

◦ Bayesian Statistical Inference | Statistical Estimation for Physics ◦

MPhil and CDT in Data Intensive Science | *University of Cambridge* | James Alvey (j.b.g.alvey@uva.nl)

Topics: Interpreting Bayes' Theorem for Inference ◦ Key Quantities in Bayesian Inference ◦ Example Applications ◦ Research Review: GW150914

◦ **Lecture Summary** | The main goal of this lecture is to introduce the concept of **Bayesian Statistical Inference**. To begin, we will very briefly recap **Bayes' theorem** with a short example, before discussing its application to **parameter estimation** problems where we want to learn something about parameters θ given data x . After this, we will discuss the key quantities in Bayesian inference: the **posterior** $p(\theta|x)$, **prior** $p(\theta)$, **likelihood** $p(x|\theta)$, and the **evidence** $p(x)$. We will use this framework to consider a couple of simple examples of parameter estimation, with associated **code examples**. Finally, we will review an example of **parameter estimation in a research context** by looking at the first detection of gravitational waves by the LIGO collaboration (GW150914). This will motivate the material in future lectures where we discuss the need for **stochastic sampling** techniques, and other modern Bayesian inference methods.

🔗 **Code examples for this lecture:** [github@james-alvey-42/BayesianInference](https://github.com/james-alvey-42/BayesianInference)

◦ **Review: Bayes' Theorem** | This first section is **review material**, so can be skipped if you have already reviewed standard applications of Bayes' theorem. As you have seen previously, a key relation in Bayesian statistics is given by **Bayes' Theorem**. Fundamentally, this relates conditional distributions $p(A|B)$ and $p(B|A)$ for two events A and B via the application of the conservation of total probability, $p(A \cap B) = p(B \cap A)$. This leads to the equality,

$$p(A|B)p(B) = p(B|A)p(A) \quad (\text{Bayes' Theorem}) \quad (1)$$

◦ **(Example) Cycling in Amsterdam** | For anyone that has spent an appreciable amount of time in Amsterdam, they will know that it rains. A lot. More concretely, on average, there is rain about 220 days of the year. As such, a very regular question is "should I bring my jacket on the bike?". There is also a weather service known as Buienradar, which is known to be one of the best rain forecasters in the country. The question we want to answer is: if Buienradar says it is going to rain, should we take our jacket? In a Bayesian context, we are trying to compute the probability $p(\text{Rains} | \text{Buienradar forecasted rain})$, which can be calculated using Bayes' theorem via,

$$p(\text{Rains} | \text{Buienradar forecasted rain}) = \frac{p(\text{Buienradar forecasted rain} | \text{Rains})p(\text{Rains})}{p(\text{Buienradar forecasted rain})} \quad (2)$$

There are two more pieces of information we need that concern how good a forecaster Buienradar typically is. Let's assume that if it actually rains, Buienradar forecasted rain 75% of the time, and on days where it doesn't rain, Buienradar forecasts rain anyway 10% of the time. In the classic disease-test context, this would be the true and false positive rates. With this we can then compute,

$$p(\text{Rains} | \text{Buienradar forecasted rain}) = \frac{75\% \times (220 \text{ days}/365 \text{ days})}{75\% \times (220 \text{ days}/365 \text{ days}) + 10\% \times (145 \text{ days}/365 \text{ days})} \simeq 92\% \quad (3)$$

In other words, the Dutch expression that "we zijn niet van suiker gemaakt"/"we aren't made of sugar", is fortunately true, and you should probably bring a coat. Note that in this case, the high probability is really driven by the relatively high occurrence of rain, rather than the particular performance of the forecasts. This example was meant to just review briefly the application of Bayes' theorem to general conditional estimation problems. From now on, we will focus exclusively on its application to **parameter estimation** and **statistical inference**.

◦ **Bayes' Theorem and Statistical Inference** | The main focus of this lecture is how to apply Bayes' theorem to problems in **statistical inference**. Of course, this is a very general task across physics, and broadly is a framework for asking the question: **what can I learn about my physics model given some observed data?** To set this up more precisely, suppose we have some model that takes parameters θ and stochastically generates simulated data x . **Parameter estimation** is then asking the question: given some observed data x , what can we infer about the underlying model parameters θ ? Making the connection to Bayes' theorem, this means that we would like to estimate the quantity $p(\theta|x)$ (the **posterior** distribution of model parameters given data). This is the key quantity of interest in parameter inference problems. If we directly apply Bayes' theorem to this quantity, we can rewrite it in the following way:

$$(\text{Posterior}) \ p(\theta|x) = \frac{(\text{Likelihood}) \ p(x|\theta) \times (\text{Prior}) \ p(\theta)}{(\text{Evidence}) \ p(x)}. \quad (4)$$

The basic premise of the Bayesian method is then that probability statements are not limited to data, but can also be applied to model parameters (and indeed models themselves). In the next section, we will go into the details of each component of this expression, including e.g the likelihood, prior, and posterior. This formulation is really the cornerstone of all Bayesian statistical inference, which has found wide application across physics. Specific examples of its application in astrophysics and cosmology include the analysis of the cosmic microwave background [1], and the estimation of source parameters (such as masses and spins) from gravitational waves [2]. We will return to this latter example below.

Bayesian vs. Frequentist Statistics (bonus material) | It is worth briefly pointing out a slightly more philosophical viewpoint on this application of Bayes' theorem. Specifically, we see explicitly that in order to do Bayesian inference, we have to allow for probability distributions over model parameters θ . On a practical level, the split between frequentist and Bayesian methods lies in whether it even makes sense to assign probabilities to parameters which are thought to be fixed but unknown – such as particle masses, cosmological parameters etc. From a frequentist perspective, no probabilities may be assigned since the value is fixed and there is no known random process associated to it. In contrast, based on a different conception of probability, “Bayesians” do allow for a probability, also called credibility in this context, to be assigned to these parameters. These constraints lead to different methods, typically centred around statistical tests in the frequentist case and Bayes' theorem in the Bayesian case. Frequentist methods then typically look at the probability of data given fixed parameters, while in pure Bayesian methods, the data is fixed and a range of possible parameters is considered. In what follows, we will take a practical approach to these conceptual issues, but see e.g. Refs. [3–5] for further reading.

◦ **Priors, Likelihoods, and Posteriors** | Before turning to a practical example of applying Bayesian inference, we will discuss each component of Eq. (4).

- **Prior** | The **prior** $p(\theta)$ is a fundamentally Bayesian concept in the sense that it assigns an initial probability, or credibility, to possible values of the model parameters θ . To carry out Bayesian inference, it is necessary to choose a prior distribution over θ , which is then updated via the observation of data x to inform some posterior belief (specified by $p(\theta|x)$). Perhaps from this discussion, it is already clear that there is some inherent user input required to carry out inference regarding choices of prior. This often creates significant debate in the literature about what is the “right” prior. Here, we will try and take a more practical approach, but it is relevant to think about the different types of choices one could make. Broadly these include attempting to choose: a physically motivated prior (e.g. that the mass of a particle should be positive), a “maximally uninformative prior” (although this can introduce some subtle issues regarding parameterisation)^a, or a parameterisation-independent prior (the Jeffreys prior [6] is a standard example of this). There is a general statement that we can make however, which is that the key science results should not depend appreciably on the prior choice, unless the prior has a clear physical motivation. We will give some concrete examples of standard prior choices below.


^ae.g. saying that an spaceship is equally likely to land anywhere on the Earth is not the same as saying that all latitudes of arrival are equally likely.

- **Likelihood** | The **likelihood** $p(x|\theta)$ is essentially the statistical definition of the forward model. It quantifies the probability of data x given model parameters θ . Depending on the particular scenario at hand, the likelihood could in principle range from an extremely complicated, hierarchical model, to a deterministic function of the model parameters. In the context of Bayesian statistics, the likelihood plays the role of updating our current credibility about model parameters given observed data x . Again, we will see some concrete examples of likelihood distributions below.
- **Posterior** | The **posterior** $p(\theta|x)$ is typically the main scientific result (although another important component is the evidence $p(x)$ if model comparison is a relevant aspect) of a Bayesian analysis. It quantifies the updated credibility of a given set of model parameters in light of observed data x . Specifically, this is obtained using Eq. (4). Note that in Eq. (4), the denominator does not depend on the model parameters θ . As such, for fixed data x , an *unnormalised* posterior distribution $p(\theta|x)$ over parameters can be obtained without evaluating $p(x)$. We will discuss below the relevant challenges, limitations, and methods required to obtain well-controlled posterior distributions, although a detailed exploration will be given in later lectures.
- **Evidence** | Although we will not focus on the **evidence** $p(x)$ in the examples below for the reason mentioned above, it nonetheless plays a key role in Bayesian model comparison, validation, and selection. Again, this will be presented in detail in later lectures, but generally it computes the overall probability of some given data x under some model (including the prior). Classically, this can be a complex quantity to compute since it is the result of a potentially high-dimensional integration over model parameters:

$$p(x) = \int d^n\theta p(x|\theta)p(\theta). \quad (5)$$

A number of techniques, the most well-known of which being **nested sampling**, have been developed to compute the evidence efficiently [7], although more modern approaches have started to emerge also [8].

◦ **Practical Examples of Bayesian Parameter Estimation** | We will now take a look at some practical examples of using Bayes' theorem in Eq. (4) to do parameter estimation. Specifically, we will begin with a simple example of flipping coins, before moving to more complex scenarios involving signal analysis. Finally, we will look at a specific **research-based** example of **gravitational wave parameter inference** in the context of GW150914.

▢ **Example #1: Flipping Coins** |  `flipping_coins.ipynb` | In this first example, we will illustrate a number of simple concepts regarding Bayesian parameter estimation. In particular, we will look at some concrete definitions of priors and likelihood, how to compute posteriors, and the behaviour of likelihoods/posteriors as we vary the data x .

The setup is simple: imagine that we have a coin that we flip n_{flips} times, noting down whether each of the individual flips is a head or a tail. The coin has some unknown probability p_{heads} of turning up heads, which we want to do Bayesian inference on. In the notation above, we assign $\theta = p_{\text{heads}}$ as our model parameter. In this case the data x is a sequence of heads and tails $x = \text{"HTHTT..."}$ (in practice, we can encode this as a sequence of 1s and 0s).

The first step is to define a **prior** on θ . Since we in principle know nothing about the coin, there are at least a few sensible choices including taking a uniform prior over all possible values $\theta \in [0, 1]$. As we mentioned above, this choice is not unique and we can (and should) check that our qualitative and quantitative analysis results are not sensitive to this. In this example we will just demonstrate the procedure with the uniform prior choice such that:

$$\theta \sim U(0, 1) \quad \text{i.e. } p(\theta) = \Theta(\theta) \times \Theta(1 - \theta) \quad (6)$$

where Θ is the heaviside step function. This is implemented in the `Prior` class in the corresponding notebook. Given the setup above, we can also write down the **likelihood** of data $x = \text{"HT..."}$ given a value of θ by considering

a sequence of independent flips. This gives:

$$p(x|\theta) = \theta^{n_{\text{heads}}} \times (1 - \theta)^{n_{\text{flips}} - n_{\text{heads}}}, \quad (7)$$

where n_{heads} is the number of heads in the given data sequence x . Similar to the case of the prior, this is implemented in the `Likelihood` class. Finally, the (unnormalised) posterior can be computed via $p(\theta|x) \propto p(x|\theta)p(\theta)$, which is implemented in the `Posterior` class.

⚙️ Programming Note: Note that we are using `jax` to code up the distributions here, this allows for `jit`-compilation of e.g. the density functions for fast evaluation.

⚙️ Programming Note: There is a subtlety on the coding side here as we increase n_{flips} in that the likelihood $p(x|\theta)$ will tend to become extremely small. At a certain point, this will exceed machine precision and we will need to approach it differently. See the exercises in the notebook to explore how this can be solved.

With these definitions, we can now explore a simple case study. Let us imagine that the coin is indeed an unbiased one, so that $\theta_{\text{true}} = 0.5$. What sort of behaviour will we observe when we do statistical inference in this setup? The first test we can do is generate lots of different realisations of the data from the true underlying model with $\theta = 0.5$. For each realisation, we can ask what our inference result would be if that were the real observation. To get a feeling for this, we can evaluate the likelihood $p(x_i|\theta)$ for each realisation x_i , as a function of θ . This is shown in the left panel of Fig. 1 for the simple case of $n_{\text{flips}} = 10$.

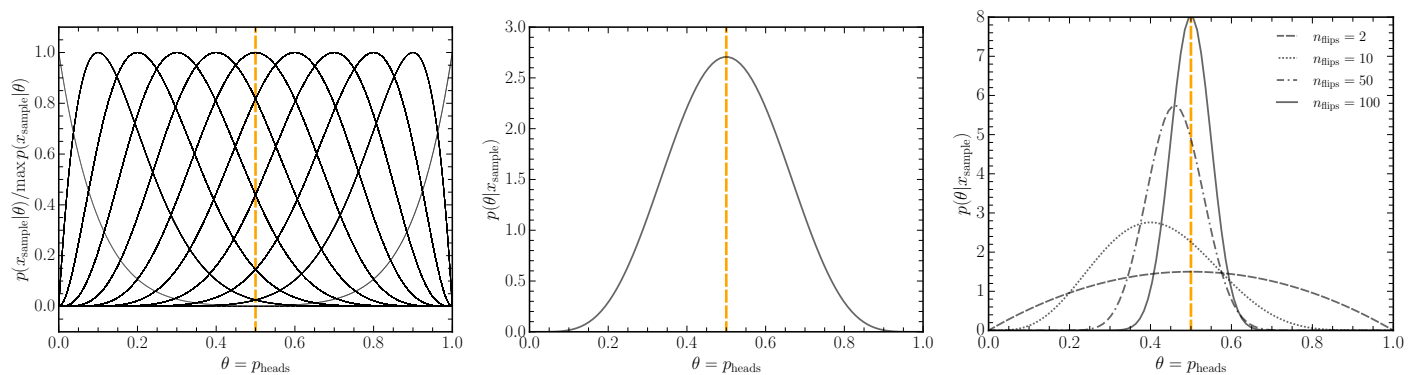



Figure 1: **Left:** Likelihood profiles as a function of θ for various data realisations generated with $\theta = 0.5$. **Centre:** Normalised posterior $p(\theta|x)$ for a particular realisation (bonus: just from looking at the posterior, what can you say about x ?). **Right:** Behaviour of the posterior estimations as we vary the number of flips n_{flips} .

? Exercise: Can you explain the discreteness in the different classes of curve? How many should there be? How does this depend on n_{flips} ? What determines the intensity of the various lines?

The final step in this example is to do some inference. Given the setup above, all we need to do is take our likelihood evaluations. The result for one data realisation with $\theta = 0.5$ is shown in the right panel of Fig. ??, but you can re-run the cell in the notebook multiple times to see the behaviour as we vary the data. We can also ask other simple questions such as: how does the precision in our determination depend on n_{flips} ? This result is also shown below, and we see that the natural intuition holds solid — more flips, better precision. We will stop this example here, but there are some additional exercises to think about in the context of the notebook.

? Exercise: Try plugging in $n_{\text{flips}} = 1000$, what happens? How can you evaluate the posterior in a more robust way (hint: consider computing $\log p(x|\theta)$)?

? Exercise: How could you check that the posterior is behaving correctly as you vary the data? (Hint: try checking how often the true value is contained in e.g. the 68% confidence interval of the posterior, how often should this be the case? This is the notion of a **coverage test**.)

▣ **Example #2: Signal Parameter Estimation** |  `signal_estimation.ipynb` | The second practical example that we will cover is in principle no more involved than the first: we have some model parameters (two this time), we choose a very un-informed prior over both, we write down the likelihood (which is particularly simple in this case), and compute the posterior. The goal of this example, however, is to act as a prelude to the **research review** that we present below. In particular, this second example has all the hallmarks of what is done in real **gravitational wave parameter inference**, albeit with a few minor simplifications.

The general idea of gravitational wave parameter inference is to look for a fixed signal template $s(\theta)$, which depends on some source parameters θ , in the presence of some additive noise n (with some distribution), so that the total data is $x = s + n$. In this simple warm up, we are going to consider the following signal and noise models. We take the signal to be a unit-variance Gaussian bump centred at some “frequency” f_0 with some amplitude A_s (we are being a bit loose with terminology and units here just to make the connection to gravitational waves). In other words, we have the template:

$$s(f|\theta = (f_0, A_s)) = \frac{A_s}{\sqrt{2\pi}} \exp\left(-\frac{(f - f_0)^2}{2}\right). \quad (8)$$

For the “detector” noise, we assume that the noise at any “frequency” is a zero-mean, unit-variance Gaussian distribution such that,

$$p(n(f)) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{n(f)^2}{2}\right) \quad (9)$$

With these definitions, we can start to get an idea of the inference problem by generating some example data. Assuming that my “detector” consists of measurements at 100 “frequencies” f_i ($i = 1, \dots, 100$) linearly spaced between $f = 1$ and $f = 20$, one example of a data realisation with $f_0 = 5$ and $A_s = 10$ is shown in the left panel of Fig. 2.

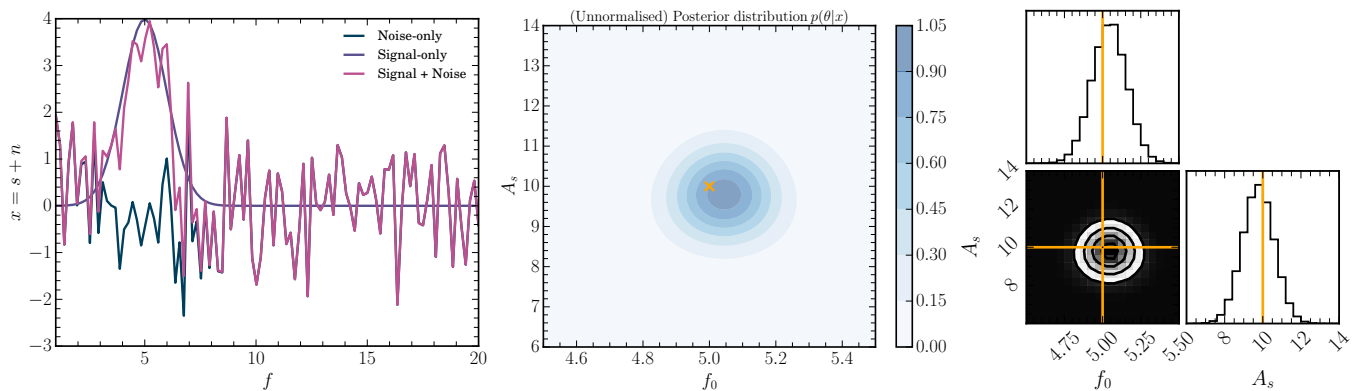


Figure 2: **Left:** Example data realisation with $f_0 = 5$ and $A_s = 10$. **Centre:** (Un-)normalised posterior $p(\theta|x)$ for a particular realisation with the true value marked with the yellow cross. **Right:** Corner plot produced using the `corner` package, with the true values indicated by the yellow lines.

We can now set up the inference problem in exactly the same way as before. Specifically, we take some uniform priors over the model parameters $f_0 \sim U[0, 21]$ and $A_s \sim U[0, 20]$. In this case we will find that we are totally **data-dominated**, i.e. that the likelihood totally swamps the prior. This makes the inference results far less sensitive to the prior choice. The **likelihood** is given by,

$$p(x = s + n|\theta) = \prod_i \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x(f_i) - s(f_i|\theta))^2}{2}\right), \quad (10)$$

where we have used the fact that for a fixed template $s(\theta)$, the data $x(f_i)$ in each frequency bin behaves like a unit-variance Gaussian random variable with mean $s(f_i|\theta)$.

? Exercise: Try and directly compute this likelihood for some fixed value of θ , what happens? How does it depend on the number of frequency bins?

If you carried out the exercise, you will find that, just as in the flipping coins example, at some point the numerics of computing this density directly blow-up. This is a very generic feature and why you will almost always see people working with **log-likelihoods** (see e.g. the maximum likelihood estimates from CMB analyses [1]) $\log p(x|\theta)$. In the present context we can easily compute this:

$$\log p(x = s + n|\theta) = -\frac{1}{2} \sum_i (x(f_i) - s(f_i|\theta))^2 + \text{const.} \quad (11)$$

where we have left out the constant factor just because this can easily be absorbed into re-normalising the posterior. This is implemented in the `Likelihood` class of the notebook. Finally, we can compute the (log-)posterior by simply summing the (log-)likelihood and (log)-prior: $\log p(\theta|x) = \log p(x|\theta) + \log p(\theta) + \text{const.}$, where again, the additional constant just translates into an additional normalisation factor in the posterior. This is implemented in the `Posterior` class.

In some sense, the rest of the example proceeds identically to the previous case, although the fact that there is now more than one model parameter allows us to highlight a standard but important construction: the **corner plot**. In the notebook, this is generated using the simple python package `corner` by weighting prior samples by the corresponding posterior density. Corner plots are a very typical, and important, output of a scientific analysis, and are a way to represent the low-dimensional behaviour of posterior distributions. In particular, they illustrate the behaviour of the 1- and 2-dimensional **marginal** posterior distributions. In this example, this means that the corner plot in Fig. 2 shows:

$$p(\theta = (f_0, A_s)|x), \quad p(f_0|x) = \int dA_s p(f_0, A_s|x), \quad p(A_s|x) = \int df_0 p(f_0, A_s|x). \quad (12)$$

From these lower dimensional marginal distributions, it is then possible to directly compute Bayesian parameter constraints/measurements, such as those for cosmological parameters, gravitational wave sources, or astrophysical data.

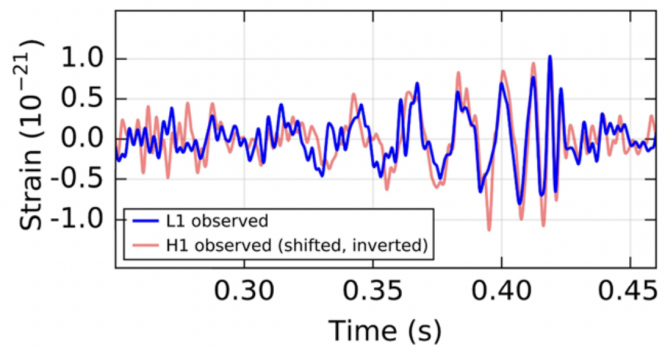
? Exercise: Try and compute the 1- and 2-sigma contours directly from the posterior by integrating. Try and compute the marginalised posteriors also (Hint: The marginal posterior for e.g. f_0 can be obtained by integrating the joint density: $p(f_0|x) = \int dA_s p(f_0, A_s|x)$).

◦ Research Review: GW150914 Parameter Inference

| On the 14th September 2015, the two LIGO detectors at Hanford and Livingston detected the first gravitational wave [9] (see inset, from Ref. [10]). There is of course, a huge amount to say regarding the significance of this observation, and the implications for our understanding of general relativity, compact objects, and astrophysics. In this short section, we will take on a smaller task and just try to understand at a basic level **how did LIGO detect a gravitational wave?** To do so, we will make some

concrete connections with the last example of finding a parameterised signal $s(\theta)$ in some noisy data $x = s + n$.

The current LIGO data analysis strategy splits into two parts: (a) first you have to find a possible signal in the data, this is done by basically comparing lots of different $s(\theta)$ to the data x and seeing which ones match



best (think of when you are at the opticians and they ask “better...or worse” with different lenses). This process (called **matched filtering**) is quite effective, and typically indicates broadly the interesting region of the data stream to analyse, but does not give precise inference. So, the next step, which we will focus on here is to (b) carry out a detailed Bayesian inference follow up to obtain precise measurements/constraints on the parameters of the gravitational wave source.

At this point, we will turn to the actual papers to make some connections with the literature. There are a number of levels to appreciate these results: the first is at the level of the basic physics of black holes and merging compact objects. This is summarised in a beautiful, and very accessible paper, see Ref. [10], which discusses the amount we can learn from the signal using basic Newtonian and GR theory in terms of the nature of the compact objects, their orbital dynamics, and the energy emitted in gravitational waves. In addition to this, the LIGO collaboration released a paper with the official announcement and characterisation of the event (Ref. [9]). This is almost certainly the paper that will appear if you google “GW150914 paper”. The paper we are going to focus on though is slightly less well-cited, but presents a more detailed picture of the Bayesian parameter estimation problem: “Improved analysis of GW150914 using a fully spin-precessing waveform model” (Ref. [11], published in *Phys. Rev. X* in 2016). For an excellent review of these paper releases, see [this blog post by Prof. Christopher Berry \(Uni. of Glasgow\)](#).

Our main interest in Ref. [11] is presented at the start of Section III (entitled *Bayesian Inference Analysis*). In particular, Eq. (2) starts by presenting Bayes’ theorem in this context (here, we lift the equations directly from the paper to match the notation):

$$p(\boldsymbol{\vartheta}|\text{model, data}) \propto \mathcal{L}(\text{data}|\boldsymbol{\vartheta}) \times p(\boldsymbol{\vartheta}). \quad (13)$$

So aside from writing the likelihood as $\mathcal{L}(\text{data}|\boldsymbol{\vartheta})$ and omitting the evidence $p(\text{data})$ as a normalisation constant, this is just Bayes’ theorem for parameter estimation. The things to ask here are: (i) what is the model? and (ii) what are the parameters $\boldsymbol{\vartheta}$? Again, we could give a whole lecture (maybe course) on gravitational waveform models, but to get the point across here, it suffices to say that the model here is a “precessing binary black hole merger waveform”. In the notation we set up above, this just means there is a model for the form of the signal template $s(\theta)$ that is tuned to described the phenomena of precessing, binary black hole mergers. The parameters $\boldsymbol{\vartheta}$ then characterise properties of the source such as the masses of the two black holes ($m_{1,2}$), their spins ($a_{1,2}$), the distance to the source (D_L), the time of arrival in the detector, and the precession of the orbit (χ_p). This model is summarised in Eq. (3) of the paper where it deconstructs the signal as measured in the detector $h_k(\boldsymbol{\vartheta})$ into a sum of + and \times -polarisation modes $h_{+/\times}(\boldsymbol{\vartheta}_{\text{intrinsic}})$ which depend on the parts of $\boldsymbol{\vartheta}$ that are intrinsic to the source (such as the masses and spins). These polarisations are projected differently onto the detector, so $h_{+/\times}(\boldsymbol{\vartheta}_{\text{intrinsic}})$ is modulated by the projection factors $F_k^{+/\times}(\boldsymbol{\vartheta}_{\text{extrinsic}})$ for each detector $k = \text{Hanford, Livingston}$. The projection factors only depend on the extrinsic parameters of the source (things like the distance to the binary, time of arrival in the detector etc.).

With the model defined, we just need to understand the prior $p(\boldsymbol{\vartheta})$ and the likelihood $\mathcal{L}(\text{data}|\boldsymbol{\vartheta})$. The prior is described just below Eq. (4) (starting “The prior probability density...”), and refers to Section II.B of Ref. [12] (if you look at this paper, you will also see a lot of similarities in terms of the description of the Bayesian inference approach). This is a nice practical illustration of some of the concrete choices that need to be made when designing priors for Bayesian inference – one test to understand the impact of their choice is described in the very next paragraph (“To assess whether the data is *informative* with regard to a source parameter...”). As far as Bayesian inference is concerned, this just leaves the likelihood $\mathcal{L}(\text{data}|\boldsymbol{\vartheta})$, defined in Eq. (4):


$$\mathcal{L}(\text{data}|\boldsymbol{\vartheta}) \propto \exp \left[-\frac{1}{2} \sum_{k=1,2} \langle h_k(\boldsymbol{\vartheta}) - d_k | h_k(\boldsymbol{\vartheta}) - d_k \rangle \right]. \quad (14)$$


Here, d_k is the actual data measured in detector k , and $\langle \cdot | \cdot \rangle$ is the “noise-weighted inner product” (see Eq. (6)

in Ref. [13]),

$$\langle h|h \rangle = 4 \int_0^\infty df \frac{h^*(f)h(f)}{S_n(f)}, \quad (15)$$

where $S_n(f)$ is the **power spectral density** (PSD) of the detector (more shortly). I hope you agree that this looks **remarkably similar** to our example above (see Eq. (10))! There are really just three differences. Firstly, the data d_k (and template h_k) are computed in Fourier space, this makes them complex numbers, and requires the inner product definition rather than the simple squared term. Secondly, they are viewed as continuous functions of frequency f , such that the sum we wrote above turns into an integral over frequency f , although this is a bit semantic since the integral is always approximated numerically. Finally, we need to relax one assumption we made about the noise in every frequency bin being identically distributed. In a real gravitational wave detector, the amplitude of the noise varies as a function of frequency, and is specified by the PSD $S_n(f)$. For detectors such as the two LIGO facilities, $S_n(f)$ becomes large at low frequencies (below about 20 Hz) and high frequencies (above $O(1000)$ Hz), leaving a sensitivity “band” in the middle where signals can be analysed.

With these (arguably minor) complications, there really is nothing more to add to the statistical framework for doing Bayesian inference for gravitational waves. If you try and follow exactly our procedure above though, you will hit a snag. Specifically, we had been a little bit naive, and somewhat blindly computed the posterior on a grid of parameter values (taken from e.g. the prior). If you try this in this example however, you will find that the number of times you need to evaluate the likelihood/prior in order to get smooth posterior estimates scales **exponentially** with the number of parameters. For example, if it takes e.g. 100 samples to get a smooth posterior distribution for one parameter, then it will take approximately 100^N samples to get a smooth distribution for N copies of this parameter. Very quickly, this becomes totally impossible computationally. So, we have to do something smarter – this is motivation for **sampling** techniques such as **Monte Carlo Markov Chains (MCMC)** and **nested sampling**, which we will focus on in future lectures. A practical example to highlight this is given in the  `high_dimensional.ipynb` notebook, along with some short exercises.

◦ **Lecture Summary** | To summarise, the **key learning objectives** from this lecture (and the  material) are:

- ❑ **Review:** Simple application of Bayes’ theorem for conditional probabilities.
- ❑ Understanding of how to apply Bayes’ theorem to problems in statistical inference.
- ❑ An appreciation for the interpretation of probability regarding models and model parameters in Bayesian statistical inference. Related discussion regarding choices of prior.
- ❑ Definitions of the prior, likelihood, and posterior in the context of Bayesian inference.
- ❑ Application of Bayesian inference to simple examples including derivation of the likelihood and choices of prior.
- ❑ Appreciation of the computational complexities involved with Bayesian inference including: computation of the likelihood/posterior, normalisation of the posterior, corner plots, and scaling with the dimensionality of the problem
- ❑ **Research Review:** Understanding of the definitions of the model, likelihood, and prior for the Bayesian inference analysis of gravitational waves. GW150914 as a concrete example.

Further Reading

- [1] PLANCK collaboration, *Planck 2018 results. VI. Cosmological parameters*, *Astron. Astrophys.* **641** (2020) A6 [1807.06209].
- [2] E. Thrane and C. Talbot, *An introduction to Bayesian inference in gravitational-wave astronomy: parameter estimation, model selection, and hierarchical models*, *Publ. Astron. Soc. Austral.* **36** (2019) e010 [1809.02293].
- [3] R.D. Cousins, *Why isn't every physicist a Bayesian?*, *Am. J. Phys.* **63** (1995) 398.
- [4] G.J. Feldman and R.D. Cousins, *A Unified approach to the classical statistical analysis of small signals*, *Phys. Rev. D* **57** (1998) 3873 [physics/9711021].
- [5] R.D. Cousins, *Lectures on Statistics in Theory: Prelude to Statistics in Practice*, 1807.05996.
- [6] H. Jeffreys, *An invariant form for the prior probability in estimation problems*, *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences* **186** (1946) 453.
- [7] G. Ashton et al., *Nested sampling for physical scientists*, *Nature* **2** (2022) [2205.15570].
- [8] N. Jeffrey and B.D. Wandelt, *Evidence Networks: simple losses for fast, amortized, neural Bayesian model comparison*, 2305.11241.
- [9] LIGO SCIENTIFIC, VIRGO collaboration, *Observation of Gravitational Waves from a Binary Black Hole Merger*, *Phys. Rev. Lett.* **116** (2016) 061102 [1602.03837].
- [10] LIGO SCIENTIFIC, VIRGO collaboration, *The basic physics of the binary black hole merger GW150914*, *Annalen Phys.* **529** (2017) 1600209 [1608.01940].
- [11] LIGO SCIENTIFIC, VIRGO collaboration, *Improved analysis of GW150914 using a fully spin-precessing waveform Model*, *Phys. Rev. X* **6** (2016) 041014 [1606.01210].
- [12] LIGO SCIENTIFIC, VIRGO collaboration, *Properties of the Binary Black Hole Merger GW150914*, *Phys. Rev. Lett.* **116** (2016) 241102 [1602.03840].
- [13] C. Cutler and E.E. Flanagan, *Gravitational waves from merging compact binaries: How accurately can one extract the binary's parameters from the inspiral wave form?*, *Phys. Rev. D* **49** (1994) 2658 [gr-qc/9402014].