Publications of the Astronomical Society of Australia (2019), **36**, e010, 12 pages
doi:10.1017/pasa.2019.2

CAMBRIDGE UNIVERSITY PRESS

Review (unsolicited)

An introduction to Bayesian inference in gravitational-wave astronomy: Parameter estimation, model selection, and hierarchical models

Eric Thrane and Colm Talbot

Centre for Astrophysics, School of Physics and Astronomy, Monash University, VIC 3800, Australia
OzGrav: The ARC Centre of Excellence for Gravitational-Wave Discovery, Clayton, VIC 3800, Australia

Abstract

This is an introduction to Bayesian inference with a focus on hierarchical models and hyper-parameters. We write primarily for an audience of Bayesian novices, but we hope to provide useful insights for seasoned veterans as well. Examples are drawn from gravitational-wave astronomy, though we endeavour for the presentation to be understandable to a broader audience. We begin with a review of the fundamentals: likelihoods, priors, and posteriors. Next, we discuss Bayesian evidence, Bayes factors, odds ratios, and model selection. From there, we describe how posteriors are estimated using samplers such as Markov Chain Monte Carlo algorithms and nested sampling. Finally, we generalise the formalism to discuss hyper-parameters and hierarchical models. We include extensive appendices discussing the creation of credible intervals, Gaussian noise, explicit marginalisation, posterior predictive distributions, and selection effects.

Keywords: methods: statistical – gravitational waves – stars: black holes – stars: neutron

(Received 07 September 2018; revised 12 January 2019; accepted 22 January 2019)

1. Preface: why study Bayesian inference?

Bayesian inference is an essential part of modern astronomy. It finds particularly elegant application in the field of gravitational-wave astronomy thanks to the clear predictions of general relativity and the extraordinary simplicity with which compact binary systems are described. An astrophysical black hole is completely characterised by just its mass and its dimensionless spin vector. The gravitational waveform from a black hole binary is typically characterised by just 15 parameters. Since sources of gravitational waves are so simple, and since we have a complete theory describing how they emit gravitational waves, there is a direct link between data and model. The significant interest in Bayesian inference within the gravitational-wave community reflects the great possibilities of this area of research.

Bayesian inference and parameter estimation are the tools that allow us to make statements about the Universe based on data. In gravitational-wave astronomy, Bayesian inference is the tool that allows us to reconstruct sky maps of where a binary neutron star merged (Abbott et al. 2017*c*), to determine that GW170104 merged 880^{+450}_{-390} Mpc away from Earth (Abbott et al. 2017*b*), and that the black holes in GW150914 had masses of 35^{+5}_{-3} $M_\odot$ and 33^{+3}_{-4} $M_\odot$ (Abbott et al. 2016*b*). We use it to determine the Hubble constant (Abbott et al. 2017*d*), to study the formation mechanism of black hole binaries (Vitale et al. 2017; Stevenson, Berry, & Mandel 2017; Talbot & Thrane 2017; Gerosa & Berti 2017; Farr et al. 2017; Wysocki, Lange, & O'Shaughnessy 2018; Lower et al. 2018), and to probe how stars die (Fishbach & Holz 2017; Talbot & Thrane 2018; Abbott et al. 2018*a*). Increasingly, Bayesian inference and parameter estimation are the language of gravitational-wave astronomy. In this note, we endeavour to provide a primer on Bayesian inference with examples from gravitational-wave astronomy.[a]

Before beginning, we highlight additional resources, useful for researchers interested in Bayesian inference in gravitational-wave astronomy. Sivia & Skilling (2006) and Gregory (2005) are useful references that are accessible to physicists and astronomers (see also the Springer Series in Astrostatistics; Manuel et al. 2012; Hilbe 2013; Chattopadhyay & Chattopadhyay 2014; Andreon & Weaver 2015). The chapter in Hilbe (2013) by Loredo discusses hierarchical models, but refers to them as 'multilevel' models (Loredo 2012). Seasoned veterans may find Gelman et al. (2013) to be a thorough reference.

[a]This review focuses on Bayesian inference applied to audio-band gravitational waves from compact binary coalescence, the only source of gravitational waves yet detected. We note in passing that Bayesian inference has been applied to study gravitational waves from rotating neutron stars (Umstätter et al. 2004; Dupuis & Woan 2005; Abbott et al. 2017*e*), bursting sources (Cornish & Littenberg 2015; Logue et al. 2012; Jade Powell et al. 2016), and stochastic backgrounds (Mandic et al. 2012; Callister et al. 2017; Abbott et al. 2018*b*). Bayesian inference methods have also been developed for space-based observatories observing at millihertz frequencies (Babak et al. 2008, 2010) and for pulsar timing arrays operating at nanohertz frequencies (Lentati et al. 2014; Vigeland & Vallisneri 2014).

Author for correspondence: Eric Thrane, Email: eric.thrane@monash.edu
Cite this article: Thrane E and Talbot C. (2019) An introduction to Bayesian inference in gravitational-wave astronomy: Parameter estimation, model selection, and hierarchical models *Publications of the Astronomical Society of Australia* **36**, e010, 1–12. https://doi.org/10.1017/pasa.2019.2

2. Fundamentals: likelihoods, priors, and posteriors

A primary aim of modern Bayesian inference is to construct a posterior distribution

$$p(\theta|d), \tag{1}$$

where θ is the set of model parameters and d is the data associated with a measurement.[b] For illustrative purposes, let us say that θ are the 15 parameters describing a binary black hole coalescence and d is the strain data from a network of gravitational-wave detectors. The posterior distribution $p(\theta|d)$ is the probability density function for the continuous variable θ given the data d. The probability that the true value of θ is between $(\theta', \theta' + d\theta)$ is given by $p(\theta'|d)d\theta'$. It is normalised so that

$$\int d\theta\, p(\theta|d) = 1. \tag{2}$$

The posterior distribution is what we use to construct credible intervals that tell us, for example, the component masses of a binary black hole event like GW150914. For details about the construction of credible intervals, see Appendix A.

According to Bayes theorem, the posterior distribution is given by

$$p(\theta|d) = \frac{\mathcal{L}(d|\theta)\,\pi(\theta)}{\mathcal{Z}}, \tag{3}$$

where $\mathcal{L}(d|\theta)$ is the likelihood function of the data given the parameters θ, $\pi(\theta)$ is the prior distribution for θ, and $\mathcal{Z}$ is a normalisation factor[c,d] called the 'evidence'

$$\mathcal{Z} \equiv \int d\theta\, \mathcal{L}(d|\theta)\,\pi(\theta). \tag{4}$$

The likelihood function is something that we choose. It is a description of the measurement. By writing down a likelihood, we implicitly introduce a noise model. For gravitational-wave astronomy, we typically assume a Gaussian-noise likelihood function that looks something like this

$$\mathcal{L}(d|\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(d-\mu(\theta))^2}{\sigma^2}\right), \tag{5}$$

where $\mu(\theta)$ is a template for the gravitational strain waveform given θ and σ is the detector noise. Note that π with no parentheses and no subscript is the mathematical constant, not a prior distribution. This likelihood function reflects our assumption that the noise in gravitational-wave detectors is Gaussian.[e] Note that the likelihood function is not normalised with respect to θ and so[f]

$$\int d\theta\, \mathcal{L}(d|\theta) \neq 1. \tag{6}$$

For a more detailed discussion of the Gaussian noise likelihood in the context of gravitational-wave astronomy, see Appendix B.

Like the likelihood function, the prior is something we get to choose. The prior incorporates our belief about θ before we carry out a measurement. In some cases, there is an obvious choice of prior. For example, if we are considering the sky location of a binary black hole merger, it is reasonable to choose an isotropic prior that weights each patch of sky as equally probable. In other situations, the choice of prior is not obvious. For example, before the first detection of gravitational waves, what would have been a suitable choice for the prior on the primary[g] black hole mass $\pi(m_1)$? When we are ignorant about θ, we often express our ignorance by choosing a distribution that is either uniform or log-uniform.[h]

While θ may consist of a large number of parameters, we usually want to look at just one or two at a time. For example, the posterior distribution for a binary black hole merger is a fifteen-dimensional[i] function that includes information about black hole masses, sky location, spins, etc. What if we want to look at the posterior distribution for just the primary mass? To answer this question we *marginalise* (integrate) over the parameters that we are not interested in (called 'nuisance parameters') so as to obtain a marginalised posterior

$$p(\theta_i|d) = \int \left(\prod_{k\neq i} d\theta_k\right) p(\theta|d) \tag{7}$$

$$= \frac{\mathcal{L}(d|\theta_i)\,\pi(\theta_i)}{\mathcal{Z}}. \tag{8}$$

The quantity $\mathcal{L}(d|\theta_i)$ is called the 'marginalised likelihood'. It can be expressed as follows:

$$\mathcal{L}(d|\theta_i) = \int \left(\prod_{k\neq i} d\theta_k\right) \pi(\theta_k)\,\mathcal{L}(d|\theta). \tag{9}$$

When we marginalise over one variable θ_a in order to obtain a posterior on θ_b, we are calculating our best guess for θ_b given uncertainty in θ_a. Speaking somewhat colloquially, if θ_a and θ_b are covariant, then marginalising over θ_a 'injects' uncertainty into the posterior for θ_b. When this happens, the marginalised posterior $p(\theta_b|d)$ is broader than the *conditional posterior* $p(\theta_b|d, \theta_a)$.

[b] By referring to 'model parameters', we are implicitly acknowledging that we begin with some model. Some authors make this explicit by writing the posterior as $p(\theta|d, M)$, where M is the model. (Other authors sometimes use I to denote the model.) We find this notation clunky and unnecessary since it goes without saying that one must always assume *some* model. If/when we consider two *distinct* models, we add an additional variable to denote the model.

[c] In this document, we use different symbols for different distributions: p for posteriors, $\mathcal{L}$ for likelihoods, and π for priors. We advocate this notation, since it highlights what is what and makes formulas easy to read. However, it is by no means standard, and some authors will use p for any and all probability distributions.

[d] For now, we treat the evidence as 'just' a normalisation factor, though, below we see that it plays an important role in model selection, and that it can be understood as a marginalised likelihood.

[e] The Gaussian noise assumption is a good starting point for describing the strain noise in gravitational-wave detectors. The combined effect of many random noise processes tends to produce nearly Gaussian strain noise. Of course, the noise description can be generalised to include non-Gaussian glitches, drift over time, and instrumental lines all of which can be described by noise parameters (see e.g. Littenberg & Cornish 2015; Röver, Meyer, & Christensen 2011).

[f] Given that the likelihood is not normalised with respect to θ, one might ask in what way it *is* normalised. The answer is that the likelihood is normalised with respect to the *data* d. Before we collect any data, the likelihood describes the chance of getting data d. It is a probability density function with units of inverse data. The integral over all possible d is unity. Once we obtain actual data, d is, of course, fixed.

[g] The 'primary' black hole is the heavier of two black holes in a binary, which is contrasted with the lighter 'secondary' black hole.

[h] A log uniform distribution is used when we do not know the order of magnitude of some quantity, for example, the energy density of primordial gravitational waves.

[i] There are eight 'intrinsic' parameters, which are fundamental properties of the binary: primary mass m_1, secondary mass m_2, primary dimensionless spin vector $\vec{s}_1$, and secondary dimensionless spin vector $\vec{s}_2$. The other seven parameters are 'extrinsic', relating to how we view the binary. The extrinsic parameters are as follows: inclination angle ι, polarisation angle ψ, phase at coalescence ϕ_c, right ascension RA, declination DEC, luminosity distance D_L, and time of coalescence t.

https://doi.org/10.1017/pasa.2019.2 Published online by Cambridge University Press

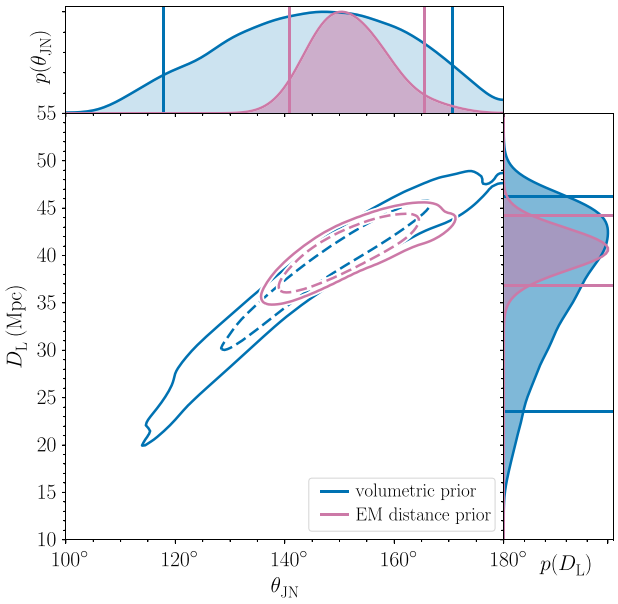

Figure 1: The joint posterior for luminosity distance and inclination angle for GW170817 from Abbott et al. (2017*a*). The blue contours show the credible region obtained using gravitational-wave data alone. The purple contours show the smaller credible region obtained by employing a relatively narrow prior on distance obtained with electromagnetic measurements. Publicly available posterior samples for this plot are available here: LIGO/Virgo (LIGO/Virgo).

The conditional posterior $p(\theta_b|d,\theta_a)$ represents a slice through the $p(\theta_b|d)$ posterior at a fixed value of θ_a.

This is nicely illustrated with an example. There is a well-known covariance between the luminosity distance of a merging compact binary from Earth D_L and the inclination angle θ_{JN}. For the binary neutron star coalescence GW170817, we are able to constrain the inclination angle much better when we use the known distance and sky location of the host galaxy compared to the constraint obtained using the gravitational-wave measurement alone.[j] Results from (Abbott et al. 2017*a*) are shown in Figure 1.

3. Models, evidence and odds

In Eq. (4), reproduced here, we defined the Bayesian evidence:

$$\mathcal{Z} \equiv \int d\theta \mathcal{L}(d|\theta)\,\pi(\theta).$$

In practical terms, the evidence is a single number. It usually does not mean anything by itself, but becomes useful when we compare one evidence with another evidence. Formally, the evidence is a likelihood function. Specifically, it is the completely marginalised likelihood function. It is therefore sometimes denoted $\mathcal{L}(d)$ with no θ dependence. However, we prefer to use $\mathcal{Z}$ to denote the fully marginalised likelihood function.

Above, we described how the evidence serves as a normalisation constant for the posterior $p(\theta|d)$. However, the evidence is also used to do model selection. Model selection answers the question: Which model is statistically preferred by the data and by how much? There are different ways to think about models. Let us return to the case of binary black holes. We may compare a 'signal model' in which we suppose that there is a binary black hole signal present in the data with a prior $\pi(\theta)$ to the 'noise model', in which we suppose that there is no binary black hole signal present. While the signal model is described by the fifteen binary parameters θ, the noise model is described by no parameters. Thus, we can define a signal evidence $\mathcal{Z}_S$ and a noise evidence $\mathcal{Z}_N$

$$\mathcal{Z}_S \equiv \int d\theta \mathcal{L}(d|\theta)\,\pi(\theta), \tag{10}$$

$$\mathcal{Z}_N \equiv \mathcal{L}(d|0), \tag{11}$$

where

$$\mathcal{L}(d|0) \equiv \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{h^2}{\sigma^2}\right). \tag{12}$$

The noise evidence $\mathcal{Z}_N$ is sometimes referred to as the 'null likelihood'.

The ratio of the evidence for two different models is called the Bayes factor. In this example, the signal/noise Bayes factor is

$$\mathrm{BF}^S_N \equiv \frac{\mathcal{Z}_S}{\mathcal{Z}_N}. \tag{13}$$

It is often convenient to work with the log of the Bayes factor:[k]

$$\log \mathrm{BF}^S_N \equiv \log(\mathcal{Z}_S) - \log(\mathcal{Z}_N). \tag{14}$$

When the absolute value of log BF is large, we say that one model is preferred over the other. The sign of log BF tells us which model is preferred. A threshold of $|\log \mathrm{BF}| = 8$ is often used as the level of 'strong evidence' in favour of one hypothesis over another (Jeffreys 1961).

The signal/noise Bayes factor is just one example of a Bayes factor comparing two models. We can calculate a Bayes factor comparing identical models but with different priors. For example, we can calculate the evidence for a binary black hole with a uniform prior on dimensionless spin and compare that to the evidence obtained using a zero-spin prior. The Bayes factor comparing these models would tell us if the data prefer spin:

$$\mathcal{Z}_{\mathrm{spin}} = \int d\theta \mathcal{L}(d|\theta)\,\pi(\theta), \tag{15}$$

$$\mathcal{Z}_{\mathrm{no\ spin}} = \int d\theta \mathcal{L}(d|\theta)\,\pi_{\mathrm{no\ spin}}(\theta), \tag{16}$$

where $\pi_{\mathrm{no\ spin}}(\theta)$ is a prior with zero spins. The spin/no spin Bayes factor is

$$\mathrm{BF}^{\mathrm{spin}}_{\mathrm{no\ spin}} = \frac{\mathcal{Z}_{\mathrm{spin}}}{\mathcal{Z}_{\mathrm{no\ spin}}}. \tag{17}$$

We may also compare two disparate signal models. For example, we can compare the evidence for a binary black hole waveform predicted by general relativity (model M_A with parameters θ) with a binary black hole waveform predicted by some other theory (model M_B with parameters ν):

$$\mathcal{Z}_A = \int d\theta \mathcal{L}(d|\theta, M_A)\,\pi(\theta), \tag{18}$$

$$\mathcal{Z}_B = \int d\nu \mathcal{L}(d|\nu, M_B)\,\pi(\nu). \tag{19}$$

The A/B Bayes factor is

$$\mathrm{BF}^A_B = \frac{\mathcal{Z}_A}{\mathcal{Z}_B}. \tag{20}$$

[j] The viewing angle $= \Theta = \min(\theta_{JN}, 180° - \theta_{JN}$ is constrained to be $< 28°$ with the electromagnetic counterpart, and $< 55°$ without it (Abbott et al. 2017*c*).

[k] A typical log evidence might be -5000, which evaluates to zero when exponentiated on a computer. Functions such as `logsumexp` can be useful for combining evidence.

https://doi.org/10.1017/pasa.2019.2 Published online by Cambridge University Press

Note that the number of parameters in ν can be different from the number of parameters in θ.

Our presentation of model selection so far has been a bit fast and loose. Formally, the correct metric to compare two models is not the Bayes factor, but rather the odds ratio

$$\mathcal{O}^A_B \equiv \frac{\mathcal{Z}_A}{\mathcal{Z}_B}\frac{\pi_A}{\pi_B}. \tag{21}$$

The odds ratio is the product of the Bayes factor with the prior odds π_A/π_B, which describes our prior belief about the relative likelihood of hypotheses A and B. In many practical applications, we set the prior odds ratio to unity, and so the odds ratio *is* the Bayes factor. This practice is sensible in many applications where our intuition tells us: until we do this measurement both hypotheses are equally likely.[l]

Bayesian evidence encodes two pieces of information. First, the likelihood tells us how well our model fits the data. Second, the act of marginalisation tell us about the size of the volume of parameter space we used to carry out a fit. This creates a sort of tension. We want to get the best fit possible (high likelihood) but with a minimum prior volume. A model with a decent fit and a small prior volume often yields a greater evidence than a model with an excellent fit and a huge prior volume. In these cases, the Bayes factor penalises the more complicated model for being too complicated.

This penalty is called an Occam factor. It is a mathematical formulation of the statement that all else equal, a simple explanation is more likely than a complicated one. If we compare two models where one model is a superset of the other—for example, we might compare general relativity and general relativity with non-tensor modes—and if the data are better explained by the simpler model, the log Bayes factor is typically modest, $\log \mathrm{BF} \approx (-2, -1)$. Thus, it is difficult to completely rule out extensions to existing theories. We just obtain ever tighter constraints on the extended parameter space.

4. Samplers

Thanks to the creation of phenomenological gravitational waveforms (called 'approximants'), it is now computationally straightforward to make a prediction about what the data d should look like given some parameters θ. That is a forward problem. Calculating the posterior, the probability of parameters θ given the data as in Eq. (3), reproduced here, is a classic inverse problem:[m]

$$p(\theta|d) = \frac{\mathcal{L}(d|\theta)\,\pi(\theta)}{\mathcal{Z}}.$$

In general, inverse problems are computationally challenging compared to forward problems. To illustrate why let us imagine that we wish to calculate the posterior probability for the 15 parameters describing a binary black hole merger. If we do this naively, we might create a grid with 10 bins in every dimension and evaluate the likelihood at each grid point. Even with this coarse resolution, our calculation suffers from 'the curse of dimensionality'. It is computationally prohibitive to carry out 10^{15} likelihood evaluations. The problem becomes worse as we add dimensions. As a rule of thumb, brute-force bin approaches become painful once one exceeds three dimensions.

The solution is to use a stochastic sampler (although recent work has shown progress carrying out these calculations using the alternative technique of iterative fitting; Pankow et al. 2015; Lange, O'Shaughnessy, & Rizzo 2018). Commonly used sampling algorithms can be split into two broad categories of method: Markov chain Monte Carlo (MCMC) (Metropolis et al. 1953; Hastings 1970) and nested sampling (Skilling 2004). These algorithms generate a list of posterior samples $\{\theta\}$ drawn from the posterior distribution such that the number of samples on the interval $(\theta, \theta + \Delta\theta) \propto p(\theta)$ (Veitch et al. 2015). Some samplers also produce an estimate of the evidence. We can visualise the posterior samples as a spreadsheet. Each column is a different parameter, for example, primary black hole mass, secondary black hole mass, etc. For binary black hole mergers, there are typically fifteen columns. Each row represents a different posterior sample.

Posterior samples have two useful properties. First, they can be used to compute expectation values of quantities of interest since (Hogg & Foreman-Mackey 2018)

$$\langle f(x)\rangle_{p(x)} = \int dx\, p(x) f(x) \approx \frac{1}{n_s}\sum_k^{n_s} f(x_k), \tag{22}$$

where $p(x)$ is the posterior distribution that we are sampling, $f(x)$ is some function we want to find the expectation value of, and the sum over k runs over n_s posterior samples. Eq. (22) will prove useful simplifying our calculation of the likelihood of data given hyper-parameters.

The second useful property of posterior samples is that, once we have samples from an N-dimensional space, we can generate the marginalised probability for any subset of the parameters by simply selecting the corresponding columns in our spreadsheet. This property is used to help visualise the output of these samplers by constructing 'corner plots', which show the marginalised one- and two-dimensional posterior probability distributions for each of the parameters. For an example of a corner plot, see Figure 1. A handy python package exists for making corner plots (Foreman-Mackey 2016).

4.1. MCMC

MCMC sampling was first introduced in Metropolis et al. (1953) and extended in Hastings (1970). For a recent overview of MCMC methods in astronomy, see Sharma (2017). In MCMC methods, particles undergo a random walk through the posterior distribution where the probability of moving to any given point is determined by the transition probability of the Markov chain. By noting the position of the particles—or 'walkers' as they are sometimes called—at each iteration, we generate draws from the posterior probability distribution.

There are some subtleties that must be considered when using MCMC samplers. First, the early-time behaviour of MCMC walkers is strongly dependent on the initial conditions. It is therefore necessary to include a 'burn-in' phase to ensure that the walker has settled into a steady state before beginning to

[l] There are some (fairly uncommon) examples where we might choose a different prior odds ratio. For example, we may construct a model in which general relativity (GR) is wrong. We may further suppose that there are multiple different ways in which it could be wrong, each corresponding to a different GR-is-wrong sub-hypothesis. If we calculated the odds ratio comparing one of these GR-is-wrong sub-hypotheses to the GR-is-right hypothesis, we would not assign equal prior odds to both hypotheses. Rather, we would assign at most 50% probability to the entire GR-is-wrong hypothesis, which would then have to be split among the various sub-hypotheses.

[m] We note here a few early papers important in the development of Bayesian inference tools for gravitational-wave astronomy. Initial implementation of MCMC methods for spinning binaries was carried out in van der Sluys et al. (2008*a*). The first demonstration of Bayesian parameter estimation for spinning binaries was performed in van der Sluys et al. (2008*b*). Veitch & Vecchio (2008) demonstrated Bayesian model selection for compact binaries.

https://doi.org/10.1017/pasa.2019.2 Published online by Cambridge University Press

accumulate samples from the posterior distribution. Once the walker has reached a steady state, the algorithm can continue indefinitely and so it is necessary for the user to define a termination condition. This is typically chosen to be when enough samples have been acquired for the user to believe an accurate representation of the posterior has been obtained. Thus, MCMC requires a degree of artistry, developed from experience.

Additionally, the positions of a walker in a chain are often autocorrelated. Because of this correlation, the positions of the walkers do not represent a faithful sampling from the posterior distribution. If no remedy is applied, the width of the posterior distribution is underestimated. It is thus necessary to 'thin' the chain by selecting samples separated by the autocorrelation length of the chain.

MCMC walkers can also fail to find multiple modes of a posterior distribution if there are regions of low posterior probability between the modes. However, this can be mitigated by running many walkers which begin exploring the space at different points. This also demonstrates a simple way to parallelise MCMC computations to quickly generate many samples. Many variants of MCMC sampling have been proposed in order to improve the performance of MCMC algorithms with respect to these and other issues. For a more in-depth discussion of MCMC methods, see e.g. chapter 11 of Gelman et al. (2013), or Hogg & Foreman-Mackey (2018). The most widely used MCMC code in astronomy is EMCEE (Foreman-Mackey et al. 2013).[n]

4.2. Nested sampling

The first widely used alternative to MCMC was introduced by Skilling (2004). While MCMC methods are designed to draw samples from the posterior distribution, nested sampling is designed to calculate the evidence. Generating samples from the posterior distribution is a by-product of the nested sampling evidence calculation algorithm. By weighting each of the samples used to calculate the evidence by the posterior probability of the sample, nested samples are converted into posterior samples.

Nested sampling works by populating the parameter space with a set of 'live points' drawn from the prior distribution. At each iteration, the lowest likelihood point is removed from the set of live points and new samples are drawn from the prior distribution until a point with higher likelihood than the removed point is found. The evidence is evaluated by assigning each removed point a prior volume and then computing the sum of the likelihood multiplied by the prior volume for each sample.

Since the nested sampling algorithm continually moves to higher likelihood regions, it is possible to estimate an upper limit on the evidence at each iteration. This is done by imagining that the entire remaining prior volume has a likelihood equal to that of the highest likelihood live point. This is used to inform the termination condition for the nested sampling algorithm. The algorithm stops when the current estimate of the evidence is above a certain fraction of the estimated upper limit.[o] Unlike MCMC algorithms nested sampling is not straightforwardly parallelisable, and posterior samples do not accumulate linearly with run time.

[n] http://dfm.io/emcee/.

[o] In practice, this is expressed as the difference between the calculated log evidence and the upper limit of the log evidence.

5. Hyper-parameters and hierarchical models

As more and more gravitational-wave events are detected, it is increasingly interesting to study the *population properties* of binary black holes and binary neutron stars. These are the properties common to all of the events in some set. Examples include the neutron star equation of state and the distribution of black hole masses. Hierarchical Bayesian inference is a formalism, which allows us to go beyond individual events in order to study population properties.[p]

The population properties of some set of events is described by the shape of the prior. For example, two population synthesis models might yield two different predictions for the prior distribution of the primary black hole mass $\pi(m_1)$. In order to probe the population properties of an ensemble of events, we make the prior for θ conditional on a set of 'hyper-parameters' Λ:

$$\pi(\theta|\Lambda). \tag{23}$$

The hyper-parameters parameterise the shape of the prior distribution for the parameters θ. An example of a (parameter, hyper-parameter) relationship is (θ = primary black hole mass m_1, Λ = the spectral index of the primary mass spectrum α). In this example

$$\pi(m_1|\alpha) \propto m_1^{\alpha}. \tag{24}$$

A key goal of population inference is to estimate the posterior distribution for the hyper-parameters Λ. In order to do this, we marginalise over the entire parameter space θ in order to obtain a marginalised likelihood:

$$\mathcal{L}(d|\Lambda) = \int d\theta\, \mathcal{L}(d|\theta)\, \pi(\theta|\Lambda). \tag{25}$$

Normally, we would call this completely marginalised likelihood an evidence, but because it still depends on Λ, we call it the likelihood for the data d given the hyper-parameters Λ. The hyper-posterior is given simply by

$$p(\Lambda|d) = \frac{\mathcal{L}(d|\Lambda)\, \pi(\Lambda)}{\int d\Lambda\, \mathcal{L}(d|\Lambda)\, \pi(\Lambda)}. \tag{26}$$

Note that we have introduced a hyper-prior $\pi(\Lambda)$, which describes our prior belief about the hyper-parameters Λ. The term in the denominator

$$\mathcal{Z}_{\Lambda} \equiv \int d\Lambda\, \mathcal{L}(d|\Lambda)\, \pi(\Lambda) \tag{27}$$

is the 'hyper-evidence', which we denote $\mathcal{Z}_{\Lambda}$ in order to distinguish it from the regular evidence $\mathcal{Z}_{\theta}$. In Appendix D we discuss posterior predictive distributions (PPD), which represent the updated prior on θ in light of the data d and given some hyper-parameterisation.

We now generalise the discussion of hyper-parameters in order to handle the case of N independent events. In this case, the total likelihood for all N events $\mathcal{L}_{\rm tot}$ is simply the product of each individual likelihood

$$\mathcal{L}_{\rm tot}(\vec{d}|\vec{\theta}) = \prod_{i}^{N} \mathcal{L}(d_i|\theta_i). \tag{28}$$

[p] Possibly the earliest papers proposing to measure *distributions* of gravitational-wave parameters are Mandel & O'Shaughnessy (2010) and Mandel (2010), while hierarchical Bayesian inference was introduced to study the population properties of sources of gravitational waves in Adams, Cornish, & Littenberg (2012).

https://doi.org/10.1017/pasa.2019.2 Published online by Cambridge University Press

Here, we use vector notation so that $\vec{d}$ is the set of measurements of N events, each of which has its own parameters, which make up the vector $\vec{\theta}$. Since we suppose that every event is drawn from the same population prior distribution—hyper-parameterised by Λ—the total marginalised likelihood is

$$\mathcal{L}_{\rm tot}(\vec{d}|\Lambda) = \prod_i^N \int d\theta_i \, \mathcal{L}(d_i|\theta_i)\, \pi(\theta_i|\Lambda). \tag{29}$$

The associated (hyper-) posterior is

$$p_{\rm tot}(\Lambda|\vec{d}) = \frac{\mathcal{L}_{\rm tot}(\vec{d}|\Lambda)\, \pi(\Lambda)}{\int d\Lambda\, \mathcal{L}_{\rm tot}(\vec{d}|\Lambda)\, \pi(\Lambda)}. \tag{30}$$

The denominator, of course, is the total hyper-evidence:

$$\mathcal{Z}_\Lambda^{\rm tot} = \int d\Lambda\, \mathcal{L}_{\rm tot}(\vec{d}|\Lambda)\, \pi(\Lambda). \tag{31}$$

We may calculate the Bayes factor comparing different hyper-models in the same way that we calculate the Bayes factor for different models.

Examining Eq. (31), we see that the total hyper-evidence involves a large number of integrals. For the case of binary black hole mergers, every event has 15 parameters, and so the dimension of the integral is $15N + M$ taking where M is the number of hyper-parameters in Λ. As N gets large, it becomes difficult to sample such a large prior volume all at once. Fortunately, it is possible to break the integral into individual integrals for each event, which are then combined through a process sometimes referred to as 'recycling'.

Thus, the total marginalised likelihood in Eq. (29) can be written as follows:

$$\boxed{\mathcal{L}_{\rm tot}(\vec{d}|\Lambda) = \prod_i^N \frac{\mathcal{Z}_\varnothing(d_i)}{n_i} \sum_k^{n_i} \frac{\pi(\theta_i^k|\Lambda)}{\pi(\theta_i^k|\varnothing)}}. \tag{32}$$

Here, the sum over k is a sum over the n_i posterior samples associated with event i. The posterior samples for each event are generated with some default prior $\pi(\theta_k|\varnothing)$. The default prior is ultimately canceled from the final answer, so it not so important what we choose for the default prior so long as it is sufficiently uninformative. Using the $\varnothing$ prior, we obtain an evidence $\mathcal{Z}_\varnothing$. In this way, we are able to analyse each event individually before recycling the posterior samples to obtain a likelihood of the data given Λ.

To see where this formula comes from, we note that

$$p(\theta_i|d_i, \varnothing) = \frac{\mathcal{L}(d_i|\theta_i)\, \pi(\theta_i|\varnothing)}{\mathcal{Z}_\varnothing(d_i)}. \tag{33}$$

Rearranging terms,

$$\mathcal{L}(d_i|\theta_i) = \mathcal{Z}_\varnothing(d_i)\, \frac{p(\theta_i|d_i, \varnothing)}{\pi(\theta_i|\varnothing)}. \tag{34}$$

Plugging this into Eq. (29), we obtain[q]

$$\mathcal{L}_{\rm tot}(\vec{d}|\Lambda) = \prod_i^N \int d\theta_i\, p(\theta_i|d_i, \varnothing)\, \mathcal{Z}_\varnothing(d_i)\, \frac{\pi(\theta_i|\Lambda)}{\pi(\theta_i|\varnothing)}. \tag{35}$$

Finally, we use Eq. (22) to convert the integral over θ_i to a sum over posterior samples, thereby arriving at Eq. (32).

All of the results derived up until this point ignore selection effects where an event with parameters θ_1 is easier to detect than an event with parameters θ_2. There are cases where selection effects are important. For example, the visible volume for binary black hole mergers scales as approximately $V \propto M^{2.1}$, which means that higher mass mergers are relatively easier to detect than lower mass mergers (Fishbach & Holz 2017). In Appendix E, we show how this method is extended to accommodate selection effects.

[q]One 'recycles' the posterior samples generated using the $\pi(\theta_i|\varnothing)$ prior in order to do something new with the hyper-parameterised prior $\pi(\theta_i|\Lambda)$.

Author ORCIDs. Eric Thrane https://orcid.org/0000-0002-4418-3895 , Colm Talbot https://orcid.org/0000-0003-2053-5582 .

Acknowledgements. This document is the companion paper to a lecture at the 2018 OzGrav Inference Workshop held July 16–18, 2018 at Monash University in Clatyon, Australia. Thank you to the organisers: Greg Ashton, Paul Lasky, Hannah Middleton, and Rory Smith. This work was supported by OzGrav through the Australian Research Council CE170100004. For helpful comments on a draft of this manuscript, we thank Sylvia Biscoveanu, Paul Lasky, Nikhil Sarin, and Shanika Galaudage. We are indebted to Will Farr who clarified our understanding of selection effects and to Patricia Schmidt who helped our understanding of phase marginalisation. We thank Rory Smith and John Veitch for drawing our attention to the explicit distance marginalisation work of Leo Singer. Finally, we thank the anonymous referee for helpful suggestions, which improved the manuscript. ET and CT are supported by CE170100004. ET is supported by FT150100281.

References

Abbott, B. P., et al. 2016a, PhRvX, 6, 041015
Abbott, B. P., et al. 2016b, PhRvL, 116, 061102
Abbott, B. P., et al. 2017a, PhRvX, 9, 011001
Abbott, B. P., et al. 2017b, PhRvL, 118, 221102
Abbott, B. P., et al. 2017c, PhRvL, 119, 161101
Abbott, B. P., et al. 2017d, Nature, 551, 85
Abbott, B. P., et al. 2017e, ApJ, 839, 12
Abbott, B. P., et al. 2018a, Binary Black Hole Population Properties Inferred from the First and Second Observing Runs of Advanced LIGO and Advanced Virgo. https://arxiv.org/abs/1811.12940
Abbott, B. P., et al. 2018b, PhRvL, 120, 201102
Adams, M., Cornish, N., & Littenberg, T. 2012, PhRvD, 86, 124032
Anderson, W. G., Brady, P. R., Creighton, J. D. E., & Flanagan, É. É. 2001, PhRvD, 63, 042003
Andreon, S. & Weaver, B. 2015, Bayesian Methods for the Physical Sciences (1st edn.; Switzerland: Springer)
Babak, S., et al. 2008, Class. Quant. Grav., 25, 184026
Babak, S., et al. 2010, Class. Quant. Grav., 27, 084009
Blackman, J., Field, S. E., Scheel, M. A., Galley, C. R., Hemberger, D. A., Schmidt, P., & Smith, R. 2017, PhRvD, 95, 104023
Callister, T., et al. 2017, PhRvX, 7, 041058
Canizares, P., Field, S. E., Gair, J. R., & Tiglio, M. 2013, PhRvD, 87, 124005
Chattopadhyay, A. K., & Chattopadhyay, T. 2014, Statistical Methods for Astronomical Data Analysis (1st edn.; New York: Springer)
Cornish, N. J., & Littenberg, T. B. 2015, Class. Quant. Grav., 32, 135012
Cutler, C., & Flanagan, É. E. 1994, PhRvD, 49, 2658
Damour, T., Iyer, B. R., & Sathyaprakash, B. S. 2005, PhRvD, 63, 044023
Dupuis, R. J., & Woan, G. 2005, PhRvD, 72, 102002
Farr, W. M. 2014, Marginalisation of the Time Parameter in Gravitational Wave Parameter Estimation. https://dcc.ligo.org/T1400460-v2/public
Farr, W. M., Stevenson, S., Miller, M. C., Mandel, I., Farr, B., & Vecchio, A. 2017, Nature, 548, 426
Fishbach, M., & Holz, D. E. 2017, ApJL, 851, L25
Fishbach, M., Holz, D. E., & Farr, W. M. 2018, ApJL, 863, L41
Foreman-Mackey, D. 2016, J. Open Source Softw., 24
Foreman-Mackey, D., Hogg, D. W., Lang, D., & Goodman, J. 2013, PASP, 125, 306
Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. 2013, Bayesian Data Analysis (3rd ed.; Boca Raton, FL: Chapman & Hall/CRC Texts in Statistical Science, Taylor & Francis), https://books.google.com.au/books?id=ZXL6AQAAQBAJ

https://doi.org/10.1017/pasa.2019.2 Published online by Cambridge University Press

Gerosa, D. & Berti, E. 2017, PhRvD, 95, 124046
Gregory, P. 2005, Bayesian Logical Data Analysis for the Physical Sciences (1st edn.; Cambridge, England: Cambridge University Press)
Hannam, M., Schmidt, P., Bohé, A., Haegel, L., Husa, S., Ohme, F., Pratten, G., & Pürrer, M. 2014, PhRvL, 113, 151101
Hastings, W. K. 1970, Biometrika, 57, 97
Hilbe, J. M., ed. 2013, Astrostatistical Challenges for the New Astronomy (1st edn.; New York: Springer)
Hogg, D. W., & Foreman-Mackey, D. 2018, ApJS, 236, 11
Jade Powell, S. E. G., Logue, J., Heng, I. S. 2016, PhRvD, 94, 123012
Jeffreys, H. 1961, Theory of Probability (3rd edn.; Oxford, England: Oxford)
Khan, S., Husa, S., Hannam, M., Ohme, F., Pürrer, M., Forteza, X. J., & Bohé, A. 2016, PhRvD, 93, 044007
Lange, J., O'Shaughnessy, R., & Rizzo, M. 2018
Lentati, L., Alexander, P., Hobson, M. P., Feroz, F., van Haasteren, R., Lee, K., & Shannon, R. M. 2014, MNRAS, 437, 3004
LIGO/Virgo, Properties of the Binary Neutron Star Merger GW170817. https://dcc.ligo.org/LIGO-P1800061/public
Littenberg, T. B., & Cornish, N. J. 2015, PhRvD, 91, 084034
Logue, J., Ott, C. D., Heng, I. S., Kalmus, P., & Scargill, J. H. C. 2012, PhRvD, 86, 044023
Loredo, T. J. 2012, Bayesian astrostatistics: A backward look to the future, https://arxiv.org/abs/1208.3036
Lower, M. E., Thrane, E., Lasky, P. D., & Smith, R. 2018, PhRvD, 98, 083028
Mandel. I. 2010, PhRvD, 81, 084029
Mandel, I., Farr, W. M., & Gair, J. R. 2018, Extracting distribution parameters from multiple uncertain observations with selection biases, https://arxiv.org/abs/1809.02063
Mandel, I., & O'Shaughnessy, R. 2010, Class. Quant. Grav., 27, 114007
Mandic, V., Thrane, E., Giampanis, S., & Regimbau, T. 2012, PhRvL, 109, 171102
Manuel, L., Eyer, S., O'Mullane, W., Ridder, J. D., eds. 2012 (1st edn.; New York: Springer)
Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E., 1953, J. Chem. Phys., 21, 1087
Ng, K. K. Y., Vitale, S., Zimmerman, A., Chatziioannou, K., Gerosa, D., & Haster, C.-J. 2018, PhRvD, 98, 083007
Pankow, C., Brady, P., Ochsner, E., & O'Shaughnessy, R. 2015, PhRvD, 92, 023002
Pürrer, M. 2014, Class. Quant. Grav., 31, 195010
Röver, C., Meyer, R., & Christensen, N. 2011, Class. Quant. Grav., 28, 015010
Sharma, S. 2017, ARA&A, 55, 213
Sidery, T., et al. 2014, PhRvD, 89, 084060
Singer, L. P., & Price, L. R. 2016, PhRvD, 93, 024013
Singer, L. P., et al. 2016, ApJL, 829, L15
Sivia, D. S., & Skilling, J. 2006 (2nd edn.; Oxford, England: Oxford)
Skilling, J. 2004, Bayesian Inference and Maximum Entropy Methods in Science and Engineering, AIP Conf. Proc., ed. Rainer Fischer, Roland Preuss, & Udo von Toussaint (Melville, NY: American Institute of Physics) 735, 395
Smith, R., Field, S. E., Blackburn, K., & Haster, C.-J., PÃijrrer, M., Raymond, V., & Schmidt, P. 2016, PhRvD, 94, 044031
Smith, R., & Thrane, E. 2018, PhRvX, 8, 021019
Stevenson, S., Berry, C. P. L., & Mandel, I. 2017, MNRAS, 471, 2801
Talbot, C., & Thrane, E. 2017, PhRvD, 96, 023012
Talbot, C., & Thrane, E. 2018, ApJ, 856, 173
Umstätter, R., Meyer, R., Dupuis, R. J., Veitch, J., Woan, G., Christensen, N. 2004, Bayesian Inference and Maximum Entropy Methods in Science and Engineering, AIP Conf. Proc., ed. Rainer Fischer, Roland Preuss, & Udo von Toussaint (Melville, NY: American Institute of Physics) 735, 336
van der Sluys, M., Raymond, V., Mandel, I., Roever, C., Christensen, N., Kalogera, V., Meyer, R., & Vecchio, A. 2008a, Class. Quant. Grav., 25, 184011
van der Sluys, M. V., et al. 2008b, ApJL, 688, L61
Veitch, J., & Del Pozzo, W. 2013, Analytic Marginalisation of Phase Parameter, https://dcc.ligo.org/LIGO-T1300326/public
Veitch, J., & Vecchio, A. 2008, PhRvD, 78, 022001
Veitch, J., et al. 2015, PhRvD, 91, 042003
Vigeland, S. J., & Vallisneri, M. 2014, MNRAS, 440, 1446
Vitale, S., Lynch, R., Sturani, R., & Graff, P. 2017, Class. Quant. Grav., 34, 03LT01
Wysocki, D., Lange, J., & O'Shaughnessy, R. 2018

Appendix A. Credible intervals

It is often convenient to use the posterior to construct 'credible intervals', regions of parameter space containing some fraction of posterior probability. (Note that Bayesian inference yields credible intervals while frequentist inference yields *confidence intervals*.) For example, one can plot one-, two-, and three-sigma contours. By definition, a two-sigma credible region includes 95% of the posterior probability, but this requirement does not uniquely determine a single credible region. One well-motivated method for constructing confidence intervals is the highest posterior density interval (HPDI) method.

We can visualise the HPDI method as follows. We draw a horizontal line through a posterior distribution and calculate the area of above the line. If we move the line down, the area goes up. If we place the line such that the area is 95%, then the posterior above the line is the HPDI two-sigma credible interval. In general, the HPDI is neither symmetric nor unimodal. The advantage of HPDI over other methods is that it yields the minimum width credible interval. This method is sometimes referred to as 'draining the bathtub'.

Another commonly used method for calculating credible intervals is to construct symmetric intervals. Symmetric credible intervals are constructed using the cumulative distribution function,

$$P(x) = \int_{-\infty}^{x} dx'\, p(x'). \tag{A1}$$

The X% credible region is the region:

$$\frac{1}{2}\left(1 - \frac{X}{100}\right) < P(x) < \frac{1}{2}\left(1 + \frac{X}{100}\right). \tag{A2}$$

While symmetric credible intervals are simpler to construct than HPDI, particularly from samples drawn from a distribution, they can be misleading for multi-modal distributions and for distributions which peak near prior boundaries.

Credible intervals are useful for testing and debugging inference projects. Before applying an inference calculation to real data, it is useful to test it on simulated data. The standard test (see e.g. Sidery et al. 2014) is to simulate data d according to parameters $\theta_{\rm true}$ drawn at random from the prior distribution $\pi(\theta)$. Then, we analyse this data in order to obtain a posterior $p(\theta|d)$. The true value should fall inside the 90% credible interval 90% of the time. Testing that this is true provides a powerful validation of the inference algorithm. Note that we do not expect the posterior to peak precisely at $\theta_{\rm true}$, just within the one-sigma region.

Appendix B. Gaussian noise likelihood

In this appendix, we introduce additional notation that is helpful for talking about the Gaussian noise likelihood frequently used in gravitational-wave astronomy. In the main body of the manuscript, d has been taken to represent data. Now, we take d to represent the Fourier transform of the strain time series $d(t)$ measured by a gravitational-wave detector. In the language of computer programming,

$$d = \texttt{fft}\left(d(t)\right)/f_s, \tag{B1}$$

where f_s is the sampling frequency and `fft` is a Fast Fourier transform. The noise in each frequency bin is characterised by the single-sided noise power spectral density $P(f)$, which is proportional to strain squared and which has units of Hz^{-1}.

The likelihood for the data in a single frequency bin j given θ is

$$\mathcal{L}(d_j|\theta) = \frac{1}{\sqrt{2\pi P_j}} \exp\left(-2\Delta f \frac{\left|d_j - \mu_j(\theta)\right|^2}{P_j}\right), \tag{B2}$$

where Δf is the frequency resolution. The factor of $2\Delta f$ comes about from a factor of $1/2$ in the normal distribution and a factor of $4\Delta f$ needed to convert the square of the Fourier transforms into units of one-sided power spectral

https://doi.org/10.1017/pasa.2019.2 Published online by Cambridge University Press

density. The template $\mu(\theta)$ is related to the metric perturbation $h_{+,\times}(\theta)$ via antenna response factors $F_{+,\times}$ (Anderson et al. 2001):

$$\mu(\theta) = F_+(\mathrm{RA}, \mathrm{DEC}, \psi)h_+(\theta) + F_\times(\mathrm{RA}, \mathrm{DEC}, \psi)h_\times(\theta). \tag{B3}$$

Gravitational-wave signals are typically spread over many (M) frequency bins. Assuming the noise in each bin is independent, the combined likelihood is a product of the likelihoods for each bin:

$$\mathcal{L}(\mathbf{d}|\theta) = \prod_j^M \mathcal{L}(d_j|\theta), \tag{B4}$$

where $\mathbf{d}$ is the set of data including all frequency bins and d_j represents the data associated with frequency bin j. If we consider a measurement with multiple detectors, the product over j frequency bins gains an additional index l for each detector. Combining data from different detectors is like combining data from different frequency bins.

It is frequently useful to work with the log likelihood, which allows us to replace products with sums of logs. The log also helps dealing with small numbers. The log likelihood is

$$\begin{aligned}\log \mathcal{L}(\mathbf{d}|\theta) &= \sum_j^M \log \mathcal{L}(d_j|\theta) \\ &= -\frac{1}{2}\sum_j \log\left(2\pi P_j\right) - 2\Delta f \sum_j \frac{(d-\mu(\theta))^2}{P_j^2} \\ &= \Psi - \frac{1}{2}\langle d-\mu(\theta), d-\mu(\theta)\rangle.\end{aligned}$$

In the last line, we define the noise-weighted inner product[r] (Cutler & Flanagan 1994):

$$\langle a, b\rangle \equiv 4\Delta f \sum_j \Re\left(\frac{a_j^* b_j}{P_j}\right), \tag{B5}$$

and the constant

$$\Psi \equiv -\frac{1}{2}\sum_j \log\left(2\pi P_j\right). \tag{B6}$$

Since constants do not change the shape of the log likelihood, we often 'leave off' this normalising term and work with log likelihood minus Ψ. This is permissible as long as we do so consistently because when we take the ratio of two evidences—or equivalently, the difference of two log evidences—the Ψ factor cancels anyway. For the remainder of this appendix, we set $\Psi = 0$. Now that we have introduced the inner product notation, we are going to stop bold-facing the data d as it is implied that we are dealing with many frequency bins.

Using the inner product notation, we may expand out the log likelihood:

$$\begin{aligned}\log \mathcal{L}(d|\theta) &= -\frac{1}{2}\left[\langle d, d\rangle - 2\langle d, \mu(\theta)\rangle + \langle \mu(\theta), \mu(\theta)\rangle\right] \\ &= -\frac{1}{2}\left[-2\log \mathcal{Z}_N - 2\kappa^2(\theta) + \rho^2_{\mathrm{opt}}(\theta)\right] \\ &= \log \mathcal{Z}_N + \kappa^2(\theta) - \frac{1}{2}\rho^2_{\mathrm{opt}}(\theta).\end{aligned} \tag{B7}$$

We see that the log likelihood can be expressed with three terms. The first is proportional to the log noise evidence:

$$-2\log \mathcal{Z}_N \equiv \langle d, d\rangle. \tag{B8}$$

For debugging purposes, it is useful to keep in mind that if we calculate $-\log \mathcal{Z}_N$ on actual Gaussian noise (with $\Psi = 0$), we expect a typical value nearly equal to the number of frequency bins M (multiplied by the number of detectors) since each term in the inner product contributes a value close to unity.[s] We skip over the second term κ^2 for a moment. The third term is the optimal matched filter signal-to-noise ratio squared:

$$\rho^2_{\mathrm{opt}} \equiv \langle \mu, \mu\rangle. \tag{B9}$$

[r]Following the convention of gravitational-wave astronomy, our inner product is real by construction. However, below it will be useful to define a complex-valued inner product; see Eq. (C10).

[s]Specifically, the distribution of an ensemble of independent $-\ln \mathcal{Z}_N$ is a normal distribution with mean M and width $M^{1/2}$ where M is the number of frequency bins (multiplied by the number of detectors). This follows from the central limit theorem.

Returning now to the second term, we express κ^2 as the product of the matched filter signal-to-noise ratio and the optimal signal-to-noise ratio:

$$\begin{aligned}\kappa^2 &\equiv \langle d, \mu\rangle \\ &= \rho_{\mathrm{mf}}\,\rho_{\mathrm{opt}},\end{aligned} \tag{B10}$$

where

$$\rho_{\mathrm{mf}} \equiv \frac{\langle d, \mu\rangle}{\langle \mu, \mu\rangle^{1/2}}. \tag{B11}$$

Readers familiar with gravitational-wave astronomy are likely acquainted with the concept of matched filtering, which is the maximum likelihood technique for gravitational-wave detection. By writing the likelihood in this way, we highlight how parameter estimation is related to matched filtering. Rapid evaluation of the likelihood function in Eq. (B7) has been made possible through reduced order methods (Smith et al. 2016; Pürrer 2014; Canizares et al. 2013).

Appendix C. Explicitly marginalised likelihoods

The most computationally expensive step in computing the likelihood for compact binary coalescences is creating the waveform template (μ in Eq. (5)). This is done in two steps. The first step is to use the *intrinsic parameters* to calculate the metric perturbation. The second (much faster) step is to use the *extrinsic parameters* to project the metric perturbation onto the detector response tensor. In some cases, it is possible to reduce the dimensionality of the inverse problem—thereby speeding up calculations and improving convergence—by using a likelihood, which explicitly marginalises over extrinsic parameters. The improvement is especially marked for comparatively weak signals, which can be important for population studies (see e.g. Smith & Thrane 2018). In this appendix, we show how to calculate $\mathcal{L}_{\mathrm{marge}}$—a likelihood, which explicitly marginalise over coalescence time, phase at coalescence, and/or luminosity distance. We continue with notation introduced in Appendix B.

C.1. Time marginalisation

In this subsection, we follow Farr (2014) to derive a likelihood, which explicitly marginalises over time of coalescence t. Given a waveform with a reference coalescence time of t_0, we can calculate the waveform at some new coalescence time t by multiplying by the appropriate phasor:

$$\mu_j(t) = \mu_j(t_0)\exp\left(-2\pi i j\frac{(t-t_0)}{T}\right), \tag{C1}$$

where $T = 1/\Delta f$ is the duration of data segment and j is the index of the frequency bin as in Appendix B. It is understood that μ is a function of whatever parameters we are not explicitly marginalising over. We can therefore write κ^2 (see Eq. B10) as

$$\begin{aligned}\kappa^2(t) &\equiv \langle d, \mu(t)\rangle \\ &= 4\Delta f \Re \sum_j^M \frac{d_j^*\mu_j(t_0)}{P_j}\exp\left(-2\pi i j\frac{(t-t_0)}{T}\right).\end{aligned} \tag{C2}$$

However, this sum is the discrete Fourier transform. By recasting this equation in terms of the fast Fourier transform `fft`, it is possible to take advantage of a highly optimised tool.

We discretise $t - t_0 = k\Delta t$ where k takes on integer values between 0 and $M = T/\Delta t$. Having made this definition, marginalising over coalescence time becomes summing over k. The variable κ^2 is a function of (discretised) coalescence time k. We can write in terms of a fast Fourier transform:

$$\begin{aligned}\kappa^2(k) &= 4\Delta f \Re \sum_j^M \frac{d_j^*\mu_j(t_0)}{P_j}\exp\left(-2\pi i j\frac{k\Delta t}{M}\right) \\ &= 4\Delta f \Re\,\mathtt{fft}_k\left(\frac{d_j^*\mu_j(t_0)}{P_j}\right),\end{aligned} \tag{C3}$$

where $\mathtt{fft}_k$ refers to the k bin of a fast Fourier transform.

https://doi.org/10.1017/pasa.2019.2 Published online by Cambridge University Press

The other terms in Eq. (B7) are independent of the time at coalescence of the template. The marginalised likelihood is therefore

$$\begin{aligned}\log \mathcal{L}^{t}_{\rm marg} &= \log \int_{t_0}^{t_0+T} dt\, \mathcal{L}(\theta, t) \\ &= \log \mathcal{Z}_N - \frac{1}{2}\rho^2_{\rm opt}(\theta) + \log \int_{t_0}^{t_0+T} dt\, e^{\kappa^2(\theta,t)}\pi(t) \\ &= \log \mathcal{Z}_N - \frac{1}{2}\rho^2_{\rm opt}(\theta) + \log \sum_{k}^{M} e^{\kappa^2(\theta,k)}\pi_k,\end{aligned} \tag{C4}$$

where π_k is the prior on the discretised coalescence time.

Caution should be taken to avoid edge effects. If we employ a naive prior, the waveform will exhibit unphysical wrap-around. Similarly, care must be taken to ensure that the time-shifted waveform is consistent with time-domain data conditioning (e.g. windowing). (This is usually not a problem for confident detections because the coalescence time is well-known and so the segment edges can be avoided.) A good solution is to choose a suitable prior, which is uniform over some values of k, but with some values set to zero in order to prevent the signal from wrapping around the edge of the data segment. Note that Eq. (C1) breaks down for when the detector changes significantly over T due to the rotation of the Earth. It can also fail in the high signal-to-noise ratio limit when the t array becomes insufficiently fine-grained.

C.2. Phase marginalisation

In this subsection, we follow Veitch & Del Pozzo (2013) (see also Veitch et al. (2015)) to derive a likelihood, which explicitly marginalises over phase of coalescence ϕ_c. To begin, we assume a gravitational-waveform approximant consisting entirely of the dominant $\ell = 2$, $|m| = 2$ modes so that[t]

$$\mu = \mu_{22} + \mu_{2-2}. \tag{C6}$$

This is a valid assumption e.g. for the widely used waveform approximants—e.g. TaylorF2 (Damour, Iyer, & Sathyaprakash 2005), IMRPhenomD (Khan et al. 2016), and IMRPhenomP (Hannam et al. 2014)—but not for waveforms that employ higher order modes (e.g. Blackman et al. 2017). Given this approximation,[u]

$$\mu(\phi_c) = e^{2i\phi_c}\mu(\phi_c = 0). \tag{C8}$$

The optimal signal-to-noise ratio $\rho_{\rm opt}$ is invariant under rotations in ϕ_c. However, the matched filter signal-to-noise ratio is not. Thus, the phase-marginalised likelihood is

$$\begin{aligned}\mathcal{L}^{\phi_c}_{\rm marg} = \mathcal{Z}_N &- \exp\left(\frac{1}{2}\rho^2_{\rm opt}\right) \\ &+ \int_0^{2\pi} d\phi_c\, \exp\left(\frac{1}{2}\langle d, \mu(\phi_c)\rangle + \frac{1}{2}\langle \mu(\phi_c), d\rangle\right)\pi(\phi_c).\end{aligned} \tag{C9}$$

[t] The variables μ_{22} and μ_{2-2} are defined as follows:

$$\begin{aligned}\mu_{\ell m} \equiv F_+ \Re\Big(h_{\ell m}(\theta)\, {}_{-2}Y_{\ell m}(\iota, \phi)\Big) \\ + F_\times \Im\Big(h_{\ell m}(\theta)\, {}_{-2}Y_{\ell m}(\iota, \phi)\Big).\end{aligned} \tag{C5}$$

They depend on the metric perturbation $h_{\ell m}$ and the antenna response functions $F_{+,\times}$. The variable ${}_{-2}Y_{\ell m}(\iota, \phi)$ is a spin-weighted spherical harmonic function, evaluated the inclination angle ι and azimuthal angle ϕ of the observer. Without loss of generality, we can set $\phi = 0$, which establishes a coordinate frame. Having defined this frame, we may rotate the binary by the phase of coalescence ϕ_c in order to change the phase of the signal observed at Earth.

[u] We emphasise that the phase at coalescence is distinct from ϕ, the azimuthal angle to the observer in the source frame, which transforms differently:

$$\mu(\phi) = e^{2i\phi}\mu_{22}(\phi = 0) + e^{-2i\phi}\mu_{2-2}(\phi = 0). \tag{C7}$$

The variable ϕ_c calibrates the time evolution of the gravitational waveform observed at Earth, while ϕ describes how the waveform varies at a fixed time for observers at different spatial locations (corresponding to different azimuthal angles).

Using Eq. (C8), we can rewrite the phase-marginalised likelihood:

$$\begin{aligned}\mathcal{L}^{\phi_c}_{\rm marg} = \int_0^{2\pi} d\phi_c\, \exp\Big(\frac{1}{2}\langle d, \mu(\phi_c = 0)\rangle_{\mathbb{C}} \exp(2i\phi_c) + \\ \frac{1}{2}\langle \mu(\phi_c = 0), d\rangle_{\mathbb{C}} \exp(-2i\phi_c)\Big)\pi(\phi_c) \\ + \cdots\end{aligned}$$

The parts that do not depend on ϕ_c are implied by the ellipsis. Here we introduce the 'complex inner product' denoted with a subscript $\mathbb{C}$:

$$\langle a, b\rangle_{\mathbb{C}} \equiv 4\Delta f \sum_j \left(\frac{a_j^* b_j}{P_j}\right), \tag{C10}$$

which is identical to the regular inner product defined in Eq. (B5) except we do not take the real part in order to preserve phase information that will be useful later on. Employing a uniform prior on ϕ_c and grouping terms, the integral can be rewritten yet again:

$$\mathcal{L}^{\phi_c}_{\rm marg} = \int_0^{2\pi} \frac{d\phi_c}{2\pi}\, \exp\Big(A\cos(2\phi_c) + B\sin(2\phi_c)\Big) + \cdots, \tag{C11}$$

where

$$A \equiv \Re\langle d, \mu(\phi_c = 0)\rangle_{\mathbb{C}}, \tag{C12}$$

$$B \equiv \Im\langle d, \mu(\phi_c = 0)\rangle_{\mathbb{C}}. \tag{C13}$$

The integral yields modified Bessel function of the first kind:

$$I_0\left(\sqrt{A^2 + B^2}\right) = \frac{1}{2\pi}\int_0^{2\pi} d\phi\, e^{Ac_\phi + Bs_\phi}. \tag{C14}$$

Thus, we have

$$\begin{aligned}\sqrt{A^2 + B^2} &= \sqrt{\Re\langle d, \mu(0)\rangle^2_{\mathbb{C}} + \Im\langle d, \mu(\phi_c = 0)\rangle^2_{\mathbb{C}}} \\ &= \left|\langle d, \mu(\phi_c = 0)\rangle_{\mathbb{C}}\right| \\ &= \left|\kappa^2_{\mathbb{C}}\right|,\end{aligned} \tag{C15}$$

where $\kappa^2_{\mathbb{C}}$ is calculated the same way as κ (Eq. B10), except we use a complex inner product. The ϕ_c marginalised likelihood becomes

$$\log\mathcal{L}^{\phi}_{\rm marg} = \log \mathcal{Z}_N - \frac{1}{2}\rho^2_{\rm opt} + \log I_0(|\kappa^2_{\mathbb{C}}|). \tag{C16}$$

We reiterate that this marginalised likelihood is valid only insofar as we trust our initial assumption, that the signal is dominated by $l = 2$, $|m| = 2$ modes.

C.3. Distance marginalisation

In this subsection, we follow Singer & Price (2016) (see also Singer et al. 2016) to derive a likelihood, which explicitly marginalises over luminosity distance D_L. Given a waveform at some reference distance $\mu(D_0)$, the waveform at an arbitrary distance is obtained by multiplication of a scale factor:

$$\mu_j(D_L) = \mu_j(D_0)\left(\frac{D_0}{D_L}\right). \tag{C17}$$

As before, it is understood that μ is a function of whatever parameters are not explicitly marginalising over. Unlike time and phase, distance affects $\rho_{\rm opt}$ in addition to κ^2 (Eq. B10),

$$\begin{aligned}\kappa^2(D_L) &= \kappa^2(D_0)\left(\frac{D_0}{D_L}\right), \\ \rho^2_{\rm opt}(D_L) &= \rho^2_{\rm opt}(D_0)\left(\frac{D_0}{D_L}\right)^2.\end{aligned} \tag{C18}$$

Note that κ^2 and $\rho_{\rm opt}$ are implicit functions of whatever parameters we are not explicitly marginalising over.

At a fixed distance, the likelihood is

$$\log\mathcal{L}(D_L) = \log\mathcal{Z}_N + \kappa^2(D_L) - \frac{1}{2}\rho^2_{\rm opt}(D_L), \tag{C19}$$

and the likelihood marginalised over luminosity distance is

$$\log\mathcal{L}^{D}_{\rm marg} = \log\mathcal{Z}_N + \log\mathcal{L}_D, \tag{C20}$$

https://doi.org/10.1017/pasa.2019.2 Published online by Cambridge University Press

where

$$\mathcal{L}_D(\kappa^2, \rho_{\rm opt}) \equiv \int dD_L\, e^{\kappa^2(D_L) - \frac{1}{2}\rho^2_{\rm opt}(D_L)} \pi(D_L). \tag{C21}$$

This integral to calculate $\log \mathcal{L}_D$ can be evaluated numerically. This explicitly marginalised form is generally true for all gravitational-waves sources. Its validity is only limited by the resolution of the numerical integral, though, cosmological redshifts adds additional complications, which we discuss in the next subsection. One can construct a pre-computed lookup table $\log \mathcal{L}_D(\rho_{\rm mf}, \rho_{\rm opt})$ to facilitate fast and precise evaluation.

C.4. Distance marginalisation with cosmological effects

There is a caveat for our discussion of distance marginalisation in the previous subsection: when considering events at cosmological distances, the prior distributions for lab-frame masses become covariant with luminosity distance D_L due to cosmological redshift. A signal emitted with source-frame mass m_s is observed with lab-frame mass given by

$$m_l = (1+z) m_s. \tag{C22}$$

In this subsection, 'mass' m is shorthand for an array of both primary and secondary mass.

Now we derive an expression for $\mathcal{L}^D_{\rm marg}$, which can be applied to cosmological distances. We start by specifying the prior on redshift and source-frame mass[v]:

$$\pi(z, m_s) = \pi(z)\pi(m_s). \tag{C23}$$

Both $\pi(z)$ and $\pi(m_s)$ can be chosen using astrophysically motivated priors (see e.g. Talbot & Thrane 2018; Fishbach & Holz 2017; Fishbach, Holz, & Farr 2018). Whatever priors we choose for $\pi(z)$ and $\pi(m_s)$, they imply some prior for the lab-frame mass:

$$\begin{aligned}\pi(z, m_l) &= \pi\left(z, m_l/(1+z)\right) \left|\frac{dm_s}{dm_l}\right| \\ &= (1+z)^{-1} \pi\left(z, m_l/(1+z)\right).\end{aligned} \tag{C24}$$

Now that we have converted the source-frame prior into a lab-frame prior, we can write down the distance-marginalised (redshift-marginalised) likelihood in terms of lab-frame quantities:

$$\mathcal{L}^z_{\rm marge}(\kappa^2, \rho_{\rm opt}) = \int dz\, \mathcal{L}(\kappa^2, \rho_{\rm opt}, z) \pi(z|m_l), \tag{C25}$$

where

$$\mathcal{L}(\kappa^2, \rho_{\rm opt}, z) = \mathcal{Z}_N\, e^{\kappa^2(D_L(z)) - \frac{1}{2}\rho^2_{\rm opt}(D_L(z))}. \tag{C26}$$

Note that κ^2 and $\rho_{\rm opt}$ are implicit functions of whatever parameters we are not explicitly marginalising over.

By creating a grid of z, we can create a look-up table for $\mathcal{L}(\kappa^2, \rho_{\rm opt}, z)$, which allows for rapid evaluation of Eq. (C25). However, this means we will also need to create a look-up table for $\pi(z|m_l)$. In order to derive this look-up table, we rewrite the joint prior on redshift and lab-frame mass can be rewritten as follows:

$$\pi(z, m_l) = \pi(z|m_l)\pi(m_l). \tag{C27}$$

The marginalised lab-mass prior is

$$\pi(m_l) \equiv \int dz\, \pi(z, m_l), \tag{C28}$$

which can be calculated numerically. (We also need this distribution to provide to the sampler.) Thus, the conditional prior we need for our look-up table is as follows:

$$\pi(z|m_l) = \pi(z, m_l)/\pi(m_l). \tag{C29}$$

[v] Many previous analyses have assumed that this distribution is separable; however, this marginalisation technique does not require this.

With look-up tables for $\mathcal{L}(\kappa^2, \rho_{\rm opt}, z)$ and $\pi(z|m_l)$, the sampler can quickly evaluate $\mathcal{L}^z_{\rm marge}$ by summing over the grid of z:

$$\mathcal{L}^z_{\rm marg}(\kappa^2, \rho_{\rm opt}) = \Delta z \sum_k \mathcal{L}(\kappa^2, \rho_{\rm opt}, z_k)\pi(z_k|m_l), \tag{C30}$$

where Δz is the spacing of the redshift grid. This allows us to carry out explicit distance marginalisation while taking into account cosmological redshift.

C.5. Marginalisation with multiple parameters

One must take care with the order of operations when implementing these marginalisation schemes simultaneously. We describe how to combine the three marginalisation techniques described above. The correct procedure is to start with Eq. (C16) and then marginalise over distance:

$$\begin{aligned}\log \mathcal{L}^{\phi,D}_{\rm marg} &= \log \mathcal{Z}_N \\ &\quad + \log \int dD_L e^{I_0(|\kappa^2_{\mathbb{C}}(D_L)|) - \frac{1}{2}\rho^2_{\rm opt}(D_L)} \pi(dD_L).\end{aligned} \tag{C31}$$

Carrying out this integral numerically, one obtains a look-up table $\log \mathcal{L}^{\phi,D}_{\rm marge}(\kappa^2_{\mathbb{C}}, \rho_{\rm opt})$, which marginalises over ϕ and D_L. Finally, we add in t marginalisation by combining the look-up table with a fast Fourier transform:

$$\mathcal{L}^{\phi,D,t}_{\rm marg}(\kappa^2_{\mathbb{C}}, \rho_{\rm opt}) = \sum_k \pi_k\, \mathcal{L}^{\phi,D}_{\rm marg}\left(\kappa^2_{\mathbb{C}}(k), \rho_{\rm opt}(k)\right). \tag{C32}$$

C.6. Reconstructing the unmarginalised posterior

While explicitly marginalising over parameters improves convergence and reduces runtime, the sampler will generate no posterior samples for the marginalised parameters. Sometimes, we want posterior samples for these parameters. In this subsection, we explain how it is possible to generate them with an additional post-processing step.

The parameter we are most likely to be interested in reconstructing is the luminosity distance D_L. Let us assume for the moment that this is the only parameter over which we have explicitly marginalised. The first step to calculate the matched filter signal-to-noise ratio $\rho_{\rm mf}$ and optimal signal-to-noise ratio $\rho_{\rm opt}$ for each sample. For one posterior sample k, the likelihood for distance is

$$\mathcal{L}_k(d|D_L) = \mathcal{Z}_N\, e^{\kappa^2(\theta_k, D_L) - \frac{1}{2}\rho^2_{\rm opt}(\theta_k, D_L)}, \tag{C33}$$

where $\kappa^2(D_L)$ and $\rho_{\rm opt}(D_L)$ are defined in Eq. (C18). (When comparing with Eq. (C18), note that we have again made explicit the dependence on θ_k = whatever parameters we are not explicitly marginalising over.) Since this likelihood is one-dimensional, it is easy to calculate the posterior for sample k using Bayes' theorem:

$$p_k(D_L|d) = \frac{\mathcal{L}(d|D_L)\pi(D_L)}{\int dL\, \mathcal{L}(d|D_L)\pi(D_L)}. \tag{C34}$$

Using the posterior, one can construct a cumulative posterior distribution for sample k:

$$P_k(D_L|d) = \int dD_L\, p_k(D_L|d). \tag{C35}$$

The integral can be carried out numerically. The cumulative posterior distribution can be used to generate random values of D_L for each posterior sample:

$$D_L = P_k^{-1}(\texttt{rand}). \tag{C36}$$

Reconstructing the likelihood or posterior when multiple parameters have been explicitly marginalised over is more complicated. However, one may use the following iterative algorithm:

1. For each sample θ_k marginalise over all originally marginalised parameters except one (λ).
2. Draw a single λ sample from the marginalised likelihood times prior.
3. Add this λ sample to the θ_k and return to step 1, this time not marginalising over λ.

https://doi.org/10.1017/pasa.2019.2 Published online by Cambridge University Press

Alternatively, one can skip the step of generating new samples in distance and calculate the likelihood of the data given D_L marginalised over all other parameters,

$$\mathcal{L}(d|D_L) = \frac{1}{n}\sum_k^n \mathcal{L}_k(d|D_L)$$
$$= \frac{\mathcal{Z}_N}{n}\sum_k^n e^{\kappa^2(\theta_k,D_L)-\frac{1}{2}\rho^2_{\rm opt}(\theta_k,D_L)}. \quad \text{(C37)}$$

This likelihood can be used in Eq. (29) to perform population inference on the distribution of source distances and/or redshifts.

Appendix D. Posterior predictive distributions

The PPD represents the updated prior on the parameters θ given the data d. Recall that the hyper-posterior $p(\Lambda|d)$ describes our post-measurement knowledge of the hyper-parameters that describe the shape of the prior distribution $\pi(\theta)$. The PPD answers the question: given this hyper-posterior, what does the distribution of $\pi(\theta)$ look like? More precisely, it is the probability that the next event will have true parameter values θ given what we have learned about the population hyper-parameters Λ:

$$p_\Lambda(\theta|d) = \int d\Lambda\, p(\Lambda|d)\,\pi(\theta|\Lambda). \quad \text{(D1)}$$

The Λ subscript helps us distinguish the PPD from the posterior $p(\theta|d)$. The hyper-posterior sample version is

$$p_\Lambda(\theta|d) = \frac{1}{n_s}\sum_k^{n_s}\pi(\theta|\Lambda_k), \quad \text{(D2)}$$

where k runs over n_s hyper-posterior samples. While the PPD is the best guess for what the distribution $\pi(\theta)$ looks like, it does not communicate information about the variability possible in $\pi(\theta)$ given uncertainty in Λ. In order to convey this information, it can be useful to overplot many realisations of $\pi(\theta|\Lambda_k)$, where Λ_k is a randomly selected hyper-posterior sample. An example of a PPD is included in Figure 2.

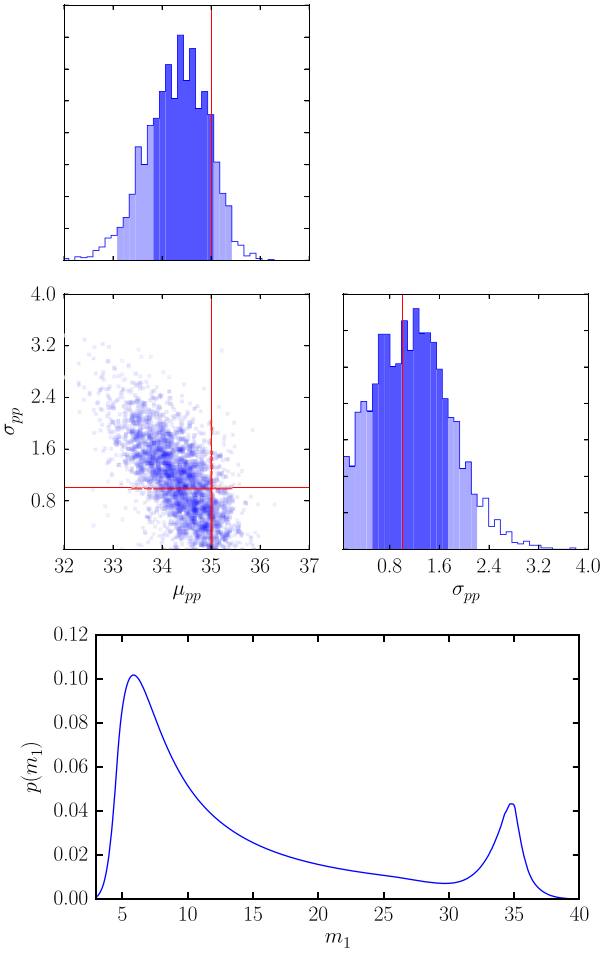

Figure 2: Top: an example corner plot from Talbot and Thrane (2018) showing posteriors for hyper-parameters $\mu_{\rm pp}$ and $\sigma_{\rm pp}$. These two hyper-parameters describe, respectively, the mean and width of a peak in the primary mass spectrum due to the presence of pulsational pair instability supernovae. Bottom: an example of a posterior predictive distribution (PPD) for primary black hole mass, calculated using the hyper-posterior distributions in the top panel (adapted from Talbot and Thrane (2018)). The PPD has a peak near $m_1 = 35$ because the hyper-posterior for $\mu_{\rm pp}$ is maximal near this value. The width of the PPD peak is consistent with the hyper-posterior for $\sigma_{\rm pp}$.

Appendix E. Selection Effects

In this section, we discuss how to carry out inference while taking into account selection effects, which arise from the fact that some events are easier to detect than others. We loosely follow the arguments from Abbott et al. (2016*a*); however, see also Mandel, Farr, & Gair (2018) and Fishbach et al. (2018). In Section E.1, we discuss selection effects in the context of an individual detection. In Section E.2, we generalise these results to populations of events.

E.1. Selection effects with a single event

Some gravitational-wave events are easier to detect than others. All else equal, it is easier to detect binaries if they are closer, higher mass (at least, up until the point that they start to go out of the observing band), and with face-on/off inclination angles. More subtle selection effects arise due to black hole spin (e.g. Ng et al. 2018). Typically, a gravitational-wave event is said to have been detected if it is observed with a matched-filter signal-to-noise ratio—maximised over extrinsic parameters $\theta_{\rm extrinsic}$—above some threshold $\rho_{\rm th}$:

$$\rho'_{\rm mf} \equiv \max_{\theta_{\rm extrinsic}}(\rho_{\rm mf}) > \rho_{\rm th}. \quad \text{(E1)}$$

Usually, $\rho_{\rm th} = 8$ for a single detector or $\rho_{\rm th} = 12$ for a ≥ 2 detector network.

Focusing on events with a $\rho_{\rm mf} > \rho_{\rm th}$ detection forces us to modify the likelihood function:

$$\mathcal{L}(\mathbf{d}|\theta, {\rm det}) = \begin{cases} \frac{1}{p_{\rm det}(\theta)}\,\mathcal{L}(\mathbf{d}|\theta) & \rho'_{\rm mf}(\theta) \geq \rho_{\rm th} \\ 0 & \rho'_{\rm mf}(\theta) < \rho_{\rm th} \end{cases}, \quad \text{(E2)}$$

where

$$p_{\rm det}(\theta) \equiv \int_{\rho'_{\rm mf}(\theta)>\rho_{\rm th}} d\mathbf{d}\,\mathcal{L}(\mathbf{d}|\theta). \quad \text{(E3)}$$

(Here, we temporarily switch to data $= \mathbf{d}$ to avoid confusing data with the differential d; we switch back to data $= d$ in a moment once we are finished with this normalisation constant.) This modification enforces the fact that we are not looking at data with $\rho'_{\rm mf} < \rho_{\rm th}$. The $p_{\rm det}$ factor ensures that the likelihood is properly normalised.

There are different ways to calculate $p_{\rm det}$ in practice. The probability density function for $\rho_{\rm mf}$ given θ—the distribution of $\rho_{\rm mf}$ arising from random noise fluctuations—is a normal distribution with mean $\rho_{\rm opt}$ and unit variance:

$$p(\rho'_{\rm mf}|\theta) = \frac{1}{2\pi}\exp\left(-\frac{1}{2}\left(\rho'_{\rm mf} - \rho_{\rm opt}(\theta)\right)^2\right), \quad \text{(E4)}$$

https://doi.org/10.1017/pasa.2019.2 Published online by Cambridge University Press

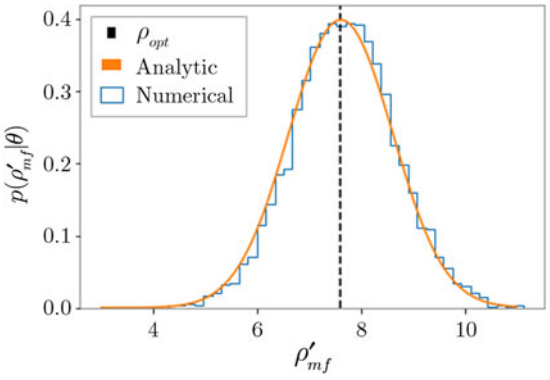

Figure 3: The distribution of matched filter signal-to-noise ratio maximised over phase for the same template in many noise realisations (blue). The distribution peaks at $\rho_{\rm opt} = 7.6$ (dashed black). The theoretical distribution (Eq. E4) is shown in orange.

see Figure 3. Thus, we have

$$p_{\rm det}(\theta) = \int_{\rho_{\rm th}}^{\infty} dx \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\Big(x - \rho_{\rm opt}(\theta)\Big)^2\right) \tag{E5}$$

$$= \frac{1}{2}\,{\rm erfc}\left(\frac{\rho_{\rm th} - \rho_{\rm opt}(\theta)}{\sqrt{2}}\right). \tag{E6}$$

Alternatively, if we are interested in the selection effects of intrinsic parameters, one may express $p_{\rm det}$ as the ratio of the 'visible volume' $\mathcal{V}(\theta)$ to the total spacetime volume $\mathcal{V}_{\rm tot}$:

$$p_{\rm det}(\theta) = \frac{\mathcal{V}(\theta)}{\mathcal{V}_{\rm tot}}. \tag{E7}$$

The visible volume is typically calculated numerically with injected signals.

E.2. Selection effects with a population of events

When considering a population of events, Eq. (E2) generalises to

$$\mathcal{L}(d, N|\Lambda, {\rm det}) = \begin{cases} \frac{1}{p_{\rm det}(\Lambda|N)}\mathcal{L}(d, N|\Lambda, R). & \rho_{\rm mf} \geq \rho_{\rm th} \\ 0 & \rho_{\rm mf} < \rho_{\rm th} \end{cases}. \tag{E8}$$

In analogy to Eq. (E7), the $p_{\rm det}$ normalisation factor can be calculated using the visible volume as a function of the hyper-parameters Λ:

$$\mathcal{V}(\Lambda) \equiv \int d\theta \mathcal{V}(\lambda)\pi(\theta|\Lambda). \tag{E9}$$

Naively, one might expect that

$$p_{\rm det}(\Lambda|N) = \left(\frac{\mathcal{V}(\Lambda)}{\mathcal{V}_{\rm tot}}\right)^N, \tag{E10}$$

but this expression is incorrect because it does not marginalise over the Poisson-distributed rate, which ends up changing the answer. Marginalising over the rate, we obtain

$$\begin{aligned} p_{\rm det}(\Lambda|N) &= \int dR \left(\frac{\mathcal{V}(\Lambda)}{\mathcal{V}_{\rm tot}}\right)^N \pi(N|R)\pi(R) \\ &= \int dR \left(\frac{\mathcal{V}(\Lambda)}{\mathcal{V}_{\rm tot}}\right)^N \left[e^{-R\mathcal{V}(\Lambda)}\frac{\mathcal{V}(\Lambda)^N R^N}{N!}\right]\pi(R) \\ &= \left(\frac{\mathcal{V}(\Lambda)}{\mathcal{V}_{\rm tot}}\right)^N \left[\int dR\, e^{-R\mathcal{V}(\Lambda)}\frac{\mathcal{V}(\Lambda)^N R^N}{N!}\right]\pi(R). \end{aligned} \tag{E11}$$

Note that $p_{\rm det}$ depends on our prior for the rate R. If we choose a uniform-in-log prior $\pi(R) \propto 1/R$, we obtain

$$p_{\rm det}(\Lambda|N) \propto \left(\frac{\mathcal{V}(\Lambda)}{\mathcal{V}_{\rm tot}}\right)^N, \tag{E12}$$

which reproduces the results from Abbott et al. (2018*a*). Note that

$$\mathcal{L}(d|\Lambda, {\rm det}) \neq \int d\theta\, \mathcal{L}(d|\theta, {\rm det})\, \pi(\theta|\Lambda). \tag{E13}$$

https://doi.org/10.1017/pasa.2019.2 Published online by Cambridge University Press