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

Eric Thrane[1,2,a] and Colm Talbot[1,2,b]

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

This note is an introduction to Bayesian inference with a focus on hierarchical models and hyper-parameters. We are writing for a audience of Bayesian novices, but we hope there will be useful insights for seasoned veterans as well. Examples are drawn from gravitational-wave astronomy, though we endeavor 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. Then, we generalize the formalism to discuss hyper-parameters and hierarchical models.

I. 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 we describe compact binary systems. An astrophysical black hole is completely characterized by just its mass and its dimensionless spin vector. From the standpoint of gravitational-wave emission neutron stars are barely more complicated; they have a tidal parameter related to the neutron star equation of state. Since sources of gravitational waves are so simple, and since we have a complete theory describing how they emit gravitational waves, there is a very 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 [1], to determine that GW170104 merged $880^{+450}_{-390}\,\mathrm{Mpc}$ away [2], and that the black holes in GW150914 had masses of $35^{+5}_{-3}\,\mathrm{M}_\odot$ and $33^{+3}_{-4}\,\mathrm{M}_\odot$ [3]. We use it to determine the Hubble constant [4], to study the formation mechanism of black hole binaries [5–11], and to probe how stars die [12, 13]. Increasingly, Bayesian inference and parameter estimation are the language of gravitational-wave astronomy. In this note, we endeavor to provide a primer on Bayesian inference with examples from gravitational-wave astronomy.

[a] eric.thrane@monash.edu
[b] colm.talbot@monash.edu

II. FUNDAMENTALS: LIKELIHOODS, PRIORS, AND POSTERIORS

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

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

Here, θ is the set of model parameters and d is the data associated with a measurement [14]. 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 the true value of θ is between $(\theta, \theta + d\theta)$ given the data d. It is normalized so that

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

The posterior distribution is what we use to construct confidence intervals that tell us, for example, the component masses of a binary black hole event like GW150914.

According to Bayes theorem, the posterior distribution is given by

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

Here, $\mathcal{L}(d|\theta)$ is the likelihood function of the data given the parameters θ, $\pi(\theta)$ is the prior distribution for θ [15] and $\mathcal{Z}$ is a normalization factor called the "evidence" [16]

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

The likelihood function is something that we choose. For gravitational-wave astronomy, we typically assume a Gaussian-noise likelihood function that looks something like this

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

Here, $\mu(\theta)$ is a template for the gravitational strain waveform given θ and σ is the detector noise. This likelihood function reflects our assumption that the noise in gravitational-wave detectors is Gaussian. Note that the likelihood function is not normalized with respect to θ and so

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

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

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 event, it is reasonable to choose an isotropic prior that weights each patch of sky as equally probable [17]. 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 black hole mass $\pi(m_1)$ [18]? When we are ignorant about θ, we often express our ignorance by choosing a distribution that is either uniform or log-uniform [19].

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 event is a fifteen-dimensional function that includes information about black hole masses, sky location, spins, etc. What if we want to look at just the posterior distribution for just the primary mass? To answer this question we *marginalize* (integrate) over the parameters that we are not interested (called "nuisance parameters") in order to obtain a marginalized posterior

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

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

The quantity $\mathcal{L}(d|\theta_i)$ is called the "marginalized likelihood"

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

When we marginalize 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 marginalizing over θ_a injects uncertainty into the posterior for θ_b. When this happens, the marginalized posterior $p(\theta_b|d)$ is significantly broader than the *conditional posterior* $p(\theta_b|\theta_a)$.

It is useful to consider a concrete example. There is a well-known covariance between the distance of a merging compact binary from earth d and the inclination of the orbital angular momentum vector with respect to the line of sight ι. For the binary neutron star coalescence GW170817, we are able to constrain the orbital inclination 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 [20].

III. MODELS, EVIDENCE AND ODDS

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

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

In practical terms, the evidence is a single number. It 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 marginalized 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 marginalized likelihood function.

Above, we described how the evidence serves as a normalization constant for the posterior $p(\theta|d)$. However, it 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 $\mathcal{Z}_N$

$$\mathcal{Z}_S \equiv \int d\theta \mathcal{L}(d|\theta)\, \pi(\theta) \tag{11}$$

$$\mathcal{Z}_N \equiv \mathcal{L}(d|0), \tag{12}$$

where

$$\mathcal{L}(d|0) \equiv \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{h^2}{\sigma^2}\right). \tag{13}$$

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

$$\mathrm{BF}^S_N \equiv \frac{\mathcal{Z}_S}{\mathcal{Z}_N}. \tag{14}$$

It is often convenient to work with the log of of Bayes factor

$$\log \mathrm{BF}^S_N \equiv \log(\mathcal{Z}_S) - \log(\mathcal{Z}_N). \tag{15}$$

When the absolute value of $\log \mathrm{BF}$ is large, we say that one model is preferred over the other. The sign of $\log \mathrm{BF}$ tells us which model is preferred. A threshold of $|\log \mathrm{BF}| = 8$ is often used as the level of "strong evidence" in favor of one hypothesis over another [21].

The signal/noise Bayes factor is just one example of a Bayes factor comparing two models. We may also compare two signal models. For example, we can compare the evidence for a binary black hole waveform predicted by general relativity (model M_1 with parameters θ) with a binary black hole waveform predicted by some other theory (model M_2 with parameters ν):

$$\mathcal{Z}_1 = \int d\theta \mathcal{L}(d|\theta, M_1)\, \pi(\theta) \tag{16}$$

$$\mathcal{Z}_2 = \int d\nu \mathcal{L}(d|\nu, M_2)\, \pi(\nu). \tag{17}$$

The 1/2 Bayes factor is

$$\mathrm{BF}^1_2 = \frac{\mathcal{Z}_1}{\mathcal{Z}_2}. \tag{18}$$

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

We can also 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 preferred spin. In this example, both hypotheses are signal hypotheses, but the two signal hypotheses are different. Indeed, we are free to compare any two hypotheses we wish.

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

$$\mathcal{O}^1_2 \equiv \frac{\mathcal{Z}_1}{\mathcal{Z}_2}\frac{\pi_1}{\pi_2}. \tag{19}$$

The odds ratio is the product of the Bayes factor with the prior odds π_1/π_2, which describes our prior belief about the relatively likelihood of hypotheses 1 and 2. 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 [22]

Bayesian model selection is a formal means of carrying out statistical inference, which corresponds nicely to the informal way that humans make inferences in daily life. To highlight this, we note that the Bayesian evidence encodes two pieces of information. First, the likelihood tells us how well our model fits the data. Second, the act of marginalization tell us about the size of the (hyper-) 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 penalizes 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 often difficult to completely rule out extensions to existing theories. However, we can use Bayesian inference to show that the data are better explained by a simpler theory.

IV. SAMPLERS

Thanks to the creation of 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

$$p(\theta|d) = \frac{\mathcal{L}(d|\theta)\, \pi(\theta)}{\mathcal{Z}}. \tag{20}$$

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 fifteen parameters describing a binary black hole merger. If we do this very naively, we might create a grid with ten 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 two or three dimensions.

The solution is to use a stochastic sampler. Commonly used sampling algorithms can be split into two broad categories of method: Markov-chain Monte Carlo (MCMC) [23, 24] and nested sampling [25]. These algorithms sample the posterior distribution, generating a list of posterior samples $\{\theta\}$ with the appropriate density to represent the posterior: $(\theta, \theta+d\theta) \propto p(\theta)$. Some samplers also produce an estimate of the evidence. We can visualize 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 [26]

$$\langle f(x)\rangle_{p(x)} = \int dx\, p(x)\, f(x) \approx \frac{1}{n_s}\sum_{k}^{n_s} f(x_k). \qquad (21)$$

Here $p(x)$ is the posterior distribution that we are sampling, $f(x)$ is some function we want to find the mean of, and n_s is the number of samples. Below, Eq. 21 will prove useful simplifying our calculation of the likelihood of data given hyper-parameters.

The second useful property is that, once we have samples from an N-dimensional space, we can generate the marginalized probability for any subset of the parameters by simply selecting the corresponding columns in our spreadsheet. This property is used to help visualize the output of these samplers by constructing "corner plots" which show the marginalized one- and two-dimensional posterior probability distributions for each of the parameters. A handy python package exists for making corner plots [27].

A. MCMC

Markov chain Monte Carlo sampling was first introduced by Metropolis *et al.* in 1953 [23] and extended by Hastings in 1970. 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 disclaimers that must be considered when using MCMC samplers. First, the early-time behavior of MCMC walkers os 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 accumulate samples from the posterior distribution. Determining when a particular chain has reached a steady state can be a challenge. 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, there is a degree of artistry applied using MCMC, developed from experience.

Additionally, the positions of a walker in a chain are autocorrelated. This correlation leads to the positions of the walker not representing a faithful sampling from the posterior distribution and leads to underestimating the width of the posterior distribution. It is thus necessary to "thin" the chain by the autocorrelation length of the chain.

Markov chain Monte Carlo 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 parallelize MCMC computations to quickly generate many samples.

Many variants of MCMC sampling have since been proposed in order to improve the performance of MCMC algorithms with respect to these and other issues. For a more in depth review of MCMC methods see [26]. The most widely used MCMC code in astronomy is EMCEE [28] [29].

B. Nested Sampling

The first widely used alternative to MCMC, was introduced by Skilling in 2004. While MCMC methods are designed to draw samples from the posterior distribution for the parameters, nested sampling is designed to calculate the evidence. Generating samples from the posterior distribution is simply a by-product of the nested sampling evidence calculation algorithm. By weighting each of the samples used to calculate the evidence by the likelihood, one can convert nested samples 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 $\times$ 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 usual 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 [30].

This method of sampling becomes very inefficient when the posterior support is only a small fraction of the whole space. It also becomes increasingly inefficient as the algorithm converges towards the maximum likelihood point. In order to improve the sampling efficiency, new points are often drawn from an ellipsoid containing the current live points. This method of sampling can still be very inefficient when the likelihood surface is multi-modal or parameters show extended covariance [31]. In order to address this issue multi-modal ellipsoidal nested sampling MULTINEST was introduced in [32], rather than

constructing a single bounding ellipsoid the points may be contained in multiple ellipsoids. MultiNest [33] is used widely in many areas of astronomy. Unlike MCMC algorithms nested sampling is not straightforwardly parallelizable, and posterior samples do not accumulate linearly with run time.

V. 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. Population properties are properties, which are common to all of the events in some set. Examples of population properties are 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.

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 some new "hyper-parameters"

$$\pi(\theta|\Lambda). \tag{22}$$

We call Λ a set of hyper-parameters. The hyper-parameters parameterize the shape of the prior distribution for the parameters θ. We can think of Λ as a knob that can be turned to change the shape of the prior for 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

$$\pi(m_1|\alpha) \propto m_1^{\alpha}. \tag{23}$$

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

$$\mathcal{L}(d|\Lambda) = \int d\theta\, \mathcal{L}(d|\theta)\, \pi(\theta|\Lambda). \tag{24}$$

Normally, we would call this completely marginalized 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

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

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

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

is they "hyper-evidence." We may refer to it as $\mathcal{Z}_\Lambda$ in order to distinguish it from the regular evidence $\mathcal{Z}_\theta$.

This section is about the study of populations of events. We now generalize 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 $\mathcal{L}_i$

$$\mathcal{L}_{\rm tot}(\vec{d}|\vec{\theta}) = \prod_i^N \mathcal{L}(d_i|\theta_i). \tag{27}$$

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-parameterized by Λ—the total marginalized likelihood is

$$\mathcal{L}_{\rm tot}(\vec{d}|\Lambda) = \prod_i^N \int d\theta_i\, \mathcal{L}(d_i|\theta_i)\, \pi(\theta_i|\Lambda). \tag{28}$$

The associated posterior is

$$p(\Lambda|\vec{d}) = \frac{\mathcal{L}_{\rm tot}(\vec{d}|\Lambda)\, \pi(\Lambda)}{\int d\Lambda\, \mathcal{L}_{\rm tot}(\vec{d}|\Lambda)\, \pi(\Lambda)}. \tag{29}$$

The denominator, of course, is they total hyper-evidence.

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

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

If we examine Eq. 30, 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 into account the M hyper-parameters Λ. As N gets large, it becomes difficult to sample such a large prior volume 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."

It turns out that the total marginalized likelihood in Eq. 28 can be written like so

$$\boxed{\mathcal{L}_{\rm tot}(\vec{d}|\Lambda) = \prod_i^N \frac{\mathcal{Z}_\emptyset(d_i)}{n_i} \sum_k^{n_i} \frac{\pi(\theta_i^k|\Lambda)}{\pi(\theta_i^k|\emptyset)}}. \tag{31}$$

Here, the sum over k is a sum over posterior samples with 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 (not favoring one value of θ over another). Using the $\varnothing$ prior, we obtain an evidence $\mathcal{Z}_\varnothing$. In this way, we are able to analyze 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

$$p(\theta_i|d_i,\varnothing) = \frac{\mathcal{L}(d_i|\theta_i)\,\pi(\theta_i|\varnothing)}{\mathcal{Z}_\varnothing} \tag{32}$$

Rearranging terms,

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

Plugging this into Eq. 28, we obtain [34]

$$\mathcal{L}_{\rm tot}(\vec{d}|\Lambda) = \prod_i^N \int d\theta_i\, p(\theta_i|d_i,\varnothing)\,\mathcal{Z}_\varnothing(d_i)\,\frac{\pi(\theta_i|\Lambda)}{\pi(\theta_i|\varnothing)}. \tag{34}$$

Finally we use Eq. 21 to convert the integral over θ_i to a sum over posterior samples, thereby arriving at Eq. 31.

VI. ACKNOWLEDGMENTS

This document is the companion note to a lecture at the 2018 OzGrav Inference Workshop held July 16-18, 2018 at Monash University in Clatyon, Australia. Thank you to the organizers: Greg Ashton, Paul Lasky, Hannah Middleton, and Rory Smith. This workshop was supported by Australian Research Council CE170100004. ET and CT are supported by CE170100004. ET is supported by FT150100281.

Appendix A: Gaussian noise likelihood in gravitational-wave astronomy

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,

$$d = \texttt{fft}\left(d(t)\right)/f_s, \tag{A1}$$

where f_s is the sampling frequency and `fft` is a Fast Fourier transform. The noise in each frequency bin is characterized 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

$$\mathcal{L}(d_j|\theta) = \frac{1}{\sqrt{2\pi P_j}}\exp\left(-2\Delta f\frac{|d_j-\mu_j(\theta)|^2}{P_j}\right). \tag{A2}$$

Here Δf is the frequency resolution. The factors of $2\Delta f$ are needed to convert the square of the Fourier transforms into units of one-sided power spectral density.

Since a typical gravitational-wave signal is spread over many frequency bins, each characterized by M independent Gaussian noise, the actual likelihood is the product of M Gaussian distributions

$$\mathcal{L}(\mathbf{d}|\theta) = \prod_j^M \mathcal{L}(d_j|\theta) \tag{A3}$$

Here $\mathbf{d}$ 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

$$\langle a,b\rangle \equiv 4\Delta f\sum_j \Re\left(\frac{a_j^* b_j}{P_j}\right), \tag{A4}$$

and the constant

$$\Psi \equiv -\frac{1}{2}\sum_j \log\left(2\pi P_j\right). \tag{A5}$$

Since constants do not change the shape of the log likelihood often "leave off" this normalizing 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 $\Psi = 0$.

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

$$\log \mathcal{L}(\mathbf{d}|\theta) = -\frac{1}{2}\left[\langle d,d\rangle - 2\langle d,\mu(\theta)\rangle + \langle \mu(\theta),\mu(\theta)\rangle\right] \tag{A6}$$

$$= -\frac{1}{2}\left[-2\log \mathcal{Z}_N - 2\rho^2_{\rm mf}(\theta) + \rho^2_{\rm opt}(\theta)\right]. \tag{A7}$$

$$= \log \mathcal{Z}_N + \rho^2_{\rm mf}(\theta) - \frac{1}{2}\rho^2_{\rm opt}(\theta) \tag{A8}$$

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

$$-2\log \mathcal{Z}_N \equiv \langle d,d\rangle. \tag{A9}$$

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 ≈ 1 [35]. The second is the matched filter signal-to-noise ratio squared

$$\rho^2_{\rm mf} \equiv \langle \mu, d\rangle. \tag{A10}$$

Last, we define the optimal matched filter signal-to-noise ratio squred

$$\rho^2_{\rm opt} \equiv \langle \mu, \mu\rangle. \tag{A11}$$

Readers familiar with gravitational-wave astronomy are likely familiar with the 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.

[1] B. P. Abbott *et al.*, Phys. Rev. Lett. **119**, 161101 (2017).
[2] B. P. Abbott *et al.*, Phys. Rev. Lett. **118**, 221102 (2017).
[3] B. P. Abbott *et al.*, Phys. Rev. Lett. **116**, 061102 (2016).
[4] B. P. Abbott *et al.*, Nature **551**, 85 (2017).
[5] S. Vitale, R. Lynch, R. Sturani, and P. Graff, Class. Quant. Grav. **34**, 03LT01 (2017).
[6] S. Stevenson, C. P. L. Berry, and I. Mandel, MNRAS **471**, 2801 (2017).
[7] C. Talbot and E. Thrane, Phys. Rev. D **96**, 023012 (2017).
[8] D. Gerosa and E. Berti, Phys. Rev. D **95**, 124046 (2017).
[9] W. M. Farr, S. Stevenson, M. C. Miller, I. Mandel, B. Farr, and A. Vecchio, Nature **548**, 426 (2017).
[10] D. Wysocki, J. Lange, and R. O'Shaughnessy, (2018), https://arxiv.org/abs/1805.06442.
[11] M. E. Lower, E. Thrane, P. D. Lasky, and R. Smith, (2018), https://arxiv.org/abs/1806.05350.
[12] M. Fishbach and D. E. Holz, Astrophys. J. Lett. **851**, L25 (2017).
[13] C. Talbot and E. Thrane, Astrophys. J. **856**, 173 (2018).
[14] 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. 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.
[15] In this document we use different symbols for different distributions: p for posteriors, $\mathcal{L}$ for likelihoods, and π for priors. We like this notation since it highlights what is what and makes formulas easy to read in our estimation. However, it is by no means standard, and some authors will use p for any and all probability distributions.
[16] For now, we will treat the evidence as "just" a normalization factor, though, below we will see that it plays an important role in model selection, and that it can be understood as a marginalized likelihood.
[17] If we include selection effects, it makes sense to use a non-isotropic prior that takes into account the anisotropic sensitivity of gravitational-wave detectors.
[18] The "primary" black hole is the heavier of two black holes in a binary, which is contrasted with the lighter "secondary" black hole.
[19] 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.
[20] The inclination angle is constrained to be $< 28^\circ with the electromagnetic counterpart, and (55^\circ)$ without it [1].
[21] H. Jeffreys, *Theory of Probability*, 3rd ed. (Oxford, Oxford, England, 1961).
[22] 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 [36].
[23] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, The Journal of Chemical Physics **21**, 1087 (1953).
[24] W. K. Hastings, Biometrika **57**, 97 (1970).
[25] J. Skilling, AIP Conf. Proc. **735**, 395 (2004).
[26] D. W. Hogg and D. Foreman-Mackey, The Astrophysical Journal Supplement Series **236**, 11.
[27] D. Foreman-Mackey, The Journal of Open Source Software **24** (2016), 10.21105/joss.00024.
[28] D. Foreman-Mackey, D. W. Hogg, D. Lang, and J. Goodman, Publications of the Astronomical Society of the Pacific **125**, 306 (2013).

[29] http://dfm.io/emcee/.
[30] In practice this is expressed as the difference between the calculated log evidence and the upper limit of the log evidence.
[31] In the context of compact binary coalescences, such covariance and multi-modality is seen between distance and binary inclination, and right ascension and declination.
[32] F. Feroz, M. P. Hobson, and M. Bridges, Mon. Not. R. Astron. Soc. **398**, 1601 (2009), arXiv:0809.3437.
[33] https://github.com/JohannesBuchner/MultiNest.
[34] By this stage, it is hopefully clear why this process is known as recycling: one "recycles" the posterior samples generated using the the $\pi(\theta_i|\emptyset)$ prior in order to do something new with the hyper-parameterized prior $\pi(\theta_i|\Lambda)$.
[35] 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) A3.
[36] T. Callister, A. S. Biscoveanu, N. Christensen, A. M. Maximiliano Isi, O. Minazzoli, T. Regimbau, M. Sakellariadou, J. Tasson, and E. Thrane, Phys. Rev. X **7**, 041058 (2017).