also be reduced by choosing the right set of parameters [116,117]. The choice of intrinsic parameters, which are often degenerate with each other, is expected to change the structure of the live volume. A complicated structure will prove more challenging to reconstruct using hypercubes. A choice of parameters that disentangles the well and poorly measured parameters will result in less complicated live volume [117].

VI. CONCLUSION

The presented analysis offers significant improvement in processing time for estimating the parameters of a CBC while producing results comparable with the contemporary samplers. Although VARAHA is currently restricted to using only aligned-spin waveform models, it has the potential to include additional physics, such as precession, eccentricity or tidal deformability. However, for most cases, we expect that GW signals can be accurately modeled as produced from aligned spin binaries since the degree of orbital precession is often difficult to measure, see, e.g., [118]. This means that often the posterior distribution simply recovers the prior [119–121]. Of course, for binaries that precess [91], this means that some information is lost.

We use a uniform prior on all the parameters but, as we calculate the likelihood for each of our samples, it is straightforward to recalculate these likelihood values with a new prior by simply dividing the probability density of the new prior with the old one and multiplying it to the likelihood value [122]. As we can also calculate the marginal likelihood the samples can naturally be used for model selection. Furthermore, owing to the reduced computational time compared to other samplers, VARAHA is a natural choice for data diagnostics in understanding the systematics or non-Gaussianity in data associated with a signal, as well as performing parameter estimation on gravitational-wave data collected by the Laser Interferometer Space Antenna (LISA) [123] and third generation detectors [124,125] where PE is likely to be slow due to the large observation times.

This code is highly parallelizable as individual threads do not communicate. It also does not have to address any problem that a usual MCMC encounters. This may include proper mixing of chains, tuning of the code and potential correlations due to low proposal acceptance rate.

ACKNOWLEDGMENTS

We thank Christopher Berry, Mark Hannam, Vivien Raymond and Martin Hoffmann for constructive comments on this work. V. T. and S. F. thank STFC for support through Grants No. ST/V005618/1 and No. ST/N005430/1. C. H. thanks the UK Research and Innovation (UKRI) Future Leaders Fellowship for support through Grant No. MR/T01881X/1. The authors are also grateful for computational resources provided by Cardiff University and LIGO Laboratory and supported by Science and Technology Facilities Council (STFC) Grant No. ST/N000064/1 and National Science Foundation Grants No. PHY-0757058 and No. PHY-0823459. This research made use of data, software and/or web tools obtained from the Gravitational Wave Open Science Center ([126]), a service of LIGO Laboratory, the LIGO Scientific Collaboration and the Virgo Collaboration. LIGO is funded by the U.S. National Science Foundation. 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. This material is based upon work supported by NSF's LIGO Laboratory which is a major facility fully funded by the National Science Foundation. VARAHA is programmed in Python and implements modules from NumPy [127], SciPy [128] and PyCBC [129]. It uses Matplotlib [130], Seaborn [131], Corner [76] and LIGO.SKYMAP ([132]) for producing result plots.

APPENDIX: IMPORTANCE SAMPLING

Assuming the parameters of the statistical model, $\boldsymbol{\theta}$, are defined using the probability density $p(\boldsymbol{\theta})$, the mean and standard deviation of parameter $\theta \in \boldsymbol{\theta}$, when marginalizing over the other parameters, is simply

$$\langle\theta\rangle = \int \theta p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}$$
$$\sigma_{\langle\theta\rangle} = \sqrt{\int (\theta - \langle\theta\rangle)^2 p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}}. \quad \text{(A1)}$$

Often these integrals are intractable and a practical way to estimate these quantities is by drawing random samples $\boldsymbol{\theta}_i$ from the true distribution $p(\boldsymbol{\theta})$. The mean and variance of θ can then be estimated from the values $\theta_i \in \boldsymbol{\theta}_i$ as

$$\bar{\theta} = \sum_i \theta_i/N,$$
$$\sigma_{\bar{\theta}} = \sqrt{(\theta_i - \bar{\theta})^2/N}, \quad \text{(A2)}$$

where i indexes the N samples drawn.

Alternatively, one can use importance sampling and estimate the integrals in Eq. (A2) by calculating the weighted mean and weighted standard deviation [133],

$$\bar{\theta} = \frac{\sum_i w_i\theta_i}{\sum_i w_i},$$
$$\sigma_{\bar{\theta}} = \sqrt{\frac{w_i(\theta_i - \bar{\theta})^2}{\sum_i w_i}}, \quad \text{(A3)}$$

where i indexes the N samples drawn from a proposal distribution $\pi(\theta)$ and w_i is the sample weight,

$$w_i = \frac{p(\boldsymbol{\theta_i})}{\pi(\boldsymbol{\theta_i})}. \quad \text{(A4)}$$

Calculating parameter expectation values and uncertainties using a limited number of samples inevitably introduces sampling errors. When performing importance sampling, these errors also depend on the choice of proposal distribution. The measurement of $\bar{\theta}$ depends on the values of the weights, and the standard error relative to the true mean behaves as

$$\sigma_{\bar{\theta}} = \sigma_\theta/\sqrt{\frac{\sum_i(w_i^2)}{(\sum_i w_i)^2}}, \quad \text{(A5)}$$

where σ_θ is the standard deviation of the parameter θ in the distribution being sampled. When samples are drawn from the true distribution, $p(\boldsymbol{\theta})$, then $w_i \equiv 1$ and the error reduces to the standard proportionality of $1/\sqrt{N}$ [134].

Consequently, an effective sample size for a given proposal distribution can be defined as [74]

$$n_{\mathrm{eff}} = \frac{(\sum_i w_i)^2}{\sum_i(w_i^2)}. \quad \text{(A6)}$$

The effective number of samples, n_{eff}, approximately represents the number of samples one would need to measure θ as accurately using samples from the true distribution. Since $n_{\mathrm{eff}} \leq N$, with equality if and only if samples are drawn from the distribution $p(\boldsymbol{\theta})$, it follows that we require a greater number of samples when drawing from an alternate distribution. A detailed discussion on effective sample size is provided in [74] and the included references.

Estimation of parameters using the Bayes equation,

$$p(\boldsymbol{\theta}|d) \propto p(d|\boldsymbol{\theta})\pi(\boldsymbol{\theta}), \tag{A7}$$

is in essence importance sampling. The proposal distribution acts as the prior distribution and the weights are replaced by a likelihood function $\mathcal{L} \equiv p(d|\boldsymbol{\theta})$ conditioned on the observed data d. Thus, we use these terms interchangeably. The posterior distribution is just the weighted prior distribution,

$$p(\boldsymbol{\theta}|d) \propto w\pi(\boldsymbol{\theta}), \tag{A8}$$

with the weights given by

$$w = \exp(\ell), \qquad \ell = \log(\mathcal{L}) - \log(\mathcal{L}_{\max}), \tag{A9}$$

where we have scaled $\mathcal{L}$, such that the maximum value of ℓ is zero. Such a scaling does not impact any discussion earlier as it gets absorbed when normalizing Eq. (A9). It helps obtain equal-weight samples after performing rejection sampling on the weights.

A point to consider is that, if rejection sampling is performed on the value of weights, it will result in a sample size of close to $\sum_i w_i$ with all the samples having equal weights of 1 [72]. However, such a procedure results in loss of information as n_{eff} is always larger than $\sum_i w_i$. Often MCMC methods are employed to sample from the posterior probability distribution. All the algorithms performing PE using MCMC methods implement some kind of rejection sampling. Although they produce equally weighted samples, they discard a good fraction of the likelihood information [73].

Using a proposal distribution that is significantly different from the true distribution when estimating parameters using importance sampling leads to most of the samples having very small weights, resulting in a severely reduced n_{eff} and a grossly inefficient analysis. If uniform priors are used, a naive estimation of likelihood inside an arbitrary big box leads to most samples having very small weights. Unlike MCMC methods, importance sampling is severely impacted by the curse of dimensionality.

[1] B. P. Abbott, R. Abbott, T. D. Abbott *et al.*, GWTC-1: A Gravitational-Wave Transient Catalog of Compact Binary Mergers Observed by LIGO and Virgo during the First and Second Observing Runs, Phys. Rev. X **9**, 031040 (2019).

[2] R. Abbott, T. D. Abbott, S. Abraham *et al.*, GWTC-2: Compact Binary Coalescences Observed by LIGO and Virgo during the First Half of the Third Observing Run, Phys. Rev. X **11**, 021053 (2021).

[3] R. Abbott, T. D. Abbott, F. Acernese *et al.*, GWTC-2.1: Deep extended catalog of compact binary coalescences observed by LIGO and Virgo during the first half of the third observing run, arXiv:2108.01045.

[4] R. Abbott, T. D. Abbott, F. Acernese *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] R. Abbott *et al.* (LIGO Scientific, Virgo, and KAGRA Collaborations), Constraints on Cosmic Strings Using Data from the Third Advanced LIGO–Virgo Observing Run, Phys. Rev. Lett. **126**, 241102 (2021).

[6] J. Calderón Bustillo, N. Sanchis-Gual, A. Torres-Forné, J. A. Font, A. Vajpeyi, R. Smith, C. Herdeiro, E. Radu, and S. H. W. Leong, GW190521 as a Merger of Proca Stars: A Potential New Vector Boson of 8.7×10^{-13} eV, Phys. Rev. Lett. **126**, 081101 (2021).

[7] J. Calderon Bustillo, N. Sanchis-Gual, S. H. W. Leong, K. Chandra, A. Torres-Forne, J. A. Font, C. Herdeiro, E. Radu, I. C. F. Wong, and T. G. F. Li, Searching for vector boson-star mergers within LIGO-Virgo intermediate-mass black-hole merger candidates, arXiv:2206.02551.

[8] 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).

[9] H. Estellés, M. Colleoni, C. García-Quirós, S. Husa, D. Keitel, M. Mateu-Lucena, M. d. L. Planas, and A. Ramos-Buades, New twists in compact binary waveform modeling: A fast time-domain model for precession, Phys. Rev. D **105**, 084040 (2022).

[10] E. Hamilton, L. London, J. E. Thompson, E. Fauchon-Jones, M. Hannam, C. Kalaghatgi, S. Khan, F. Pannarale, and A. Vano-Vinuales, Model of gravitational waves from precessing black-hole binaries through merger and ringdown, Phys. Rev. D **104**, 124027 (2021).

[11] J. E. Thompson, E. Fauchon-Jones, S. Khan, E. Nitoglia, F. Pannarale, T. Dietrich, and M. Hannam, Modeling the gravitational wave signature of neutron star black hole coalescences, Phys. Rev. D **101**, 124059 (2020).

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

[13] A. Matas *et al.*, Aligned-spin neutron-star–black-hole waveform model based on the effective-one-body approach and numerical-relativity simulations, Phys. Rev. D **102**, 043023 (2020).

[14] V. Varma, S. E. Field, M. A. Scheel, J. Blackman, D. Gerosa, L. C. Stein, L. E. Kidder, and H. P. Pfeiffer, Surrogate models for precessing binary black hole simulations with unequal masses, Phys. Rev. Res. **1**, 033015 (2019).

[15] G. Riemenschneider, P. Rettegno, M. Breschi, A. Albertini, R. Gamba, S. Bernuzzi, and A. Nagar, Assessment of

consistent next-to-quasicircular corrections and postadiabatic approximation in effective-one-body multipolar waveforms for binary black hole coalescences, Phys. Rev. D **104**, 104045 (2021).

[16] T. Dietrich, A. Samajdar, S. Khan, N. K. Johnson-McDaniel, R. Dudi, and W. Tichy, Improving the NRTidal model for binary neutron star systems, Phys. Rev. D **100**, 044003 (2019).

[17] B. P. Abbott *et al.* (LIGO Scientific Collaboration, Virgo Collaboration), Properties of the Binary Black Hole Merger GW150914, Phys. Rev. Lett. **116**, 241102 (2016).

[18] N. Christensen and R. Meyer, Parameter estimation with gravitational waves, Rev. Mod. Phys. **94**, 025001 (2022).

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

[20] V. Tiwari and S. Fairhurst, The emergence of structure in the binary black hole mass distribution, Astrophys. J. Lett. **913**, L19 (2021).

[21] R. Abbott *et al.*, The Population of Merging Compact Binaries Inferred Using Gravitational Waves Through GWTC-3, Phys. Rev. X **13**, 011048 (2023).

[22] V. Tiwari, Exploring features in the binary black hole population, Astrophys. J. **928**, 155 (2022).

[23] T. A. Callister, C.-J. 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).

[24] J. Roulet and M. Zaldarriaga, Constraints on binary black hole populations from LIGO-Virgo detections, Mon. Not. R. Astron. Soc. **484**, 4216 (2019).

[25] S. Vitale, S. Biscoveanu, and C. Talbot, The orientations of the binary black holes in GWTC-3, arXiv:2204.00968.

[26] 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).

[27] The LIGO Scientific Collaboration and the Virgo Collaboration, Constraining the p-Mode-g-Mode Tidal Instability with GW170817, Phys. Rev. Lett. **122**, 061104 (2019).

[28] M. Agathos, J. Meidam, W. Del Pozzo, T. G. F. Li, M. Tompitak, J. Veitch, S. Vitale, and C. Van Den Broeck, Constraining the neutron star equation of state with gravitational wave signals from coalescing binary neutron stars, Phys. Rev. D **92**, 023012 (2015).

[29] R. Abbott *et al.* (LIGO Scientific, VIRGO, and KAGRA Collaborations), Constraints on the cosmic expansion history from GWTC-3, arXiv:2111.03604.

[30] T. A. Apostolatos, C. Cutler, G. J. Sussman, and K. S. Thorne, Spin-induced orbital precession and its modulation of the gravitational waveforms from merging binaries, Phys. Rev. D **49**, 6274 (1994).

[31] N. Yunes, K. G. Arun, E. Berti, and C. M. Will, Post-circular expansion of eccentric binary inspirals: Fourier-domain waveforms in the stationary phase approximation, Phys. Rev. D **80**, 084001 (2009).

[32] B. J. Owen, Search templates for gravitational waves from inspiraling binaries: Choice of template spacing, Phys. Rev. D **53**, 6749 (1996).

[33] C. Cutler and É. E. Flanagan, Gravitational waves from merging compact binaries: How accurately can one extract the binary's parameters from the inspiral waveform?, Phys. Rev. D **49**, 2658 (1994).

[34] E. Poisson and C. M. Will, Gravitational waves from inspiraling compact binaries: Parameter estimation using second-post-Newtonian waveforms, Phys. Rev. D **52**, 848 (1995).

[35] L. P. Singer and L. R. Price, Rapid Bayesian position reconstruction for gravitational-wave transients, Phys. Rev. D **93**, 024013 (2016).

[36] J. Veitch *et al.*, Parameter estimation for compact binaries with ground-based gravitational-wave observations using the LALInference software library, Phys. Rev. D **91**, 042003 (2015).

[37] C. M. Biwer, C. D. Capano, S. De, M. Cabero, D. A. Brown, A. H. Nitz, and V. Raymond, PyCBC Inference: A Python-based parameter estimation toolkit for compact binary coalescence signal, Publ. Astron. Soc. Pac. **131**, 024503 (2019).

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

[39] 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).

[40] R. J. E. Smith, G. Ashton, A. Vajpeyi, and C. Talbot, Massively parallel Bayesian inference for transient gravitational-wave astronomy, Mon. Not. R. Astron. Soc. **498**, 4492 (2020).

[41] G. Ashton and C. Talbot, Bilby-MCMC: An MCMC sampler for gravitational-wave inference, Mon. Not. R. Astron. Soc. **507**, 2037 (2021).

[42] J. Skilling, Nested sampling for general Bayesian computation, Bayesian Anal. **1**, 833 (2006).

[43] N. Metropolis and S. Ulam, The Monte Carlo method, J. Am. Stat. Assoc. **44**, 335 (1949).

[44] C. Pankow, P. Brady, E. Ochsner, and R. O'Shaughnessy, Novel scheme for rapid parallel parameter estimation of gravitational waves from compact binary coalescences, Phys. Rev. D **92**, 023002 (2015).

[45] J. Lange, R. O'Shaughnessy, and M. Rizzo, Rapid and accurate parameter inference for coalescing, precessing compact binaries, arXiv:1805.10457.

[46] J. Wofford, A. Yelikar, H. Gallagher, E. Champion, D. Wysocki, V. Delfavero, J. Lange, C. Rose, V. Valsan, S. Morisaki, J. Read, C. Henshaw, and R. O'Shaughnessy, Expanding RIFT: Improving performance for GW parameter inference, Phys. Rev. D **107**, 024040 (2023).

[47] G. Ashton and T. Dietrich, The use of hypermodels to understand binary neutron star collisions, Nat. Astron. **6**, 961 (2022).

[48] C. Hoy, Accelerating multimodel Bayesian inference, model selection, and systematic studies for gravitational wave astronomy, Phys. Rev. D **106**, 083003 (2022).

[49] V. Tiwari, S. Klimenko, V. Necula, and G. Mitselmakher, Reconstruction of chirp mass in searches for gravitational wave transients, Classical Quantum Gravity **33**, 01LT01 (2016).

[50] H. Gabbard, C. Messenger, I. S. Heng, F. Tonolini, and R. Murray-Smith, Bayesian parameter estimation using conditional variational autoencoders for gravitational-wave astronomy, Nat. Phys. **18**, 112 (2022).

[51] S. R. Green and J. Gair, Complete parameter inference for GW150914 using deep learning, Mach. Learn. Sci. Tech. **2**, 03LT01 (2021).

[52] S. R. Green, C. Simpson, and J. Gair, Gravitational-wave parameter estimation with autoregressive neural network flows, Phys. Rev. D **102**, 104057 (2020).

[53] M. Dax, S. R. Green, J. Gair, J. H. Macke, A. Buonanno, and B. Schölkopf, Real-Time Gravitational Wave Science with Neural Posterior Estimation, Phys. Rev. Lett. **127**, 241103 (2021).

[54] 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, Phys. Rev. Lett. **130**, 171403 (2023).

[55] H. Shen, E. A. Huerta, E. O'Shea, P. Kumar, and Z. Zhao, Statistically-informed deep learning for gravitational wave parameter estimation, Mach. Learn. Sci. Tech. **3**, 015007 (2022).

[56] M. J. Williams, J. Veitch, and C. Messenger, Nested sampling with normalizing flows for gravitational-wave inference, Phys. Rev. D **103**, 103006 (2021).

[57] B. Farr, C. P. L. Berry, W. M. Farr, C.-J. Haster, H. Middleton, K. Cannon, P. B. Graff, C. Hanna, I. Mandel, C. Pankow, L. R. Price, T. Sidery, L. P. Singer, A. L. Urban, A. Vecchio, J. Veitch, and S. Vitale, Parameter estimation on gravitational waves from neutron-star binaries with spinning components, Astrophys. J. **825**, 116 (2016).

[58] C. P. L. Berry, I. Mandel, H. Middleton, L. P. Singer, A. L. Urban, A. Vecchio, S. Vitale, K. Cannon, B. Farr, W. M. Farr, P. B. Graff, C. Hanna, C.-J. Haster, S. Mohapatra, C. Pankow, L. R. Price, T. Sidery, and J. Veitch, Parameter estimation for binary neutron-star coalescences with realistic noise during the Advanced LIGO era, Astrophys. J. **804**, 114 (2015).

[59] R. Abbott *et al.*, Prospects for observing and localizing gravitational-wave transients with Advanced LIGO, Advanced Virgo and KAGRA, Living Rev. Relativity **23**, 3 (2020).

[60] M. Pürrer, Frequency-domain reduced order models for gravitational waves from aligned-spin compact binaries, Classical Quantum Gravity **31**, 195010 (2014).

[61] P. Canizares, S. E. Field, J. Gair, V. Raymond, R. Smith, and M. Tiglio, Accelerated Gravitational Wave Parameter Estimation with Reduced Order Modeling, Phys. Rev. Lett. **114**, 071104 (2015).

[62] S. Vinciguerra, J. Veitch, and I. Mandel, Accelerating gravitational wave parameter estimation with multi-band template interpolation, Classical Quantum Gravity **34**, 115006 (2017).

[63] Y. Setyawati, M. Pürrer, and F. Ohme, Regression methods in waveform modeling: A comparative study, Classical Quantum Gravity **37**, 075012 (2020).

[64] N. J. Cornish, Heterodyned likelihood for rapid gravitational wave parameter inference, Phys. Rev. D **104**, 104054 (2021).

[65] B. P. Abbott, R. Abbott, T. D. Abbott *et al.*, GW151226: Observation of Gravitational Waves from a 22-Solar-Mass Binary Black Hole Coalescence, Phys. Rev. Lett. **116**, 241103 (2016).

[66] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), GW170817: Observation of Gravitational Waves from a Binary Neutron Star Inspiral, Phys. Rev. Lett. **119**, 161101 (2017).

[67] W. K. Hastings, Monte Carlo sampling methods using Markov chains and their applications, Biometrika **57**, 97 (1970).

[68] J. Skilling, Nested sampling, in *American Institute of Physics Conference Series*, edited by R. Fischer, R. Preuss, and U. V. Toussaint (2004), Vol. 735, pp. 395–405, 10.1063/1.1835238.

[69] E. Higson, W. Handley, M. Hobson, and A. Lasenby, Dynamic nested sampling: An improved algorithm for parameter estimation and evidence calculation, Stat. Comput. **29**, 891 (2019).

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

[71] A. Dvoretzky, J. Kiefer, and J. Wolfowitz, Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator, Ann. Math. Stat. **27**, 642 (1956).

[72] G. Casella, C. P. Robert, and M. T. Wells, Generalized accept-reject sampling schemes, Lect. Notes Monogr. Ser. **45**, 342 (2004).

[73] D. van Ravenzwaaij, P. Cassey, and S. D. Brown, A simple introduction to Markov-chain Monte Carlo sampling, Psychon. Bull. Rev. **25**, 143 (2018).

[74] L. Martino, E. Victor, and S. Carlos, Effective sample size for importance sampling based on discrepancy measures, Signal Process. **131**, 386 (2017).

[75] S. W. Nydick, The Wishart and inverse Wishart distributions (2012), https://swnydick.github.io/assets/reports/Wishart_Distribution.pdf.

[76] D. Foreman-Mackey, CORNER.PY: Scatterplot matrices in Python, J. Open Source Software **1**, 24 (2016).

[77] B. Zackay, L. Dai, and T. Venumadhav, Relative binning and fast likelihood evaluation for gravitational wave parameter estimation, arXiv:1806.08792.

[78] N. Leslie, L. Dai, and G. Pratten, Mode-by-mode relative binning: Fast likelihood estimation for gravitational waveforms with spin-orbit precession and multiple harmonics, Phys. Rev. D **104**, 123030 (2021).

[79] C. Mills and S. Fairhurst, Measuring gravitational-wave higher-order multipoles, Phys. Rev. D **103**, 024042 (2021).

[80] S. Husa, S. Khan, M. Hannam, M. Pürrer, F. Ohme, X. J. Forteza, and A. Bohé, Frequency-domain gravitational waves from nonprecessing black-hole binaries. I. New numerical waveforms and anatomy of the signal, Phys. Rev. D **93**, 044006 (2016).

[81] S. Khan, S. Husa, M. Hannam, F. Ohme, M. Pürrer, X. Jiménez Forteza, and A. Bohé, Frequency-domain gravitational waves from nonprecessing black-hole binaries. II. A phenomenological model for the advanced detector era, Phys. Rev. D **93**, 044007 (2016).

[82] A. Taracchini, Y. Pan, A. Buonanno, E. Barausse, M. Boyle, T. Chu, G. Lovelace, H. P. Pfeiffer, and M. A. Scheel, Prototype effective-one-body model for nonprecessing spinning inspiral-merger-ringdown waveforms, Phys. Rev. D **86**, 024011 (2012).

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

[84] 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).

[85] G. Pratten, S. Husa, C. Garcia-Quiros, M. Colleoni, A. Ramos-Buades, H. Estelles, and R. Jaume, Setting the cornerstone for a family of models for gravitational waves from compact binaries: The dominant harmonic for non-precessing quasicircular black holes, Phys. Rev. D **102**, 064001 (2020).

[86] C. Kalaghatgi, M. Hannam, and V. Raymond, Parameter estimation with a spinning multimode waveform model, Phys. Rev. D **101**, 103004 (2020).

[87] F. H. Shaik, J. Lange, S. E. Field, R. O'Shaughnessy, V. Varma, L. E. Kidder, H. P. Pfeiffer, and D. Wysocki, Impact of subdominant modes on the interpretation of gravitational-wave signals from heavy binary black hole systems, Phys. Rev. D **101**, 124054 (2020).

[88] R. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Properties and astrophysical implications of the $150M_{\odot}$ binary black hole merger GW190521, Astrophys. J. Lett. **900**, L13 (2020).

[89] R. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), GW190412: Observation of a binary-black-hole coalescence with asymmetric masses, Phys. Rev. D **102**, 043015 (2020).

[90] R. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), GW190814: Gravitational waves from the coalescence of a 23 solar mass black hole with a 2.6 solar mass compact object, Astrophys. J. Lett. **896**, L44 (2020).

[91] M. Hannam *et al.*, General-relativistic precession in a black-hole binary, Nature (London) **610**, 652 (2022).

[92] N. V. Krishnendu and F. Ohme, Interplay of spin-precession and higher harmonics in the parameter estimation of binary black holes, Phys. Rev. D **105**, 064012 (2022).

[93] J. Aasi *et al.* (LIGO Scientific Collaboration), Advanced LIGO, Classical Quantum Gravity **32**, 074001 (2015).

[94] F. Acernese *et al.* (VIRGO Collaboration), Advanced Virgo: A second-generation interferometric gravitational wave detector, Classical Quantum Gravity **32**, 024001 (2015).

[95] T. Akutsu *et al.* (KAGRA Collaboration), Overview of KAGRA: Calibration, detector characterization, physical environmental monitors, and the geophysics interferometer, Prog. Theor. Exp. Phys. **2021**, 05A102 (2021).

[96] S. Babak *et al.*, Searching for gravitational waves from binary coalescence, Phys. Rev. D **87**, 024033 (2013).

[97] B. F. Schutz, Networks of gravitational wave detectors and three figures of merit, Classical Quantum Gravity **28**, 125023 (2011).

[98] B. Allen, W. G. Anderson, P. R. Brady, D. A. Brown, and J. D. E. Creighton, FINDCHIRP: An algorithm for detection of gravitational waves from inspiraling compact binaries, Phys. Rev. D **85**, 122006 (2012).

[99] S. Fairhurst, Localization of transient gravitational wave sources: Beyond triangulation, Classical Quantum Gravity **35**, 105002 (2018).

[100] S. A. Usman *et al.*, The PyCBC search for gravitational waves from compact binary coalescence, Classical Quantum Gravity **33**, 215004 (2016).

[101] C. Messick *et al.*, Analysis framework for the prompt discovery of compact binary mergers in gravitational-wave data, Phys. Rev. D **95**, 042001 (2017).

[102] T. Adams, D. Buskulic, V. Germain, G. M. Guidi, F. Marion, M. Montani, B. Mours, F. Piergiovanni, and G. Wang, Low-latency analysis pipeline for compact binary coalescences in the advanced gravitational wave detector era, Classical Quantum Gravity **33**, 175012 (2016).

[103] Q. Chu, M. Kovalam, L. Wen, T. Slaven-Blair, J. Bosveld, Y. Chen, P. Clearwater, A. Codoreanu, Z. Du, X. Guo, X. Guo, K. Kim, T. G. F. Li, V. Oloworaran, F. Panther, J. Powell, A. S. Sengupta, K. Wette, and X. Zhu, SPIIR online coherent pipeline to search for gravitational waves from compact binary coalescences, Phys. Rev. D **105**, 024023 (2022).

[104] S. Fairhurst, Triangulation of gravitational wave sources with a network of detectors, New J. Phys. **11**, 123006 (2009); **13**, 069602(E) (2011).

[105] E. Baird, S. Fairhurst, M. Hannam, and P. Murphy, Degeneracy between mass and spin in black-hole-binary waveforms, Phys. Rev. D **87**, 024035 (2013).

[106] S. Morisaki and V. Raymond, Rapid parameter estimation of gravitational waves from binary neutron star coalescence using focused reduced order quadrature, Phys. Rev. D **102**, 104020 (2020).

[107] L. Pathak, A. Reza, and A. S. Sengupta, Rapid reconstruction of compact binary sources using meshfree approximation, arXiv:2210.02706.

[108] T. Sidery, B. Aylott, N. Christensen *et al.*, Reconstructing the sky location of gravitational-wave detected compact binary systems: Methodology for testing and comparison, Phys. Rev. D **89**, 084060 (2014).

[109] J. Aasi, B. P. Abbott, R. Abbott, T. Abbott *et al.*, Advanced LIGO, Classical Quantum Gravity **32**, 074001 (2015).

[110] F. Acernese, M. Agathos, K. Agatsuma *et al.*, Advanced Virgo: A second-generation interferometric gravitational wave detector, Classical Quantum Gravity **32**, 024001 (2015).

[111] S. Khan, F. Ohme, K. Chatziioannou, and M. Hannam, Including higher order multipoles in gravitational-wave models for precessing binary black holes, Phys. Rev. D **101**, 024056 (2020).

[112] S. Fairhurst, R. Green, C. Hoy, M. Hannam, and A. Muir, Two-harmonic approximation for gravitational waveforms from precessing binaries, Phys. Rev. D **102**, 024055 (2020).

[113] A. D. Viets, M. Wade, A. L. Urban, S. Kandhasamy, J. Betzwieser, D. A. Brown, J. Burguet-Castell, C. Cahillane, E. Goetz, K. Izumi, S. Karki, J. S. Kissel, G. Mendell, R. L. Savage, X. Siemens, D. Tuyenbayev, and A. J. Weinstein, Reconstructing the calibrated strain signal in the Advanced

LIGO detectors, Classical Quantum Gravity **35**, 095015 (2018).

[114] N. J. Cornish, Fast Fisher matrices and lazy likelihoods, arXiv:1007.4820.

[115] K. W. K. Wong, M. Isi, and T. D. P. Edwards, Fast gravitational wave parameter estimation without compromises, arXiv:2302.05333.

[116] E. Lee, S. Morisaki, and H. Tagoshi, Mass-spin reparametrization for a rapid parameter estimation of inspiral gravitational-wave signals, Phys. Rev. D **105**, 124057 (2022).

[117] J. Roulet, S. Olsen, J. Mushkin, T. Islam, T. Venumadhav, B. Zackay, and M. Zaldarriaga, Removing degeneracy and multimodality in gravitational wave source parameters, Phys. Rev. D **106**, 123015 (2022).

[118] R. Green, C. Hoy, S. Fairhurst, M. Hannam, F. Pannarale, and C. Thomas, Identifying when precession can be measured in gravitational waveforms, Phys. Rev. D **103**, 124023 (2021).

[119] P. Schmidt, M. Hannam, and S. Husa, Towards models of gravitational waveforms from generic binaries: A simple approximate mapping between precessing and nonprecessing inspiral signals, Phys. Rev. D **86**, 104063 (2012).

[120] S. Fairhurst, R. Green, C. Hoy, M. Hannam, and A. Muir, The two-harmonic approximation for gravitational waveforms from precessing binaries, Phys. Rev. D **102**, 024055 (2020).

[121] C. Hoy, C. Mills, and S. Fairhurst, Evidence for subdominant multipole moments and precession in merging black-hole-binaries from GWTC-2.1, Phys. Rev. D **106**, 023019 (2022).

[122] V. Tiwari, S. Fairhurst, and M. Hannam, Constraining black hole spins with gravitational-wave observations, Astrophys. J. **868**, 140 (2018).

[123] P. Bender, A. Brillet, I. Ciufolini, A. Cruise, C. Cutler, K. Danzmann, W. Folkner, J. Hough, P. McNamara, M. Peterseim *et al.*, LISA. Laser Interferometer Space Antenna for the detection and observation of gravitational waves. An international project in the field of fundamental physics in space (Max-Planck-Institut für Quantenoptik, 1998), https://pure.mpg.de/rest/items/item_52082_1/component/file_52083/content.

[124] M. Punturo *et al.*, The Einstein Telescope: A third-generation gravitational wave observatory, Classical Quantum Gravity **27**, 194002 (2010).

[125] D. Reitze, R. X. Adhikari, S. Ballmer, B. Barish, L. Barsotti, G. Billingsley, D. A. Brown, Y. Chen, D. Coyne, R. Eisenstein *et al.*, Cosmic Explorer: The U.S. contribution to gravitational-wave astronomy beyond LIGO, Bull. Am. Astron. Soc. **51**, 035 (2019).

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

[127] C. R. Harris *et al.*, Array programming with NumPy, Nature (London) **585**, 357 (2020).

[128] P. Virtanen *et al.* (SciPy 1.0 Contributors), SciPy 1.0: Fundamental algorithms for scientific computing in Python, Nat. Methods **17**, 261 (2020).

[129] A. Nitz *et al.*, gwastro/pycbc: v2.0.5 release of PyCBC (2022).

[130] J. D. Hunter, Matplotlib: A 2d graphics environment, Comput. Sci. Eng. **9**, 90 (2007).

[131] M. L. Waskom, Seaborn: Statistical data visualization, J. Open Source Software **6**, 3021 (2021).

[132] https://lscsoft.docs.ligo.org/ligo.skymap.

[133] L. Held and D. Bové, *Likelihood and Bayesian Inference: With Applications in Biology and Medicine*, Statistics for Biology and Health (Springer, Berlin, 2020).

[134] V. Tiwari, Estimation of the sensitive volume for gravitational-wave source populations using weighted Monte Carlo integration, Classical Quantum Gravity **35**, 145009 (2018).