One, No One, and One Hundred Thousand

Inferring the properties of a population in presence of selection effects

Salvatore Vitale * 1,2

1Department of Physics and Kavli Institute for Astrophysics and Space Research, Massachusetts Institute of Technology, 77 Massachusetts Ave, Cambridge, MA 02139, USA
2LIGO Laboratory, Massachusetts Institute of Technology, 185 Albany St, Cambridge, MA 02139, USA

July 14, 2020

Abstract

This document contains a pedagogical introduction to hierarchical Bayesian inference in presence of selection effects. The main equations are obtained from first principles, step by step. Two examples are provided which show how things work, in practice. While I had in mind gravitational-wave astrophysics as I drafted these notes, they should be generic enough to be useful to anyone who wants to learn about the inference on the properties of a population of sources.

1 Are these notes for you?

One of the most exciting prospects when working with an experiment that is detecting a set of sources is to use those sources to learn about the properties of the underlying population. That is, every source that is detected will have its own set of unique properties that make it unique and special, but taken all together they might tell us something about the physical, astrophysical, biophysical, ..., cosmological mechanism that is generating all the sources. The analysis can be made complicated by the fact that not all sources are equally easy to detect [1 2], due to the sensitivity of the detector or to other constraints. If not properly accounted for, these selection effects will result in a biased analysis of the population, with a bias that will usually grow with the population size [3]. This is usually called a selection bias, or Malmquist bias.

The set of sources can be really anything. A very incomplete list includes neutrinos [4], exoplanets [5 6], gamma ray bursts [7 8]; or even be an heterogeneous set of different messengers [9 10].

Given what I do for a living, these notes will mostly focus on transient gravitational-wave (GWs) signals emitted by compact binary coalescences [11 12], but they should be generic enough to be useful for many scientists dealing with population analysis in presence of selection effects.

In these notes you will not find any new scientific results, or new tricks that have not been suggested and exploited elsewhere. The reason why I wrote this document is to have all the relevant information in the same place, with a consistent notation, and at a pedagogical level, where no steps are skipped, and the final expressions are derived from first principles. I will use these for future students or summer students in my group, but I hope they can be of use for others. In the body I

*svitale@mit.edu
will refer to scientific papers were similar derivations or expressions are used, though the notation might be slightly different. I tried to keep my notation consistent throughout this document. Since it eventually became 50 pages long, I will likely have failed. Feel free to drop me an email if you see anything weird or wrong, I will update these notes every once in a while.

These notes might be useful for you if you want to learn in more details the physics and statistics behind some of the gravitational-wave literature, or really any branch of science dealing with similar, common, issues. This document is certainly not the only one available online that focus on hierarchical analysis with a pedagogical focus: the interested reader should also be sure to check out Refs. [4, 3, 13, 14].

The rest of this document is organized as follows: in Section 2 we present the basics of population, sources and hyper parameters; in Section 3 we introduce selection effects and show how to calculate the fraction of detectable sources; in Section 4 and Section 5 we calculate the likelihood and the posterior density function of the population hyper parameters, respectively; in Section 6 we go through a simple one dimensional example, while a more complicated example inspired by gravitational-wave detections is given in Section 7. A useful glossary, including a list of all the main symbols used in the document, is provided at the end.

A version of this document with white text over a black background can be found following the “other formats” link in the main Arxiv page.

2 Source production, rates, populations

There is a physical process that generates sources at a rate:

$$S_t(t) \equiv \frac{dN^s(t)}{dt}$$

(1)

where $N^s(t)$ is the number of sources at time $t$. Throughout these notes, we will use an index or letter “$s$” (possibly capitalized) to refer to source counts, rates, etc. When we need to refer to background events, we will instead use a letter or index “b”.

If the source rate is constant in time one can write:

$$\frac{dN^s(t)}{dt} = S_t \implies N^s(t) = (t - t_0) S_t + N^s(t_0)$$

(2)

with $t_0$ an arbitrary reference time.

It often is useful to consider different types of differential source rates. For example, when studying the mergers of compact binaries through gravitational-wave emission [15], one is usually interested in the number of mergers, per unit time per unit comoving volume [16, 17, 18, 11]. To keep the notation compact and easy too generalize, we will just add more symbols to the underscore list:

$$\frac{d^2N^s(t)}{dtdV_c} \equiv S_{t,V_c}(t)$$

(3)

So that, more generally, if we need to refer to the differential source rate over a set of parameters belonging to a vector $\vec{\kappa}$ we will write:

$$S_{\kappa_1,\ldots,\kappa_n}(t) \equiv S_{\vec{\kappa}}(t) \equiv \frac{d^nN^s(t)}{d\kappa_1\ldots d\kappa_n} \equiv \frac{dN^s(t)}{d\vec{\kappa}}$$

(4)

(Hereafter, we will not explicitly report the time dependence in every equation, but only in those where more than one time variable appear).
If each of the sources can be characterized by a set of parameters $\vec{\theta}$ (e.g. their masses, distances, etc) we can talk of the fraction of the $N^*$ sources with parameters in the range $[\vec{\theta}, \vec{\theta} + d\vec{\theta}]$:

$$p(\vec{\theta}) = \frac{1}{N^*} \frac{dN^*}{d\vec{\theta}} = \frac{1}{N^*} S_{\vec{\theta}}$$

(5)

so that

$$\delta N^*(\vec{\theta}) \equiv N^* p(\vec{\theta}) \delta \vec{\theta}$$

(6)

is the number of events with parameters in the range $[\vec{\theta}, \vec{\theta} + \delta \vec{\theta}]$.

These quantities are normalized as one would expect:

$$\int d\vec{\theta} S_{\vec{\theta}} = N^*$$

(7)

and

$$\int d\vec{\theta} p(\vec{\theta}) = 1$$

(8)

Now, in general there will exist some underlying factors – for example, the average metallicity at some redshift - that can affect both the total number of sources, or their time rate, and the distribution of the individual source parameters $\vec{\theta}$. We will call $\vec{\lambda}$ the set of parameters that control the shape of the distribution of the source parameters, and treat the overall source number $N^*$ as a free scale parameter. When necessary, we can refer to the shape and scale parameters together as:

$$\Lambda \equiv \{\vec{\lambda}, N^*\}$$

These are often called hyper parameters (to contrast them with the individual source parameters, $\vec{\theta}$). The distribution of events’ parameters, thus becomes a function of $\vec{\lambda}$:

$$p(\vec{\theta}) \rightarrow \pi(\vec{\theta} | \vec{\lambda})$$

(9)

We will refer to $\pi(\vec{\theta} | \vec{\lambda})$ as the population prior, and use the symbol $\pi$ instead of $p$ to indicate that it is a prior distribution (in a Bayesian sense), i.e. a distribution that conveys what we believe we know about $\vec{\theta}$ before we analyze the data. Note that a prior distribution might be informed by previously analyzed data: what data this distribution is prior to should be evident from the context.

In some of the literature, the population prior is indicated with an explicit “pop” label, but I will not do it since the presence of $\vec{\lambda}$ on the right side of the conditional bar makes this unambiguous.

This makes most of the interesting quantities depend on the hyper parameters. For example the number of sources per unit $\vec{\theta}$, per unit time, and per unit time-$\vec{\theta}$:

$$S_{\vec{\theta}}(\Lambda) = N^* \pi(\vec{\theta} | \vec{\lambda}) ; \quad S_t(\Lambda) = \frac{dN^*(\Lambda)}{dt} ; \quad S_{t, \vec{\theta}}(\Lambda) = \frac{d^2N^*(\Lambda)}{dt d\vec{\theta}}$$

(10)

In general I will not systematically report the dependence on $\Lambda$, $N^*$ or $\vec{\lambda}$ explicitly, unless needed, as it is basically everywhere and would only make formulae cumbersome. If in doubt, you can always refer to the Glossary, where the dependence of various key variables is explicitly spelled out.

One talks of hierarchical inference when a collection of detections from the population is used to infer $\Lambda$.

---

1 In these notes we will only consider two layers of inference
It is hierarchical because two levels of inference are involved: one must first characterize the individual sources that are detected, i.e. measure $\hat{\theta}$ for all sources, then use them all to learn something about the population parameters $\Lambda$.

### 3 Selection effects

In general, these observations will represent a biased sample of the true underlying population since some sources will be harder, or impossible, to detect. Therefore, an important question we need to address is the following: given a source with a specific set of parameters $\hat{\theta}$, would we be able to detect it, or not? The answer is not always certainly yes or not, because our detector is noisy, and depending the noise realization the same source might be more or less easy to detect. The best we can do is to fold in the analysis the fact that the noise is not entirely predictable, and calculate the probability that a source with parameters $\hat{\theta}$ is detectable:

$$
p(\text{det}|\hat{\theta}) = \int_{\mathcal{D}} d\mathcal{D} \; p(\text{det}|\mathcal{D}\hat{\theta})p(\mathcal{D}|\hat{\theta}) \tag{11}
$$

where $\mathcal{D}$ is the set of all possible data, i.e. the possible noise realizations and we have used the sum rule of probability, since $\mathcal{D}$ represents a complete set of mutually exclusive outcomes. We can split the integration domain into two disjoint sets: one corresponding to the data that will yield a detection statistic $\rho_{\mathcal{D}}$ (e.g. a signal-to-noise ratio) above some pre-determined threshold:

$$
\mathcal{D}_\uparrow \equiv \mathcal{D} : \rho_{\mathcal{D}} \geq \rho_{\text{thr}}
$$

and its complement

$$
\mathcal{D}_\downarrow \equiv \mathcal{D} : \rho_{\mathcal{D}} < \rho_{\text{thr}}
$$

where $\rho_{\text{thr}}$ is the threshold detection statistic. Here and in the rest of these notes we will use an upward arrow, as in $\mathcal{D}_\uparrow$, to refer to quantities that are associated with detectable sources. Similarly, a downward arrow is used for quantities associated with undetectable sources. In some of the literature the label “det” is used. I prefer the arrow because “det” could refer to both detectable and detected, which are very different concepts, which I want to keep well separated. Everything that is detected was detectable, but the opposite is not necessarily true (e.g. a detector might only be online 50% of the time, which would halve the number of detected signals over some time range, but leave the number of detectable source unchanged). The only exception will be $p(\text{det}|\hat{\theta})$, which for consistency I should have called something like $p(\rho_\uparrow|\hat{\theta})$, but the use of “p-det” is so widespread in the gravitational-wave literature that it would make many of you a disservice if I changed it here.

Having partitioned the data this way, we can write:

$$
p(\text{det}|\hat{\theta}) = \int_{\mathcal{D}} d\mathcal{D} \; p(\text{det}|\mathcal{D}\hat{\theta})p(\mathcal{D}|\hat{\theta}) = \int_{\mathcal{D}_\uparrow} d\mathcal{D} \; p(\text{det}|\mathcal{D}\hat{\theta})p(\mathcal{D}|\hat{\theta}) + \int_{\mathcal{D}_\downarrow} d\mathcal{D} \; p(\text{det}|\mathcal{D}\hat{\theta})p(\mathcal{D}|\hat{\theta}) \tag{12}
$$

By construction, $p(\text{det}|\mathcal{D}\hat{\theta}) = 1 \; (p(\text{det}|\mathcal{D}\hat{\theta}) = 0)$ if $\mathcal{D} \in \mathcal{D}_\uparrow \; (\mathcal{D} \in \mathcal{D}_\downarrow)$ thus the second integral is zero and one has:

---

*That is: $\mathcal{D}$ is not the data that was actually collected, but rather all the data that might possibly have been collected.*
This expression gives us the probability that a source with parameters $\vec{\theta}$ is detectable. We often need the total number of sources that are detectable by our instruments over some period of time, irrespective of their parameters. This can be done by calculating the number of sources produced per unit $\vec{\theta}$, multiply by the probability of detecting those sources, and integrating over all possible $\vec{\theta}$. This gives for the the number of detectable sources, given some population hyper parameters:

$$N^* \equiv \int d\vec{\theta} \, p(\text{det}|\vec{\theta}) = N^* \int d\vec{\theta} \, \pi(\vec{\theta}|\lambda)p(\text{det}|\vec{\theta}) \equiv N^* \alpha(\vec{\lambda})$$

(15)

where we have defined the fraction of detectable events:

$$\alpha(\vec{\lambda}) \equiv \frac{N^*}{N^*} = \int d\vec{\theta} \, \pi(\vec{\theta}|\lambda)p(\text{det}|\vec{\theta})$$

(16)

In going from the central to the right expression in Eq. (15) we have used Eq. (5)

Notice that while the probability that a source with true parameters $\vec{\theta}$ is detectable does (obviously) depend on the parameters, the criterium used to decide if a piece of data contains a signal or not is based on the data, not on the parameters (which we would not usually know, anyways). In the expressions above this is explicitly indicated by the fact that the partition of the integrals is made according to some function of the data, $\rho_D$.

4 Hierarchical likelihood

The main goal of the analyst is usually to calculate a posterior distribution for the population hyper parameters:

$$p(\Lambda|D\mathcal{H}_\Lambda)$$

(17)

where $D$ here indicates all of the available data produced by our detector(s) and $\mathcal{H}_\Lambda$ is a proposition that represents all of the assumptions, settings, modeling, population parameters and range, population priors, etc, used in our hyper analysis (when I need to refer to a generic set of settings I will use $\mathcal{H}$ without any subscript). It is important to remember that something like $\mathcal{H}$ is always present on the right of your conditional bars, even when you don’t write it down. Usually, scientific papers provide in the body of the paper details about their analysis (start and stop time of the data segments used, data processing tools, windowing, sensitivity of the detector, the detection criterium, how much coffee the analyst had in the morning, etc): all of these factors, and more, should be indicated explicitly. With this in mind, I am going to do just like you do and won’t write $\mathcal{H}$ down explicitly in each and every equation. I will still write down occasionally, especially if nothing else is on the right hand side of the conditional bar, or when things might otherwise be ambiguous because several possible settings are at play. But I will get back to $\mathcal{H}$ in a few pages, when we hit an apparent contradiction and obtain two numbers that look like they should be the same, but are not.

---

But you can clearly see why it would be prohibitive to explicitly write them all on the right of a conditional bar in every equation.
In Eq. (17) we are using a different symbol for the data than we did in the previous section, since before we wanted to refer to all possible realizations of the noise that could theoretically be produced, while keeping the source the same. Here \( D \) is the actual realization of the data our experiment recorded. In the case of gravitational-wave detectors, for example, \( D \) would be a time series reporting the strain.

In this section I will closely follow the derivation provided by Loredo and Lamb in the context of hierarchical inference of neutrino observations [4], making minor changes to make the notation uniform with mine. I am not going to add much to what you can find in that (great) paper. Thus all the credit goes to them, and all the typos to me.

We imagine dividing the whole data set into a large sequence of small chunks, or segments, each of duration \( \delta t \). The actual duration of \( \delta t \) is not important, as long as it is small enough that the probability of having more than one trigger (see definition in the next paragraph, or in the Glossary) in a time \( \delta t \) is negligible. We will exploit this soon.

At this point it's also worth reminding ourselves that, as explained in the previous section, we have identified a detection criterium \( \rho > \rho_{\text{thr}} \) based on the data, that we can use to decide if any piece of data contains something more than noise, or not.

If we use that criterium in all the small chunks of \( D \), there will be two types of data chunks: those for which \( \rho \) is above threshold, and the others. Notice that finding that a piece of data contains something above threshold, as just defined, does not automatically imply that it contains an interesting astrophysical (or physical) signal. Unfortunately, most detectors suffer from background transient events, which might trigger the detection criteria, but are not what we sought (in gravitational-wave detectors, those are usually called glitches [19]). From now on I will use detection uniquely to refer to astrophysical signals and trigger to anything that rises above the detection threshold, no matter of whether it’s astrophysical or a background.

Following [4], we use the index \( i \) to label the data chunks that have been found to contain a trigger. Therefore, \( d_i \) is the \( i \)-th data chunk with a trigger, and covers the time range \( \delta t_i \equiv [t_i, t_i + \delta t] \). We will use the index \( j \) to label data chunks without triggers. Therefore \( \bar{d}_j \) is the piece of data included in the time interval \( \delta t_j \equiv [t_j, t_j + \delta t] \), and it does not contain any trigger.

The collection of \( \{d_i\}, \{\bar{d}_j\}, \forall \ i, j \) represents a complete and non-overlapping partition of the dataset \( D \). We can write the hyper posterior using Bayes theorem as:

\[
p(\Lambda | D_H) = p(D|\Lambda_H) \frac{\pi(\Lambda|H_H)}{p(D|H_H)}
\]

As in the rest of these notes, we use \( \pi \) for priors: \( \pi(\Lambda|H_H) \) thus represents the probability density function for \( \Lambda \) that we use before looking at the current set of GW data. \( p(D|H_H) \) is the evidence of the data marginalized over the hyper parameters (also called the hyper evidence):

\[
p(D|H_H) \equiv \int d\Lambda \ p(D|\Lambda_H)\pi(\Lambda|H_H)
\]

and usually won’t play a role unless we are trying to do model selection among different population models [11]. We will get back to the hyper evidence later, but for now let’s focus on the likelihood of the data, \( p(D|\Lambda_H) \). If we can assume that the noise has an autocorrelation length much smaller than \( \delta t \), then different data segments are statistically independent, and the likelihood factorizes:
\[ p(D|\Lambda H_\Lambda) = p(\{d_i\}, \{\bar{d}_j\}|\Lambda H_\Lambda) = \left[ \prod_{i=1}^{N^{\text{tr}}} p(d_i|\Lambda H_\Lambda) \right] \left[ \prod_{j=1}^{N^{\text{nt}}} p(\bar{d}_j|\Lambda H_\Lambda) \right] \] (20)

where the labels \( \text{tr} \) and \( \text{nt} \) stand for “trigger” and “no-trigger”.

To explicitly calculate the likelihood, we must make some assumption on the generation mechanism of both real sources and background events. We will assume that the probability for a source or a glitch to happen in any given data chunk is independent of what happened in other data chunks. This is a reasonable assumption in most cases. Certainly it is a good assumption for GW detectors at their current sensitivity\(^4\). It is probably not a great assumption for glitches in GW detectors, as they sometimes come in groups \(^5\) (e.g. due to elevated seismic activity, or a faulty connection), but we will for the moment still stick to this assumption, because it is simple and allows making all the important points\(^5\). Under these hypotheses, both the distribution of the number of sources and backgrounds is Poissonian.

If \( N^{\text{exp}} \) sources are expected to occur on average in a given time interval, the distribution of the number of source actually generated, \( N^{\text{act}} \), is

\[ p(N^{\text{act}}|N^{\text{exp}}) = \text{poiss}(N^{\text{act}}|N^{\text{exp}}) = \frac{e^{-N^{\text{exp}}}(N^{\text{exp}})^{N^{\text{act}}}}{N^{\text{act}}!} \] (21)

where \( \text{poiss}(O|E) \) is the Poisson probability for \( O \) events occurring given that \( E \) were expected.

Please notice that we have not yet said anything about whether these \( N^{\text{act}} \) sources are all detectable or not, these are not necessarily all detections!

Ok, but what is the expected number of sources in each of our time intervals \( \delta t \)? It can be simply written as the integral of the time-rate over the relevant time interval:

\[ N^{\text{exp}} = \int_{\delta t} dt S(t) = \int_{\delta t} dt \frac{dN^s}{dt} \] (22)

But, remember, because the production mechanism depends on the hyper parameters, so does \( N^{\text{exp}} \). Let’s write down this explicitly just this once:

\[ N^{\text{exp}}(\Lambda) = \int_{\delta t} dt S(t)(\Lambda) = \int_{\delta t} dt \frac{dN^s(\Lambda)}{dt} \] (23)

As long as the time rate is not a strong function of time, and can be treated as approximately constant over the small time interval \( \delta t \), we can simplify this expression as:

\[ N^{\text{exp}}(\Lambda) \simeq S(t) \delta t \] (24)

We will assume that this is always the case and use this simplification hereafter.

There is a small catch here: usually we will be interested in the expected number of sources generated in some time-interval \emph{as measured by the sources clocks}. When the sources are distributed at cosmological distances, or are moving relativistically, there will be conversions to be done between the differential rate per unit time in the rest frame of the source, and the one in the detector frame. We will do this explicitly later, in Section 7.

\(^4\)It might become less robust of an assumption when the sensitivity of the detectors is such that a significant number of events will be gravitationally lensed, which will make what happens in the chunk \( d_i \) not totally unrelated to what happened other segments.

\(^5\)If the background production relevant for your detector clearly violates these assumptions, and you know how to model it, you can modify most of the results below accordingly.
Similar expressions hold for the distribution of the actual number of background events:

\[ p(N_{\text{act}}^b|N_{\text{exp}}^b) = \text{poiss}(N_{\text{act}}^b|N_{\text{exp}}^b) = \frac{e^{-N_{\text{exp}}^b} (N_{\text{exp}}^b)^{N_{\text{act}}^b}}{N_{\text{act}}^b!} \]  

(25)

where the expected number of background events over the time interval \( \delta t \) is

\[ N_{\text{exp}}^b = \int \delta t \frac{dN^b}{dt} = \int \delta t B_t \]  

(26)

In these expression, \( B_t(t) \) is defined as the number of background events per unit time, which we assume can be measured experimentally, or predicted theoretically or empirically. As for the source rate, as long as \( B(t) \) does not vary too quickly, we can write the expected number of background events in the data chunk as:

\[ N_{\text{exp}}^b \simeq \delta t B_t \]  

(27)

We will assume this is always the case \(^6\).

### 4.1 Calculating the likelihood of data without a trigger

We now have everything we need to explicitly calculate Eq. (20). Please, make yourself comfortable, be sure you have a tea or your favorite drink nearby: this will take a while. We start by looking at a generic data chunk \( \bar{d}_j \) where no triggers were found. We need to calculate \( p(\bar{d}_j|\Lambda) \). This term is the probability of not reporting any trigger in the data chunk, given that the true value of the hyper parameters is \( \Lambda \), plus all other modeling and assumptions we have made \(^7\).

There are multiple reasons why no trigger might be found in the data: maybe nothing actually happened in that time interval and thus there was nothing to detect in the first place; or maybe a source or a background event \( \textit{did} \) occur, but their characteristics were such that the data did not match our criteria to claim a detection.

Let’s formalize this mathematically.

Following \(^4\), I will introduce the following propositions:

\[ S_0 \equiv \text{“No astrophysical source occurred in the time interval of interest”} \]
\[ S_1 \equiv \text{“Exactly one astrophysical source occurred in the time interval of interest”} \]
\[ B_0 \equiv \text{“No background event occurred in the time interval of interest”} \]
\[ B_1 \equiv \text{“Exactly one background event occurred in the time interval of interest”} \]

These prepositions are agnostic about the parameters of the trigger. Please notice the use of the word “occurrence”, which does not imply either detectability or detection (See the Glossary at the end of this document).

We can write the likelihood for non-detection by using the sum rule and the propositions above:

\[ p(\bar{d}_j|\Lambda) = p(\bar{d}_j|S_0B_0 + S_0B_1 + S_1B_0)|\Lambda) \]  

(28)

\(^6\)In some cases, \( B_t \) might actually be time-independent, in which cases this equation is exactly valid.

\(^7\)Remember I’m dropping \( H_\Lambda \) on the right of the conditional bar, until I’ll regret it
For us to be able to write this as an identity, it should be the case that \( (S_0B_0 + S_0B_1 + S_1B_0) = 1 \), while theoretically more than one trigger could occur during the time \( \delta t \). However, because we picked \( \delta t \) to be very small, the probability of more than one trigger occurring in that time is so small that we treat it as zero.

We therefore can use the sum rule considering only the tree terms above, and expand as:

\[
\begin{align*}
p(\bar{d}_j | \Lambda) & = p(\bar{d}_j (S_0B_0 + S_0B_1 + S_1B_0) | \Lambda) \\
& = p(\bar{d}_j S_0B_0 | \Lambda) + p(\bar{d}_j S_0B_1 | \Lambda) + p(\bar{d}_j S_1B_0 | \Lambda) \quad (29)
\end{align*}
\]

where we used the fact that all of the four propositions above are mutually exclusive, within our analysis setup.

We now proceed and calculate each of these three terms.

### 4.1.1 The good: \( p(\bar{d}_j S_0B_0 | \Lambda) \)

This is the probability of no trigger being found in the data and no signals occurring and no background occurring, given the hyper parameters.

We can use the product rule and write it as:

\[
p(\bar{d}_j | S_0B_0 \Lambda) = p(\bar{d}_j | S_0B_0 \Lambda) p(S_0B_0 | \Lambda) = p(\bar{d}_j | S_0B_0 \Lambda) p(S_0 | \Lambda) p(B_0 | H) \quad (30)
\]

where in the last step we have used the fact that \( S_0 \) and \( B_0 \) are independent to factorize their joint probability, and the fact that \( B_0 \) does not depend on the astrophysical source hyper-parameters to write \( p(B_0 | \Lambda) = p(B_0 | H) \)

The first term, \( p(\bar{d}_j | S_0B_0 \Lambda) \), is trivial: the probability of nothing raising above the detection threshold given that nothing happened is 100%:

\[
p(\bar{d}_j | S_0B_0 \Lambda) = 1
\]

(notice how \( \Lambda \) is not really playing any role here, since no source occurred, and it could have been removed from the right hand side of the conditional bar).

Let’s now move to \( p(B_0 | H) \), that is the probability that no background occurred in the \( j \)-th time chunk. Using the Poisson distribution, Eq. (25), this is the probability of 0 occurrences given that \( B_0 \delta t \) were expected:

\[
p(B_0 | H) = \text{poiss}(0 | N_b \exp = B_0 \delta t) = e^{-B_0 \delta t} \quad (31)
\]

Similarly for \( p(S_0 | \Lambda) \):

\[
p(S_0 | \Lambda) = \text{poiss}(0 | N_s \exp = S_0 \delta t) = e^{-S_0 \delta t} \quad (32)
\]

and putting all together we have:

\[
p(\bar{d}_j S_0B_0 | \Lambda) = e^{-[B_0 + S_0] \delta t} \quad (33)
\]

---

8 I’m adding a generic \( H \) on the right of a conditional bar, because it horrifies me to see nothing there.
4.1.2 The bad: \( p(\tilde{d}_jS_1B_0|\Lambda) \)

Let’s use the product rule on this term but without fully expanding all terms:

\[
p(\tilde{d}_jS_1B_0|\Lambda) = p(\tilde{d}_jS_1|B_0\Lambda)p(B_0|\Lambda) = p(\tilde{d}_jS_1|\Lambda)p(B_0|\Lambda)
\]

(34)

The last term, the probability of no background event occurring, has already been calculated in the previous subsection, Eq. (31):

\[
p(B_0|\Lambda) = e^{-B_0\delta t}
\]

(35)

Let’s now look at the first term, the probability that one source occurs and no triggers are reported, given \( \Lambda \).

The proposition \( S_1 \) does not specify the parameters of the source that occurred, but in order to do explicit calculations we need to specify some value of parameters. This can be achieved by marginalizing over all of the possible source parameters:

\[
p(\tilde{d}_jS_1|\Lambda) = \int d\vec{\theta}p(\tilde{d}_jS_1|\vec{\theta},\Lambda)
\]

(36)

and using the product rule:

\[
p(\tilde{d}_jS_1|\Lambda) = \int d\vec{\theta}p(\tilde{d}_jS_1|\vec{\theta},\Lambda) = \int d\vec{\theta}p(\tilde{d}_j|S_1\vec{\theta},\Lambda)p(S_1\vec{\theta}|\Lambda).
\]

(37)

\( p(S_1\vec{\theta}|\Lambda) \) is the probability density that one and only one event occurred, and that it had parameters within a small interval \( \delta\vec{\theta} \) around \( \vec{\theta} \). The probability itself is:

\[
p(S_1\vec{\theta}|\Lambda)\delta\vec{\theta}
\]

(38)

We need to be careful when calculating this: this is not just the Poisson probability for the occurrence of one source with parameters \( \vec{\theta} \) when \( N^s_{\exp}(\vec{\theta},\Lambda) \) sources with parameters \( \vec{\theta} \) were expected. This would be:

\[
p(\text{One source occurrence with parameters } \vec{\theta}|\Lambda) = \text{poiss}(1|N^s_{\exp}(\vec{\theta},\Lambda)) = N^s_{\exp}(\vec{\theta},\Lambda)e^{-N^s_{\exp}(\vec{\theta},\Lambda)}(\delta t\delta\vec{\theta}S_{t,\vec{\theta}})^e(\delta t\delta\vec{\theta}S_{t,\vec{\theta}})\equiv(\delta t\delta\vec{\theta}S_{t,\vec{\theta}})\equiv(\delta t\delta\vec{\theta}S_{t,\vec{\theta}})\equiv(\delta t\delta\vec{\theta}S_{t,\vec{\theta}})
\]

(40)

Where we have used the syntax of Eq. (4) to write the number of events in an interval \( \delta t \) and \( \delta\vec{\theta} \) as:

\[
N^s_{\exp}(\vec{\theta},\Lambda) = \int d\vec{\theta}d\vec{\theta}d\theta_\delta(\vec{\theta},\Lambda)d\theta_\delta d\theta_\delta = \frac{d^2N^s}{d\theta_\delta d\theta_\delta}(\vec{\theta},\Lambda)d\theta_\delta d\theta_\delta \equiv d^2\theta_\delta S_{t,\vec{\theta}}(\vec{\theta},\Lambda)
\]

(41)

In fact Eq. (38) expresses a much stronger requirement: that one event with parameters in the interval \( \delta\vec{\theta} \) around \( \vec{\theta} \) happened and that no other events with any other values of the parameters happened. The simple Poisson probability we just calculated doesn’t say anything about what else might have happened in that time interval, thus we need to explicitly calculate this extra term:
\[ p(\text{No source occurrence with any parameters different from } \vec{\theta} | \Lambda) \quad (42) \]

This looks like a Poisson probability for zero occurrences, but what is the number of expected events with parameters other than \( \vec{\theta} \)? That it’s going to be the total number of expected events minus the number of expected events with parameters \( \vec{\theta} \):

\[ N^s_{\exp}(\sim \vec{\theta}) = S_t \delta t - S_{t, \vec{\theta}} \delta t \delta \vec{\theta} \quad (43) \]

and Eq. (42) becomes:

\[ p(\text{No source occurrence with any parameters different from } \vec{\theta} | \Lambda) = \text{poiss}(0 | N^s_{\exp}(\sim \vec{\theta})) \quad (44) \]

The quantity we want to calculate, Eq. (38) is thus the logical “and” of the propositions expressed in Equation (40) and Eq. (42):

\[ p(S_1 \vec{\theta} | \Lambda) \delta \vec{\theta} = p(\text{One source occurrence with parameters } \vec{\theta} | \Lambda) \times \]

\[ \times p(\text{No source occurrence with any parameters different from } \vec{\theta} | \Lambda) = \]

\[ = (\delta t \delta \vec{\theta} S_{t, \vec{\theta}}) e^{- (\delta t \delta \vec{\theta} S_{t, \vec{\theta}})} e^{- N^s_{\exp}(\sim \vec{\theta})} \]

\[ = (\delta t \delta \vec{\theta} S_{t, \vec{\theta}}) e^{- (\delta t \delta \vec{\theta} S_{t, \vec{\theta}})} e^{- [S_t \delta t - S_{t, \vec{\theta}} \delta t \delta \vec{\theta}]} \]

\[ = \delta t \delta \vec{\theta} S_{t, \vec{\theta}} e^{- S_t \delta t} \quad (46) \]

so that:

\[ p(S_1 \vec{\theta} | \Lambda) = \delta t S_{t, \vec{\theta}} e^{- S_t \delta t} \quad (47) \]

Notice that the term in the exponent is the expected number of event occurrences in the time interval irrespective of their parameters!

Finally, we need to calculate \( p(\bar{d}_j | S_1 \vec{\theta} \Lambda) \), this is the probability of recording a piece of data without any trigger above the threshold, given that one source occurred with parameters in the interval \([\vec{\theta}, \vec{\theta} + \delta \vec{\theta}]\). This is one minus the probability of detecting the trigger:

\[ p(\bar{d}_j | S_1 \vec{\theta} \Lambda) = 1 - p(\text{The trigger is detected} | S_1 \vec{\theta} \Lambda) \quad (48) \]

Notice that we can drop \( S_1 \) and \( \Lambda \) on the right of the conditional bar of the last term, since \( \vec{\theta} \) already tell us all we need:

\[ p(\text{The trigger is detected} | S_1 \vec{\theta} \Lambda) = \]

\[ p(\text{The trigger is detected} | \vec{\theta}) = \]

\[ p(\text{The trigger is detectable} | \vec{\theta}) = \]

\[ p(\text{det} | \vec{\theta}) \quad (49) \]

In the third line we have switched “detected” with “detectable” because in the calculation we are carrying out a piece of data \((\bar{d}_j)\) was collected. This excludes the possibility that a detectable
source might have occurred but not being detected. Therefore, this is exactly the same quantity we defined earlier, in Eq. (11) and Eq. (14).

Putting all together:

\[
p(\bar{d}_j S_1 B_0 | \Lambda) = e^{-[B_t + S_t] \delta t} \int d\hat{\theta} \left[ 1 - p(\det | \bar{\theta}) \right] S_{t, \hat{\theta}}
\]

(50)

4.1.3 The ugly: \( p(\bar{d}_j S_0 B_1 | \Lambda) \)

That was long, but satisfactory. We are nearly done, there is only one term to calculate (At this point some of you, like me, might have forgotten what it is that we are originally calculating... that would be \( p(d_j | \Lambda) \)).

Let’s again use the product rule:

\[
p(\bar{d}_j S_0 B_1 | \Lambda) = p(\bar{d}_j S_0 B_1 | H) p(S_0 | \Lambda) = p(\bar{d}_j S_0 B_1 | H) \]

(51)

where we have used the fact that no source occurred, and hence we cannot learn anything about the data from \( S_0 \) and \( \Lambda \), to simplify one of the terms.

We already calculated \( p(S_0 | \Lambda) \) in the previous subsections, Eq. (32):

\[
p(S_0 | \Lambda) = e^{-S_t \delta t}
\]

(52)

Let’s consider the other term, \( p(\bar{d}_j B_1 | H) \). This is the probability that a background event occurred and that no trigger was reported (\( H \) reminds us of the other things we assumed but are not explicitly listing: no sources occurred, our search algorithm was configured in a certain way and uses a \( \rho_{thr} \) to select triggers, etc). To make progress we need to follow a procedure similar to the one in Section 4.1.2, and expand this across all the possible \( \hat{\theta} \):

\[
p(\bar{d}_j B_1 | H) = \int d\hat{\theta} p(\bar{d}_j B_1 | H) = \int d\hat{\theta} p(\bar{d}_j | B_1 H) p(B_1 | H)
\]

(53)

This might seem puzzling at first: why are we using the astrophysical parameters of the sources, \( \hat{\theta} \), now that only background events are being considered? What we are doing, really, is to assume that background events, or at least the ones we care about, i.e. the ones that can be “detected”, can be characterized with the same parameters that are used to describe the astrophysical sources. This is not entirely correct, and if we had a better model and parametrization for the background, we should use that instead. However, If we think about it, this is not totally crazy either: since those background events are detected by an algorithm that was looking for an astrophysical signal with some unknown set of parameters \( \hat{\theta} \), it is not too far fetched to assume that the background event too can be described by similar parameters. Obviously, in the case of a background, the meaning of the measured \( \hat{\theta} \) is not immediately clear. Assuming the background can be described with the same parameters as the sources, has the advantage of making the math easier, since we can literally follow the same steps of the previous subsection and calculate \( p(B_1 | H) \) as the product of two terms: the probability of one background event occurring with parameters in \( [\hat{\theta}, \hat{\theta} + \delta \hat{\theta}] \) and the probability of zero background events happening with any other parameters. The result is:

\[
p(B_1 | H) = \delta t B_{t, \hat{\theta}} e^{-B_t \delta t}
\]

(54)

where
\[ B_{t, \theta} \equiv \frac{d^2 N^b}{dtd\theta} \]  

is the differential rate of background events per unit time and per unit \( \theta \).

There is only one term left to calculate, \( p(\bar{d}_j | B_1 \bar{\theta}H) \) i.e. the probability that a background event occurring with parameters \( \bar{\theta} \) would not trigger a detection. This is one minus the probability that it would trigger a detection:

\[ p(\bar{d}_j | B_1 \bar{\theta}H) = 1 - p(\text{det} | B_1 \bar{\theta}H) \]

Here we are going to assume that our experiment and search algorithm are blind about the true nature of the trigger, whether it’s an astrophysical source or a background event. If a background and a real source occurred with the same set of parameters \( \theta \) in two pieces of data with the same noise realization, they will be both classified as a detection or a non-detection. This is a quite good assumption: if the detector or the algorithm had a way of distinguish these two situations, we would use that as a tool to get rid of background events. The very fact that we are here talking about them implies that at least a sneaky fraction of the background events looks and walks like our real sources!

This implies:

\[ p(\text{det} | B_1 \bar{\theta}H) = p(\text{det} | \bar{\theta}H) \]

which is the same probability of detection we used in the previous section, and introduced in Eq. (11). Putting all together:

\[ p(\bar{d}_j S_0 B_1 | \Lambda) = e^{-[S_t + B_t] \delta t} \delta t \int d\bar{\theta} \left[ 1 - p(\text{det} | \bar{\theta}) \right] B_{t, \bar{\theta}} \]  

(56)

4.1.4 Putting the pieces together

Ok, we now have all the pieces we need to calculate Eq. (29) (do you remember...? The number number we actually wanted to calculate?).

\[ p(\bar{d}_j | \Lambda) = p(\bar{d}_j S_0 B_0 | \Lambda) + p(\bar{d}_j S_0 B_1 | \Lambda) + p(\bar{d}_j S_1 B_0 | \Lambda) \]  

(57)

Using Eq. (33), Eq. (50) and Eq. (56):

\[
\begin{align*}
p(\bar{d}_j | \Lambda) & = e^{-[B_t + S_t] \delta t} \\
& + e^{-[B_t + S_t] \delta t} \int d\bar{\theta} \left[ 1 - p(\text{det} | \bar{\theta}) \right] S_{t, \bar{\theta}} \\
& + e^{-[S_t + B_t] \delta t} \int d\bar{\theta} \left[ 1 - p(\text{det} | \bar{\theta}) \right] B_{t, \bar{\theta}} \\
& = e^{-[B_t + S_t] \delta t} \left[ 1 + \delta t \int d\bar{\theta} \left[ 1 - p(\text{det} | \bar{\theta}) \right] \left( B_{t, \bar{\theta}} + S_{t, \bar{\theta}} \right) \right]
\end{align*}
\]  

(58-61)

We can simplify this expression considerably if we recognize that the second term in the big square brackets is much smaller than 1. Mathematically, this happens because we have taken \( \delta t \ll 1 \), which
implies $\delta t B_t \ll 1$ and $\delta t S_t \ll 1$. Physically, this means that most of the times there are no triggers in our data.

We can exploit this taking the logarithm of the non-detection probability:

$$\log p(\bar{d}_j|\Lambda) \simeq -\delta t \left[ B_t + S_t \right] + \delta t \int d\tilde{\theta} \left[ 1 - p(\det|\tilde{\theta}) \right] \left( B_t,\tilde{\theta} + S_t,\tilde{\theta} \right)$$

(62)

Where we have used $\log(1 + x) \simeq x$ for $x \ll 1$.

The terms with the differential rates of sources and background, when multiplied by 1 can be trivially integrated, and exactly cancel out the two terms outside of the integral, leaving only:

$$\log p(\bar{d}_j|\Lambda) \simeq -\delta t \int d\tilde{\theta} p(\det|\tilde{\theta}) \left( B_t,\tilde{\theta} + S_t,\tilde{\theta} \right)$$

(63)

and switching back to the actual probability:

$$p(\bar{d}_j|\Lambda) \simeq e^{-\delta t \int d\tilde{\theta} p(\det|\tilde{\theta})(B_t,\tilde{\theta} + S_t,\tilde{\theta})}$$

(64)

A last simplification can be made if the instrument noise is stationary, so that $p(\det|\tilde{\theta})$ is time independent:

$$\int d\tilde{\theta} p(\det|\tilde{\theta})S_{t,\tilde{\theta}} = \int d\tilde{\theta} p(\det|\tilde{\theta}) \frac{d^2 N^s}{dt d\tilde{\theta}}$$

$$= \frac{d}{dt} \int d\tilde{\theta} p(\det|\tilde{\theta}) \frac{d N^s}{d\tilde{\theta}} = \frac{dN^s}{dt}$$

(65)

This is the time rate of detectable astrophysical sources. One can similarly define a rate of detectable backgrounds and finally write:

$$p(\bar{d}_j|\Lambda) \simeq e^{-\delta t \left( \frac{dN^s}{dt} + \frac{dN^b}{dt} \right)}$$

(66)

This misleadingly simple expression, obtained with so much use of virtual ink, contains the selection effect of our instruments.

### 4.2 Calculating the likelihood of data with a trigger

Those of us with a significant attention span, will remember that we are only half-done, because we still need to calculate the probability of finding a trigger in the data chunk $d_i$:

$$p(d_i|\Lambda)$$

A trigger can be reported because a background event occurs, or because a source occurs. We can thus use the same propositions introduced in the previous section and write:

$$p(d_i|\Lambda) = p(d_i(S_0B_0 + S_0B_1 + S_1B_0)|\Lambda)$$

$$= p(d_iS_0B_0|\Lambda) + p(d_iS_0B_1|\Lambda) + p(d_iS_1B_0|\Lambda)$$

$$= p(d_iS_0B_1|\Lambda) + p(d_iS_1B_0|\Lambda)$$

(67)
The term $p(d_i, S_0 B_0 | \Lambda)$ is zero because we cannot possibly have recorded a trigger if nothing occurred! But what about noise fluctuations and false positive? Those are background events, and thus would be captured by $B_1$!

We therefore have two terms to calculate. Much of the heavy lifting was done in the previous section, thus this will be much faster.

4.2.1 We got a source: $p(d_i S_1 B_0 | \Lambda)$

Just as we did in Section 4.1.2 we can use the product rule, and marginalize over all possible source parameters. We get:

\[
p(d_i S_1 B_0 | \Lambda) = p(B_0 | \mathcal{H}) p(d_i S_1 | \Lambda B_0) \\
= p(B_0 | \mathcal{H}) \int d\vec{\theta} p(d_i S_1 | \vec{\theta} \Lambda) \\
= p(B_0 | \mathcal{H}) \int d\vec{\theta} p(d_i | S_1 \vec{\theta} \Lambda) p(S_1 \vec{\theta} | \Lambda) \tag{68}
\]

Both the first and the last terms have already been calculated, Eq. (31) and Eq. (47).

The remaining term $p(d_i | S_1 \vec{\theta} \Lambda)$ is the probability of getting the data $d_i$ given that a source occurred with parameters $\vec{\theta}$. This is just the likelihood of the data:

\[
p(d_i | S_1 \vec{\theta} \Lambda) = p(d_i | \vec{\theta}) \tag{69}
\]

the very same function used to calculate $p(\text{det} | \vec{\theta})$ in Eq. (14).

Putting all together:

\[
p(d_i S_1 B_0 | \Lambda) = e^{-[S_1 + B_1] \delta t} \int d\vec{\theta} p(d_i | \vec{\theta} \Lambda) S_1 \vec{\theta} \tag{70}
\]

Notice that the integral here knows nothing about selection effects: the integration range must be performed across all possible $\vec{\theta}$ that the population allows for, regardless of whether they are detectable or not. Similarly, the likelihood function is the “usual” likelihood function, not truncated or affected in any way by the existence of threshold effects.

4.2.2 We got a background: $p(d_i B_1 S_0 | \Lambda)$

We proceed exactly as in the previous subsection and obtain:

\[
p(d_i B_1 S_0 | \Lambda) = e^{-[S_1 + B_1] \delta t} \int d\vec{\theta} p(d_i | \vec{\theta} \Lambda) B_1 \vec{\theta} \tag{71}
\]

Here again, we have assumed that the detector is blind to whether something with parameters $\vec{\theta}$ is a source or a background, hence obtaining the equalities:

\[
p(d_i | S_1 \vec{\theta} \Lambda) = p(d_i | B_1 \vec{\theta} \Lambda) = p(d_i | \vec{\theta})
\]
4.2.3 Putting the pieces together

We can now add the terms obtained in the last two subsections and write Eq. (67) as:

\[ p(d_i|\Lambda) = e^{-[S_t+B_t]\delta t} \int d\tilde{\theta} p(d_i|\tilde{\theta}) \left( S_{t,\tilde{\theta}} + B_{t,\tilde{\theta}} \right) \]  

(72)

4.3 Finishing off

We now have all we need to write down the posterior of \( \Lambda \), Eq. (20). We are going to use the equations derived in the previous sections, explicitly reporting the times at which time-dependent quantities are evaluated:

\[ p(D|\Lambda H_{\Lambda}) = p(\{d_i\}, \{\bar{d}_j\}|\Lambda H_{\Lambda}) = \prod_{i=1}^{N_{tr}} p(d_i|\Lambda H_{\Lambda}) \prod_{j=1}^{N_{nt}} p(\bar{d}_j|\Lambda H_{\Lambda}) \]

\[ = \left[ \prod_{i=1}^{N_{tr}} e^{-[S_t(t_i)+B_t(t_i)]\delta t} \int d\tilde{\theta} p(d_i|\tilde{\theta}) \left( S_{t,\tilde{\theta}}(t_i) + B_{t,\tilde{\theta}}(t_i) \right) \right] \times \left[ \prod_{j=1}^{N_{nt}} e^{-\delta t \left( \frac{dN_{s}^+}{dt}(t_j) + \frac{dN_{b}^+}{dt}(t_j) \right)} \right] \]

(73)

We have not put any indices in \( \delta t \) since we have assumed the data was partitioned in intervals with the same duration (this condition can be relaxed quite trivially, if necessary).

We can write the product of exponentials as:

\[ \prod_{i=1}^{N_{tr}} e^{-[S_t(t_i)+B_t(t_i)]\delta t} = e^{-\delta t \sum_{i=1}^{N_{tr}} [S_t(t_i)+B_t(t_i)]} \equiv e^{-f_{tr} dt[S_t(t)+B_t(t)]} \]

(74)

where we have defined \( T^{tr} \equiv N^{tr} \delta t \) as the total amount of time during which triggers were reported. Similarly for the segments without triggers:

\[ \prod_{j=1}^{N_{nt}} e^{-\delta t \left( \frac{dN_{s}^+}{dt}(t_j) + \frac{dN_{b}^+}{dt}(t_j) \right)} = e^{-\delta t \sum_{j=1}^{N_{nt}} \left( \frac{dN_{s}^+}{dt}(t_j) + \frac{dN_{b}^+}{dt}(t_j) \right)} \]

(75)

where \( T^{nt} \equiv N^{nt} \delta t \), and we note that in both this and the previous expression, the integral will cover disjoint data segments.

If the rate of background and astrophysical source are small enough that most of the times nothing is occurring:

\[ T^{tr} \ll T^{nt} \]

and we can neglect the exponential in Eq. (74), while keeping the integral over non-triggering times, Eq. (75). Furthermore, in Eq. (75) we can also change the integration range from \( T^{nt} \) to the total experiment time \( T \), since the two only differ by the total time with triggers, which is small.
Therefore, we finally obtain:

\[
p(D|\Lambda H) = e^{-f_t \int_0^T \left( \frac{dN_s^\uparrow}{dt}(t) + \frac{dN_b^\uparrow}{dt}(t) \right)} \\
\times \prod_{i=1}^{N_{tr}} \delta t \int d\vec{\theta} p(d_i|\vec{\theta}) \left( S_{t_i,\vec{\theta}}(t_i) + B_{t_i,\vec{\theta}}(t_i) \right)
\]

(76)

This is equation A26 of [4], modulo a different syntax.

4.4 Special cases

The expression we found can be simplified a bit further if more assumptions can be made about the background or source rates.

4.4.1 Constant background rate

If the background rate is constant in time, then we have:

\[
\int_T dt B_t = TB_t = N_b
\]

(77)

This is the total number of background events over the experiment time, i.e. the experiment time multiplied by the constant rate of background triggers. Similarly, the total number of detectable background triggers over the duration of the experiment becomes

\[
\int_T dt \frac{dN_b^\uparrow}{dt} = T \frac{dN_b^\uparrow}{dt} = N_b^\uparrow
\]

(78)

Finally, (cfr Eq. (55)) the differential background rate per unit \(\vec{\theta}\) is now constant in time:

\[
B_{t_i,\vec{\theta}}(t) = B_{t_i,\vec{\theta}} = \frac{B_{\vec{\theta}}}{T}
\]

(79)

and the likelihood simplifies as:

\[
p(D|\Lambda H) = e^{-N_b^\uparrow - f_t \int_0^T \left( \frac{dN_s^\uparrow}{dt}(t) \right)} \\
\times \prod_{i=1}^{N_{tr}} \delta t \int d\vec{\theta} p(d_i|\vec{\theta}) \left( S_{t_i,\vec{\theta}}(t_i) + \frac{B_{\vec{\theta}}}{T} \right)
\]

(80)

4.4.2 Constant background and source rate

Let’s now suppose that the rate of source occurrence is also constant in time:

\[
S_t(t) = S_t = \frac{N_s}{T}
\]

and the term in the exponential becomes:

\[
\int_T dt \frac{dN_s^\uparrow}{dt} = T \frac{dN_s^\uparrow}{dt} = N_s^\uparrow
\]

(81)
which is the total number of detectable sources over the duration of the experiment. The likelihood thus simplifies as:

\[
p(D|\Lambda H_\Lambda) = e^{-\left(N^b + N^s\right)} \\
\times \prod_{i=1}^{N^{tr}} \delta t \int d\bar{\theta} p(d_i|\bar{\theta}) \left(S_{\bar{\theta}} + B_{\bar{\theta}}\right)
\]  

(82)

Where now the exponential contains the total number of expected triggers, irrespective of their nature. For each chunk with a trigger, we integrate the likelihood times the differential background and source rate per unit \(\bar{\theta}\), without any selection effect.

4.4.3 Constant source rate and no background

Finally, the most optimistic scenario, which is also the one that yields the most compact expression, is the one where we can assume both that the source rate is constant in time and that the background rate is zero. Our friends instrument scientists and our search algorithms are so good that we have excluded the possibility that our detection criteria can be met by anything else besides the occurrence of a real source.

In this wonderful world, we have:

\[
p(D|\Lambda H_\Lambda) = e^{-N^s} \prod_{i=1}^{N^{tr}} \int d\bar{\theta} p(d_i|\bar{\theta}) dN^s(\bar{\theta})
\]  

(83)

This is the fundamental equation for hierarchical inference used in most of the gravitational-wave literature. Please notice that I will be using \(N^{tr}\) for the product even if we are assuming that there is no background, thus all triggers are sources. This is because I want to avoid any possible confusion with the true total number of sources \((N^s)\). This is fully consistent with our definition of trigger, see the Glossary at the end of this document.

5 Hierarchical posterior

Let’s now derive the posterior on \(\Lambda\). We will work with the simplest of the possible cases, no background and constant source rate (see [21] for an analysis that accounts for background). In this case we can elaborate Eq. (83) as:

\[
p(D|\Lambda N^s) = e^{-N^s} \prod_{i=1}^{N^{tr}} \int d\bar{\theta} \frac{dN^s}{d\bar{\theta}}(\bar{\lambda}) p(d_i|\bar{\theta})
\]  

(84)

\[
= e^{-N^s} \prod_{i=1}^{N^{tr}} N^s \int d\bar{\theta} \pi(\bar{\theta}|\bar{\lambda}) p(d_i|\bar{\theta})
\]  

(85)

\(^9\)We have removed the term \(\delta t/T\) from inside the product, since it’s a constant term that does not depend on \(\Lambda\), and thus won’t affect its inference.
We notice that in this expression the only place where selection effects are present is in the term \( N^* \). In particular, the likelihood for the data of individual sources does not include any selection effect.

Using Eq. (15) one can write

\[
N^* = \frac{N^* \alpha(\vec{\lambda})}{\alpha(\vec{\lambda})}
\]  

which gives:

\[
p(D|\vec{\lambda}N^*) = e^{-N^{*\tau}} \prod_{i=1}^{N^*} \int d\vec{\theta} \pi(\vec{\theta}|\vec{\lambda}) p(d_i|\vec{\theta})
\]

\[
= e^{-N^{*\tau}} (N^*)^{N^{\text{tr}}} \prod_{i=1}^{N^*} \int d\vec{\theta} \pi(\vec{\theta}|\vec{\lambda}) \frac{p(d_i|\vec{\theta})}{\alpha(\vec{\lambda})}
\]

\[
= e^{-N^{*\tau}} (N^*)^{N^{\text{tr}}} \alpha(\vec{\lambda})^{-N^{\text{tr}}} \prod_{i=1}^{N^*} p(d_i|\vec{\lambda})
\]

\[
= e^{-N^{*\tau}} (N^*)^{N^{\text{tr}}} \alpha(\vec{\lambda})^{-N^{\text{tr}}} p(D|\vec{\lambda})
\]

With the likelihood in hand, we can easily write the joint posterior of \( \Lambda \) given the data:

\[
p(\Lambda|D) = p(\vec{\lambda}N^*)|D) = \frac{p(D|\vec{\lambda}N^*)}{p(D|\mathcal{H}_\Lambda)} \pi(N^*) \pi(\vec{\lambda})
\]

where we have used Bayes theorem, and used the symbol \( \pi \) for distributions that don’t depend on the data. We have also written

\[
\pi(\vec{\lambda}N^*) = \pi(\vec{\lambda}|N^*) \pi(N^*) = \pi(\vec{\lambda}) \pi(N^*)
\]

since \( N^* \) and \( \vec{\lambda} \) are independent variables by construction in our analysis. When writing the evidence of the data, \( p(D|\mathcal{H}_\Lambda) \), we have restored the explicit dependence on all settings and choices made in the hyper analysis, \( \mathcal{H}_\Lambda \), since soon we will need to deal with another model. In this context, one often call \( p(D|\mathcal{H}_\Lambda) \) the hyper evidence, a number that properly normalizes the hyper posterior:

\[
p(D|\mathcal{H}_\Lambda) = \int d\Lambda p(D|\Lambda) \pi(\Lambda) = \int d\vec{\lambda} dN^* p(D|\vec{\lambda}N^*) \pi(N^*) \pi(\vec{\lambda})
\]

(you can check that if you integrate Eq. (91) over \( \Lambda \), and use this definition of hyper evidence, you get 1, as it should be the case for a probability density.)

We thus finally have the fundamental equation:

\[
p(\vec{\lambda}N^*)|D) = \frac{\pi(N^*) \pi(\vec{\lambda})}{p(D|\mathcal{H}_\Lambda)} e^{-N^{*\tau}} (N^*)^{N^{\text{tr}}} \alpha(\vec{\lambda})^{-N^{\text{tr}}} \prod_{i=1}^{N^*} p(d_i|\vec{\lambda})
\]

\[\text{10} \mathcal{H}_\Lambda \text{ should be on the right side of the conditional bar of all terms really, but the explicit presence of } N^*, \vec{\lambda} \text{ or } \Lambda \text{ in the other terms makes it clear which model is being used} \]
This equation depends on the number of detected triggers \( N^{tr} \) over the experiment time and the number of sources that were expected above the detection threshold over the experiment time \( N^s \). Notice that these two numbers are the same due to the Poissonian nature of the problem (If in doubt, go back to Section 4!). Selection effects are contained explicitly in \( \alpha(\bar{\lambda}) \) and implicitly in \( N^s \). The individual events likelihood \( p(d_i|\bar{\lambda}) \) will be further discussed below. The hyper evidence, \( p(D|H_\Lambda) \), could be dropped while doing parameter estimation, since it does not depend on the hyper parameters. It cannot be dropped if one wants to do model selection, i.e. compare how well two different models for the populations (each with its number of parameters and population priors) describe the data \([11, 22, 23, 24, 25, 26, 27]\).

Now, suppose one does not actually care about the value of the overall scale \( N^s \), but only about the shape parameters \( \bar{\lambda} \). The posterior for those given the data can be obtained from Eq. (93) by marginalizing over \( N^s \):

\[
p(\bar{\lambda}|D) = \frac{\pi(\bar{\lambda})}{p(D|H_\Lambda)} \int dN^s \left[ e^{-N^s} (N^s)^{N^{tr}} \alpha(\bar{\lambda})^{-N^{tr}} p(D|\bar{\lambda}) \right] \pi(N^s) \tag{94}
\]

Notice that the expected number of detectable sources \( N^s \) depends on the expected total number of sources \( N^s \), Eq. [15], which we can use to write:

\[
p(\bar{\lambda}|D) = \frac{\pi(\bar{\lambda})}{p(D|H_\Lambda)} \int dN^s \left[ e^{-N^s} \alpha(\bar{\lambda}) \left( N^s \alpha(\bar{\lambda}) \right)^{N^{tr}} \alpha(\bar{\lambda})^{-N^{tr}} p(D|\bar{\lambda}) \right] \pi(N^s) \tag{95}
\]

\[
= \frac{\pi(\bar{\lambda})}{p(D|H_\Lambda)} p(D|\bar{\lambda}) \int dN^s e^{-N^s} \alpha(\bar{\lambda}) \left( N^s \right)^{N^{tr}} \pi(N^s) \tag{96}
\]

The marginalization can be carried out analytically in the simple case where the prior on \( N^s \) is a Jeffrey prior \( \pi(N^s) = (N^s)^{-1} \); then:

\[
\int dN^s e^{-N^s} \alpha(\bar{\lambda}) N^s N^{tr} \pi(N^s) = \int dN^s e^{-N^s} \alpha(\bar{\lambda}) \left( \frac{N^s}{N^{tr}} \right)^{N^{tr}} \pi(N^s) \tag{97}
\]

\[
\int dN^s e^{-N^s} \alpha(\bar{\lambda}) N^s N^{tr} - 1 = \left( y \equiv N^s \alpha \right) = \int \frac{dy}{\alpha(\bar{\lambda})} e^{-y} \left[ \frac{y}{\alpha(\bar{\lambda})} \right]^{N^{tr}-1} \tag{98}
\]

\[
= \alpha(\bar{\lambda})^{-N^{tr}} \int dy e^{-y} y^{N^{tr}-1} = \alpha(\bar{\lambda})^{-N^{tr}} \Gamma(N^{tr}-1) \tag{99}
\]

Where \( \Gamma \) is the gamma function. Plugging this back we have:

\[
p(\bar{\lambda}|D) = \frac{\pi(\bar{\lambda})}{p(D|H_\Lambda)} p(D|\bar{\lambda}) \alpha(\bar{\lambda})^{-N^{tr}} \Gamma(N^{tr}-1) \tag{100}
\]

Remembering that \( p(D|\bar{\lambda}) \) is the product of \( N^{tr} \) terms we have:

\[
p(\bar{\lambda}|D) = \frac{\pi(\bar{\lambda})}{p(D|H_\Lambda)} \Gamma(N^{tr}-1) \prod_{i=1}^{N^{tr}} \frac{p(d_i|\bar{\lambda})}{\alpha(\bar{\lambda})} \tag{101}
\]

\footnote{Remember we already assumed that no background happens, thus \( N^{tr} \) is actually a source count.
To calculate this expression, we need for each source the likelihood of the data given \( \vec{\lambda} \), 
\( p(d_i|\vec{\lambda}) \). This can be written in a more useful form by using the sum rule to marginalize over all possible values of the source parameters:

\[
p(\vec{\lambda}|D) = \frac{\pi(\vec{\lambda})}{p(D|H_\Lambda)} \Gamma(N^{tr} - 1) \prod_{i=1}^{N^{tr}} \int d\vec{\theta} p(d_i|\vec{\theta}\vec{\lambda}) \pi(\vec{\theta}|\vec{\lambda}) \alpha(\vec{\lambda}) \tag{102}
\]

\[
P(\vec{d}|H_{PE}) = \int d\vec{\theta} p(\vec{d}|\vec{\theta}) \pi(\vec{\theta}|H_{PE}) \tag{103}
\]

In the last step we have eliminated \( \vec{\lambda} \) from the right side of the conditional bar 
\( p(d_i|\vec{\theta}\vec{\lambda}) \) since

the likelihood only depends on the source parameters, and not on the population hyper parameters. Also notice how the selection effects are now entirely contained in \( \alpha(\vec{\lambda}) \) (cfr Eq. (20) of [28]).

How do we calculate \( p(d_i|\vec{\theta}) \)?

The most common scenario is that we have a set of posterior samples 
\( p(\vec{\theta}|d_i) \), produced by either ourselves, a colleague, or a scientific collaboration, using a Markov Chain Monte Carlo (MCMC) or similar algorithm, since in general these are high-dimensional non-trivial distributions to sample [29, 30]. If the sampling prior used to generate those samples is \( \pi(\vec{\theta}|H_{PE}) \) then posterior and likelihood are related as:

\[
p(d|\vec{\theta}) = p(\vec{\theta}|d, H_{PE}) \frac{p(d|H_{PE})}{\pi(\vec{\theta}|H_{PE})} \tag{104}
\]

where I have used \( H_{PE} \) (“PE” stands for Parameter Estimation) for the modeling used to characterize the individual events and \( p(d|H_{PE}) \) is the evidence of the data obtained with the sampling algorithm:

\[
p(d|H_{PE}) = \int d\vec{\theta} p(d|\vec{\theta}) \pi(\vec{\theta}|H_{PE}) \tag{105}
\]

Please notice that \( H_{PE} \) and \( H_\Lambda \) imply different prior beliefs in the distribution of \( \vec{\theta} \): the former model assumes that the parameters \( \vec{\theta} \) are distributed as \( \pi(\vec{\theta}|H_{PE}) \), i.e. the sampling prior, while the latter assumes \( p(\vec{\theta}|\vec{\lambda}) \), i.e. the population prior. This means that in general the hyper evidence, Eq. (92) is not equal to the evidence calculated by the sampling algorithm:

\[
p(D|H_\Lambda) = \prod_{i=1}^{N^{tr}} p(d_i|H_\Lambda) \neq \prod_{i=1}^{N^{tr}} p(d_i|H_{PE}) = p(D|H_{PE}) \tag{106}
\]

A last point worth stressing before we continue. **If** the sampling prior \( \pi(\vec{\theta}|H_{PE}) \) is uniform in **all** parameters then:

\[
\pi(\vec{\theta}|H_{PE}) = c \implies p(d|\vec{\theta}) \propto p(\vec{\theta}|d) \tag{107}
\]

and one could use the draws from the posterior as if they were draws from the likelihood function. Unfortunately, in general, at least some of the sampling priors are non-trivial, and must be taken into account when going from posterior to likelihood, as we saw in Eq. (104) (one often says that they need to “remove” the prior).
With this in mind, let’s use Bayes theorem inside the product, \(103\):

\[
p(\tilde{\lambda}|D) = \frac{\pi(\tilde{\lambda})}{p(D|\mathcal{H}_\Lambda)} \Gamma(N^{tr} - 1) \prod_{i=1}^{N^{tr}} \frac{1}{\alpha(\lambda)} \int d\tilde{\theta} p(d_i|\tilde{\theta}) \pi(\tilde{\theta}|\tilde{\lambda})
\]

\[
\Rightarrow \frac{\pi(\tilde{\lambda})}{p(D|\mathcal{H}_\Lambda)} \Gamma(N^{tr} - 1) \prod_{i=1}^{N^{tr}} \frac{1}{\alpha(\lambda)} \int d\tilde{\theta} \frac{p(\tilde{\theta}|d_i)p(d_i|\mathcal{H}_{PE})\pi(\tilde{\theta}|\tilde{\lambda})}{\pi(\tilde{\theta}|\mathcal{H}_{PE})}
\]

\[
\Rightarrow \pi(\tilde{\lambda}) \Gamma(N^{tr} - 1) \frac{p(D|\mathcal{H}_{PE})}{p(D|\mathcal{H}_\Lambda)} \prod_{i=1}^{N^{tr}} \frac{1}{\alpha(\lambda)} \int d\tilde{\theta} \frac{p(\tilde{\theta}|d_i)\pi(\tilde{\theta}|\tilde{\lambda})}{\pi(\tilde{\theta}|\mathcal{H}_{PE})}
\]

where in the first equation we have used the fact that the single event likelihood does not logically depend on \(\tilde{\lambda}\). In the second row \(\pi(\tilde{\theta}|\mathcal{H}_{PE})\) is the sampling prior for \(\tilde{\theta}\) used when drawing posterior samples (using e.g. an MCMC sampler, as mentioned above). The ratio of the MCMC evidence over the hyper evidence is ugly to see, but luckily is not really important as long as we are only trying to measure the population hyper parameters of one model \(\mathcal{H}_\Lambda\) and not comparing different models, since this ratio does not depend on the actual value of the hyper parameters. This is why it is only seldom explicitly reported in the literature.

The integral in Eq. \(110\) can be simplified by noticing that it represents the expectation of the ratio of the population prior over the sampling prior. The exact expectation can be replaced with the approximate sample mean \(31\), i.e. with a discrete sum:

\[
\int d\tilde{\theta} \frac{p(\tilde{\theta}|\mathcal{H}_{PE})}{\pi(\tilde{\theta}|\mathcal{H}_{PE})} \approx \frac{1}{N_{\text{samples}}} \sum_{j=0}^{N_{\text{samples}}} f(\bar{\theta}^j) |_{\bar{\theta}^j \sim p(\tilde{\theta}|\mathcal{H}_{PE})}
\]

which is valid for any function \(f(\tilde{\theta})\) (the symbol \(\sim\) in this context means “drawn from the distribution”), as long as the number of available posterior samples \(N_{\text{samples}}\) is large enough (the larger \(N_{\text{samples}}\), the better the approximation).

Therefore, assuming we are given \(N_{\text{samples}}\) posterior samples from the posterior distribution of \(\tilde{\theta}\) for each of the \(N^{tr}\) sources, then we can write (cfr Eq. (14) of \(3\)):

\[
\prod_{i=1}^{N^{tr}} \frac{1}{\alpha(\lambda)} \int d\tilde{\theta} p(\tilde{\theta}|d_i) \frac{\pi(\tilde{\theta}|\tilde{\lambda})}{\pi(\tilde{\theta}|\mathcal{H}_{PE})} = \prod_{i=1}^{N^{tr}} \frac{1}{N_{\text{samples}} \alpha(\lambda)} \sum_{j=0}^{N_{\text{samples}}} \left[ \frac{\pi(\tilde{\theta}|\tilde{\lambda})}{\pi(\tilde{\theta}|\mathcal{H}_{PE})} \right] |_{\bar{\theta}^j \sim p(\tilde{\theta}|d_i,\mathcal{H}_{PE})}
\]

where \(\bar{\theta}^j\) is the \(j\)-th posterior sample from the posterior distribution of the \(i\)-th source. Since we are calculating an average, one often uses angle brackets:

\[
\frac{1}{N_{\text{samples}}} \sum_{j=0}^{N_{\text{samples}}} \left[ \frac{\pi(\tilde{\theta}|\tilde{\lambda})}{\pi(\tilde{\theta}|\mathcal{H}_{PE})} \right] |_{\bar{\theta}^j \sim p(\tilde{\theta}|d_i,\mathcal{H}_{PE})} = \left\langle \frac{\pi(\tilde{\theta}|\tilde{\lambda})}{\pi(\tilde{\theta}|\mathcal{H}_{PE})} \right\rangle |_{\bar{\theta} \sim p(\tilde{\theta}|d_i,\mathcal{H}_{PE})}
\]

So finally:

\[
p(\tilde{\lambda}|D) = \pi(\tilde{\lambda}) \Gamma(N^{tr} - 1) \alpha^{-N^{tr}} \frac{p(D|\mathcal{H}_{PE})}{p(D|\mathcal{H}_\Lambda)} \prod_{i=1}^{N^{tr}} \left\langle \frac{\pi(\tilde{\theta}|\tilde{\lambda})}{\pi(\tilde{\theta}|\mathcal{H}_{PE})} \right\rangle |_{\bar{\theta} \sim p(\tilde{\theta}|d_i,\mathcal{H}_{PE})}
\]
It is worth stressing and remembering that the selection effects are here included in the detectable fraction \( \alpha(\bar{\lambda}) \), and that this term comes from marginalizing over the total number of sources, not from the individual events likelihood.

### 5.1 What we don’t say

Suppose that the detected signals can be characterized by two unknown parameters \( \tilde{\theta} = (x, y) \), but we really only care about \( x \) because it is the parameter that will tell us about the underlying population parameters \( \bar{\lambda} \). In other words, the distribution of the \( x \) parameters of the population of events does depend on \( \bar{\lambda} \), \( \pi(x|\bar{\lambda}, \mathcal{H}_\bar{\lambda}) \), but we believe the distribution of \( y \) is fixed and known: \( \pi(y|\mathcal{H}_\bar{\lambda}) \). Or, more concretely, we are using someone else’s public data release, but they only gave us some of the marginalized \( x \) posteriors, not the whole cloud of points.

Can we calculate the \( \bar{\lambda} \) posterior by only integrating over \( x \), and neglecting \( y \)? Looking at Eq. (101) one might be tempted to use the sum rule and write:

\[
p(d_i|\bar{\lambda}) = \int dx \ p(d_i, x|\bar{\lambda}) = \int dx \ p(d_i, x) \pi(x|\bar{\lambda})
\]

\[
= \int dx \ p(x|d_i) p(d_i) \frac{\pi(x|\bar{\lambda}, \mathcal{H}_\bar{\lambda})}{\pi(x|\mathcal{H}_{PE})} = p(d_i) \left\langle \frac{\pi(x|\bar{\lambda}, \mathcal{H}_\bar{\lambda}) \pi(y|\mathcal{H}_\bar{\lambda})}{\pi(x|\mathcal{H}_{PE})} \right\rangle_{x \sim p(x|d_i)} \tag{115}
\]

where \( \pi(x|\mathcal{H}_{PE}) \) is the sampling prior on \( x \) used when sampling the posterior distribution. Now, this seems a perfectly legitimate equation; but one could also integrate on both \( x \) and \( y \) and obtain:

\[
p(d_i|\bar{\lambda}) = \int dx \ dy \ p(d_i, x, y|\bar{\lambda}) = p(d_i) \left\langle \frac{\pi(x|\bar{\lambda}, \mathcal{H}_\bar{\lambda}) \pi(y|\mathcal{H}_\bar{\lambda})}{\pi(x,y|\mathcal{H}_{PE})} \right\rangle_{x, y \sim p(x,y|d_i)} \tag{116}
\]

Let’s assume for simplicity that the sampling prior factorizes, \( \pi(x,y|\mathcal{H}_{PE}) = \pi(x|\mathcal{H}_{PE}) \pi(y|\mathcal{H}_{PE}) \):

\[
p(d_i|\bar{\lambda}) = p(d_i) \left\langle \frac{\pi(x|\bar{\lambda}, \mathcal{H}_\bar{\lambda}) \pi(y|\mathcal{H}_\bar{\lambda})}{\pi(x|\mathcal{H}_{PE}) \pi(y|\mathcal{H}_{PE})} \right\rangle_{x, y \sim p(x,y|d_i)} \tag{117}
\]

Clearly, the two expressions in the last term of Eq. (115) and Eq. (117) are in general different, because the values of \( x \) used in the two cases are the same, no matter of whether we use the samples from the 2D \( x \sim p(x,y|d_i) \) or from the marginalized distribution \( x \sim p(x|d_i) \), but now there is an extra term \( \pi(y|\mathcal{H}_{\bar{\lambda}})/\pi(y|\mathcal{H}_{PE}) \) which need not be equal to 1.

But the left hand sides look the same, so what’s going on? The problem is that we have not explicitly written down all of the assumptions on the right side of the conditional bars.

What we want for the hierarchical analysis, Eq. (101) is really the probability of the data given \( \bar{\lambda} \) and our model \( \mathcal{H}_\bar{\lambda} \). The model comes with set of priors on the distribution of population parameters, all parameters:

\[
p(d_i|\bar{\lambda}) = p(d_i|\bar{\lambda}_\mathcal{H}_\bar{\lambda}) = p(d_i|\bar{\lambda} \land \pi(\tilde{\theta}|\lambda)) = p(d_i|\bar{\lambda} \land \pi(x|\bar{\lambda}, \mathcal{H}_\bar{\lambda}) \land \pi(y|\mathcal{H}_\bar{\lambda})) \tag{118}
\]

In fact, this is the very reason why we have to “divide out” the sampling priors in Eq. (117): we want to calculate this likelihood under the hypothesis that the prior being used is the population one \( \pi(\tilde{\theta}|\bar{\lambda}, \mathcal{H}_\bar{\lambda}) \), but we have been handled samples that were produced using the prior \( \pi(\tilde{\theta}|\mathcal{H}_{PE}) \).

Instead, in Eq. (115) we calculated:
\[ p(d_i|\lambda) = p(d_i|\lambda \wedge \pi(x|\lambda H_{\lambda})) \]  \hspace{1cm} (119)

Because nothing is said about the prior on the other variable, this is implicitly assuming that the prior on \( y \) is still the one used while sampling the parameter space, i.e. \( \pi(y|H_{PE}) \):

Eq. (115) \leftrightarrow p(d_i|\lambda \wedge \pi(x|\lambda, H_{\lambda}) \wedge \pi(y|H_{PE})) \hspace{1cm} (120)

This is why the right hand sides of Eq.(115) and Eq. (117) look like they are going to give two different results: in general, they will!

The only parameters that we could not explicitly account for while calculating the posterior of the population hyperparameters are those for which the population prior is exactly equal to the sampling prior. This can be seen directly comparing the two equations above: both the LHSs (Eq. (120) and Eq. (118)) and the RHSs, Eq.(115) and Eq. (117) become identical if \( \pi(y|H_{\lambda}) = \pi(y|H_{PE}) \). But notice that this still implies that the relevant prior distributions are separable, \( \pi(x|H_{PE})\pi(y|H_{PE}) \), which might not always be true! If the priors are more complicated, and correlations exist, it is even hard to simplify things so that only some of the prior distributions survive. Therefore Eq. (116) is always what we want to do, while Eq. (115) might be sometimes what we want to do, if we are very careful and lucky.

One might still wonder: why care if a constant (i.e. not depending on \( \lambda \), which is the quantity we are interested in) multiplicative term is missing? Can’t this be accounted for later, while enforcing the proper normalization of the \( \lambda \) posterior? The answer is no, it would be a problem. This can be seen best by looking at the RHS of Eq. (112), which in our 2D example reads:

\[ \prod_{i=1}^{N_{tr}} \left( \frac{1}{\frac{N_{samples}}{N_{samples}}} \sum_{j=0}^{N_{samples}} \frac{\pi_i(x^j|H_{H_{\lambda}})\pi(y^j|H_{\lambda})}{\pi(x^j|H_{PE})\pi(y^j|H_{PE})} \right)_{x^j, y^j \sim p(x,y|d_i)} \]  \hspace{1cm} (121)

(where I’m neglecting the \( \alpha \) term, which is not important for this discussion, and assuming all priors factorize.)

If \( \pi(y^j|H_{\lambda}) \neq \pi(y^j|H_{PE}) \) in general the ratio \( \pi(y^j|H_{\lambda})/\pi(y^j|H_{PE}) \) will be different from 1, and it will be for different samples and different events. Thus if we entirely neglected \( y \), we would be adding a different, and wrong, numerical weighting coefficient for each sample in these sums.

What if someone is only handled marginalized posteriors? For example, the LIGO [32]-Virgo [33] collaboration (LVC) currently only releases in low latency a estimate of the sky position and distance, but not of the masses, of the candidate compact binary coalescence (CBC) sources it detects [34]. In absence of this information one cannot correctly remove the mass prior, hence their inference on any hyper parameters, even if not directly linked to the masses, would be wrong unless their population prior for the masses were identical to what the LVC used to produce samples (and assuming prior factorizes). If that is the case, then masses can be entirely dropped out while calculating the likelihood term above, but, again, care must be taken to be sure that the very same prior is used as the population prior when calculating selection effects, i.e. \( \alpha(\lambda) \), or else biases will be introduced. A common mistake is to use posterior samples which have already been marginalized over some parameter that is not of interest (e.g. \( y \) in the example above), hence implicitly using a prior for \( y \) equal to the sampling prior, but then use a different prior when calculating \( \alpha(\lambda) \).

\[ \text{Notice that the evidences on the right hand side of Eq. (115) and Eq. (117) are also in general different, for the same reason.} \]
6 A simple 1-dimensional example

Let’s consider a simple but instructive 1D example, that shows how many of these pieces come to play, in practice.

The problem is as follows. There is an infinity of audio speakers arranged along the x-axis. An audio detector is placed at some distance, Fig. 1 and when one speaker at position \( x \) emits a sound \( s \), the detector records:

\[
d = x + n
\]  

where \( x \) is the true position of the speaker that emitted a signal, and \( n \sim \mathcal{N}(0, 1) \) is a random number drawn from a normal distribution with mean equal to 0 and standard deviation equal to 1 meter. Here we are just assuming, arbitrarily, that the noise associated with recording the data has this typical width of 1 meter. We can think of this noise term as due to e.g. winds that affect the perceived direction of the incoming audio wave.

More realistic cases will differ from this in a few ways: a) the noise properties (e.g. its sigma) might be time dependent b) it might be different for different events c) the data might not directly contain the variable of interest, but rather a function of it (i.e. \( d = n + h(x) \), with \( h \) some potentially complicated function, see Section 7) and d) there might be more than one parameter that affects the detected data, all or some of which might be of interest to us, and all or some of which might depend on a higher hyper-level.

However, because of a nearby mountain range, there exists a maximum value of \( x \), \( x_{\text{max}} \), such that detectors placed at \( x > x_{\text{max}} \) would not produce a measurable sound.

However, it is important to remember that we do not know \( x \), the best we can do is to use the data to set a threshold. Since the noise in our example is a zero-mean process it’s reasonable to use as a detection criteria:

\[
d \leq x_{\text{max}} \leftrightarrow \text{detection}
\]  

But we should not forget that there is noise! Thus even a source with \( x_{\text{true}} = x_{\text{max}} \) will produce a detectable signal 50% of the time (50% since the noise is symmetric). Let’s pause for a second and stress a couple of important points here

- Barring rare exceptions, the selection criteria is based on the data, not on the true value of the parameter. In general we do not know the parameters! The data is all we have. The data are our friend.
- Because of the noise, sources with a true value of the parameter that fail the threshold for detection can be promoted to detections. Conversely, sources that would meet the criteria, can fail it because of the noise.

This also means that unless there is absolutely no noise (i.e. the noise is so small that we are de facto measuring the true value of the parameters), the selection effect will not be sharp feature, a Heaviside function in our simple 1D case, but rather a sigmoid that smoothly turns off.

Let’s see this explicitly. First, given that the noise is gaussian, the likelihood for the data \( d \) given \( x \) is:

\[
p(d|\theta) = \frac{1}{\sqrt{2\pi}} \exp \left[ -\frac{1}{2} (d - x)^2 \right]
\]  

I am writing these notes in a period of my life where I discovered the joys of Arduino, as you can clearly see.
Figure 1: A sketch of the problem. The red bell represents the density of speakers that are playing (my limited artistic skills prevent me from making a drawing where the speakers are actually drawn more densely in that region).

(here there is, implicitly, also the sigma of the noise, which is equal to 1 meter). Thus in our scenario the likelihood for the event is a gaussian, seen as a function of $x$, centered at $d$.

We had previously defined the probability of a source being detectable given the true value of the parameter, Eq. (14):

$$p(\text{det}|x) = \int_{D_\gamma} d d p(d|x) \quad (125)$$

now, this integral is made over the detectable data, which are the data for which $d < x_{\text{max}}$, that is:

$$p(\text{det}|x) = \int_{D_\gamma} d d p(d|x) = \int_{-\infty}^{x_{\text{max}}} dd \frac{1}{\sqrt{2\pi}} \exp \left[-\frac{1}{2}(d-x)^2\right] \quad (126)$$

But this is just the cumulative distribution of a normal distribution with mean $x$, and can be written in terms of the error function $\text{erf}$:

$$p(\text{det}|x) = \frac{1}{2} \left[ 1 + \text{erf} \left( \frac{x_{\text{max}} - x}{\sqrt{2}} \right) \right] \quad (127)$$

We see that $p(\text{det}|x_{\text{max}}) = 0.5$, as mentioned above: half of the sources with true value of $x$ at threshold are detectable because of the noise. The overall curve is indeed a sigmoid, Figure 2.

Let’s pause here for a second to be sure something is clear: we said earlier that the criteria we use to decide if a piece of data contains a trigger must only depend on the data, not on true true value of the parameter, Eq. (123). However, the probability that any given value of $x$ results in a detection does depend on $x$, Eq. (125). There is no contradiction: in Eq. (125) $x$ is on the right hand side of the conditional bar, and hence it’s assumed known. Whether and when that particular $x$ result in a detection is still determined by the data: given $x$ one can calculate all possible datas that
Figure 2: The probability $p(\text{det}|x)$ of detecting the sound from a speaker at true position $x$. Here $x_{\text{max}} = 5$

can be contain it, i.e. all possible noise realizations, and adds up those that result in a detectable piece of data.

Let’s now consider inference on the property of the speakers, given a set of detections. We will assume that a friend of ours has setup the system such that over a long period of time\textsuperscript{14} a set of speakers will play, one at the time. The positions of the speakers that will play are scattered around some central value $\mu_T$, and are draws from a normal distribution:

$$ \pi(x|\mu, \sigma) = N(x - \mu, \sigma). $$

Our friend has told us the value of $\sigma$, which we can thus treat as known, but has challenged us to measure the true value of $\mu_T$, given the data recorded by the detector.

We have all the tools we need to meet the challenge. The mathematics has been laid out in the previous sections, and can be directly reused with the understanding that our only shape hyper-parameter\textsuperscript{15} is:

$$ \tilde{\lambda} = \{ \mu_T \} $$

The posterior on $\mu_T$ is given by Eq. (103). We first need to calculate the detection efficiency $\alpha(\tilde{\lambda})$:

$$ \alpha(\mu_T) = \int dx \pi(x|\mu_T)p(\text{det}|x) $$

which is shown in Figure 3 for $\mu_T = 4.3$, $\sigma = 2$ and $x_{\text{max}} = 5$.

With this in hand, let’s consider the pieces of Eq. (103) that depend on $\mu_T$, and neglect everything else since it will only act as a constant term with respect to $\mu_T$, that we can fix by normalizing the posterior at the end. We have:

\textsuperscript{14}Long enough that the probability that two speakers play in the same small data chunk is negligible and we can use the math we developed so far.

\textsuperscript{15}We are not trying to infer the total number of detectors that played.
Figure 3: The fraction of detectable events $\alpha(\mu_\lambda)$ as a function of the $\mu_\lambda$. Here $\sigma_\lambda = 2$ and $x_{\text{max}} = 5$

$$p(\mu_\lambda|D) \propto \pi(\mu_\lambda) \prod_{i=1}^{N} \frac{\int dx \ d_i(x) \pi_\lambda(x|\mu_\lambda)}{\alpha(\mu_\lambda)}$$  \hspace{1cm} (128)

Because both the individual event likelihood and the population prior are Gaussians, we can actually solve the numerator analytically (yeah!)

$$\int dx \ d_i(x) \pi_\lambda(x|\mu_\lambda) = \frac{1}{2\pi\sigma_\lambda} \int dx \ e^{-\frac{1}{2}(x-\mu_\lambda)^2} e^{-\frac{1}{2\sigma_\lambda^2}(x-\mu_\lambda)^2}$$  \hspace{1cm} (129)

Before we continue, let me stress again that in here the integration range is the full and unadulterated range that $x$ can theoretically take, including values which are above our threshold $x_{\text{max}}$. As seen before, the selection effects do not enter the numerator of the hyper posterior, but only through $\alpha(\lambda)$.

In this case, this is of great help because it allows us to get rid of the integral using standard Gaussian integrals. In fact if

$$f(x) = \frac{1}{\sqrt{2\pi\sigma_f}} e^{-\frac{(x-\mu_f)^2}{2\sigma_f^2}}$$

and

$$g(x) = \frac{1}{\sqrt{2\pi\sigma_g}} e^{-\frac{(x-\mu_g)^2}{2\sigma_g^2}}$$

Then:

$$f(x)g(x) = \frac{A}{\sqrt{2\pi\sigma_{fg}}} \exp \left[ -\frac{(x-\mu_{fg})^2}{2\sigma_{fg}^2} \right]$$

where
\[ \mu_{fg} = \frac{\mu_f \sigma_g^2 + \mu_g \sigma_f^2}{\sigma_f^2 + \sigma_g^2} \]

and

\[ \sigma_{fg} = \sqrt{\frac{\sigma_f^2 \sigma_g^2}{\sigma_f^2 + \sigma_g^2}} \]

with amplitude

\[ A = \frac{1}{\sqrt{2\pi (\sigma_f^2 + \sigma_g^2)}} \exp \left[ -\frac{(\mu_f - \mu_g)^2}{2 \left( \sigma_f^2 + \sigma_g^2 \right)} \right] \]

In our case:

\[ \frac{1}{2\pi \sigma_\lambda} \int dx \ e^{-\frac{1}{2} (d_i - x)^2} e^{-\frac{1}{2\sigma_\lambda^2} (x - \mu_\lambda)^2} = \frac{1}{\sqrt{2\pi \left( 1 + \sigma_\lambda^2 \right)}} \exp \left[ -\frac{(d_i - \mu_\lambda)^2}{2 \left( 1 + \sigma_\lambda^2 \right)} \right] \] (130)

where we have used

\[ \frac{1}{\sqrt{2\pi \sigma_{fg}}} \int dx \ \exp \left[ -\frac{(x - \mu_{fg})^2}{2\sigma_{fg}^2} \right] = 1 \]

We thus see that each detection contributes to the overall posterior for \( \mu_\lambda \) with a term proportional to:

\[ \exp \left[ -\frac{(d_i - \mu_\lambda)^2}{2 \left( 1 + \sigma_\lambda^2 \right)} \right] \] (131)

and the overall posterior for \( \mu_\lambda \) is:

\[ p(\mu_\lambda | D) \propto \pi(\mu_\lambda) \prod_{i=1}^{N_{\text{tr}}} \exp \left[ -\frac{(d_i - \mu_\lambda)^2}{2 \left( 1 + \sigma_\lambda^2 \right)} \right] \frac{1}{\alpha(\mu_\lambda)} \] (132)

The width of the \( \mu_\lambda \) posterior is clearly determined by how many detections we have, with more events resulting in a sharper peak. Similarly, smaller \( \sigma_\lambda \) help, since a) the smaller \( \sigma_\lambda \) the more the typical data will be close to \( \mu_\lambda^2 \), making more events contribute useful information b) the denominator in the single-event exponent will be smaller, meaning that the tails of the single-event likelihoods will contribute less and less.

[If we had not fixed the noise sigma of each event to be the same and equal to 1 meter, a smaller sigma noise would also make the inference better, for the same identical reasons].

We notice that the term \( 1/\alpha(\mu_\lambda) \) gives a boost to \( \mu_\lambda \)'s which result in more hard-to-detect events. Of course, the price to pay for making the denominator smaller is that the numerator will also decrease if we are considering \( \mu_\lambda \)'s that are far from our data. The \( \mu_\lambda \) posterior is thus determined by the delicate balance between these two factors.

In this simple example we have actually integrated the likelihood analytically, to obtain Eq. (132).

A possible implementation of this algorithm would thus be:
• Choose $\mu^T, \sigma, x_{max}$.

• Generate $N$ random numbers $\vec{x}_T \sim N(\mu^T, \sigma)$, those are the true positions of the speakers that will be playing, and we do not have access to those directly.

• Generate $N$ noisy data sets by adding to each $x$ in $\vec{x}_T$ a random draw from a normal distribution: $\vec{\xi} = N(\vec{x}_T, 1m)$ with unit standard deviation.

• Only retain the data that satisfy the detection criteria:

$$\forall d_i \in \vec{\xi}, d_i \in D \text{ if } d_i < x_{max}$$

$D$ is our catalog of $N^{tr}$ detected sources that will be used for the inference.

• Grid the $\mu$ axis with enough points, and at each point calculate $\alpha(\mu)$ and Eq. (131) for each event.

• Pick a prior for $\mu$, flat is just fine, and calculate Eq. (132) [Note: because we marginalized over $x$ analytically, we don’t actually have to create synthetic posteriors for the individual events - only their measured data $d_i$ matters, Eq. (132)].

The reason why we managed to proceed so far using analytical calculations is because the problem in hand is quite simple. In general the problem will be too complicated for the math to carried out analytically, due to higher dimensionality, correlations, non-trivial priors, etc. In that case, a good starting point is Eq. (128):

$$p(\mu|D) \propto \pi(\mu) \prod_{i=1}^{N^{tr}} \frac{\int dx \ p(d_i|x)\pi(x|\mu)}{\alpha(\mu)} \tag{133}$$

$$\propto \pi(\mu) \prod_{i=1}^{N^{tr}} \frac{\int dx \ p(x|d_i)\pi(x|\mu)}{\alpha(\mu)} \tag{134}$$

$$\propto \pi(\mu) \prod_{i=1}^{N^{tr}} \frac{1}{\alpha(\mu)} \langle \frac{\pi(x|\mu)}{\pi(x|H_{PE})} \rangle_{x \sim p(x|d_i)} \tag{135}$$

where in the last line we are averaging the ratio of the population prior over the sampling prior (i.e. the prior used to produce the posterior samples $p(x|d_i)$) calculated at values of $x$ fairly drawn from the posterior samples.

Assuming that for each source a set of $N_{samples}$ samples have been drawn from the posterior of each trigger then we can write this as:

$$p(\mu|D) \propto \frac{\pi(\mu)}{\alpha(\mu)^{N^{tr}}} \prod_{i=1}^{N^{tr}} \frac{1}{N_{samples}} \sum_{j=1}^{N_{samples}} \frac{\pi(x|\mu)}{\pi(x|H_{PE})} \bigg|_{x \sim p(x|d_i)} \tag{136}$$

If we want to do this with our problem, and generate samples instead of using the analytic integration over $x$, we would do (see Appendix A of [35] for a more complicated problem):

• Choose $\mu^T, \sigma, x_{max}$.\]
• Generate $N$ random numbers $\mathbf{x}_T \sim N(\mu_T, \sigma_T)$, those are the true positions of the speakers that will be playing, and we do not have access to those directly.

• Generate $N$ noisy data sets by adding to each $x$ in $\mathbf{x}_T$ a random drawing from a normal distribution: $\mathbf{\tilde{x}} = N(\mathbf{x}_T, 1m)$

• Only retain the data that satisfy the detection criteria:

$$\forall d_i \in \mathbf{\tilde{x}}, d_i \in D \iff d_i < x_{\text{max}}$$

$D$ is our catalog of $N^{\text{tr}}$ detected sources that will be used for the inference.

• For each $d_i \in D$ draw $N_{\text{samples}}$ samples from the likelihood distribution Eq. (124):

$$\{d_1^i, d_2^i \ldots d_{N_{\text{samples}}}^i\} \sim N(d_i; 1m)$$

• If a flat sampling prior on $x$ is being assumed, then these are also samples from the posterior of $x$, and can be used directly in Eq. (136). Otherwise something like importance resampling must be used.

• Pick a prior for $\mu, \lambda$, flat is just fine, grid the $\mu, \lambda$ axis with enough points, and at each point calculate $\alpha(\mu, \lambda)$ and then Eq. (136).

In the left panel of Figure 4 we show an example of the inference on $\mu_T$ obtained with 150 sources. In this case the true mean of the population prior, $\mu_T$, is smaller than the threshold $x_{\text{max}}$.

The 150 individual likelihoods are shown in yellow (the cut of the x-axis at -2m is just for graphical reasons, and not a cut on the sources or samples used). The thick vertical line is $x_{\text{max}}$. It’s worth stressing again how it is entirely possible and perfectly fine that samples from individual events extend to values larger than $x_{\text{max}}$: all those samples must be used to calculate the numerator of Eq. (136).

![Figure 4](image-url)

Figure 4: (left:) The inference on the central position of the gaussian distribution of speakers that have been turned on, $\mu$. The true value of all relevant parameters is given in the title. The blue dot-dashed and cyan dashed curves are the posteriors obtained either by numerical or analytical average of the likelihood samples. (right:) the same, but with true value of $\mu$ larger than $x_{\text{max}}$. 

31
The black line (right y-panel) is the detection efficiency \( \alpha(\mu_\lambda) \), already shown in Figure 3. Finally, the blue dot-dashed and cyan dashed curves are the posteriors on \( \mu_\lambda \) obtained either with the fully numerical method, Eq. (136), or the analytical marginalization over the samples, Eq. (132), respectively.

It is worth stressing that this inference can be carried on even if most of the speakers are not detectable, right panel of Figure 4. In this case \( \mu_\lambda^T = 8.3 \) and only \( \sim 7\% \) of the speakers produce a detectable sound. Studying the figure it might at first look surprising that the posterior peaks at the true value, in a region where no data is available, if not the tail of the distribution of the detectable events. What is happening is that, as seen above, each time we try a value of \( \mu_\lambda \) we need too divide by \( \alpha(\mu_\lambda) \) to the power \( N^{tr} \). Thus, since \( \alpha(\mu^T_\lambda) \approx 0.07 \) while \( \alpha(x_{max}) = 0.5 \), for example, the product of the tails of the likelihoods at \( \mu^T_\lambda \) is enhanced by a factor of \( (0.5/0.07)^{150} = 1.2 \times 10^{128} \) (!!!) relative to the product of the likelihoods at \( x_{max} \).

The fact that inference can be carried out correctly even if the population is such that most events cannot be detected should not be surprising. In fact, it is quite common. In gravitational-wave astrophysics, the signal-to-noise ratio - which is the detection criterium \(^{16}\) - decreases as \( 1/d_L \), where \( d_L \) is the source luminosity distance. If sources are placed with a distance distribution that goes like \( d_\beta L \), then the true value of \( \beta \) is such that most events are in fact undetectable \(^{17}\).

7 A gravitational-wave example

In this section we are going to look at a more complicated example: population inference on the mass distribution of binary black holes. We will assume that all detected binary black holes emitting GWs are members of the same population, and that their masses are draws from Gaussian distributions of unknown means. While the details of the model are discussed below, let me stress here that we are considering a single population. In general, it might be the case that more than one population exists (e.g. some of the black hole binaries might have formed in galactic fields and others in globular clusters). This can be dealt using a model for each of the populations, and treat the number of sources produced in each as one of the unknowns being measured from the data. \(^{18}\)

If enough sources are detected, and if the various populations have some distinguishing features, one will be able to measure both number of sources produced by each population, and their characteristic parameters \(^{36, 24, 37, 26, 25, 11}\). A multi-population analysis would then proceed along the same lines of the one described below, but it would be computationally much more expensive due to the larger number of parameters.

We will work under the simplifying assumptions that one GW detector has collected data for a total time \( T \), and a search algorithm has been searching for binary black hole coalescences, finding a certain number of triggers. No background exists, and we are not interested on the total number of events, which is marginalized using a Jeffreys prior. Under these assumptions, the posterior for the population parameters is given by Eq. (110).

In our simple model, black holes have no spins, and they inspiral in quasi-circular orbits. If this is the case, each GW depends on 9 unknown parameters: the two masses (\( m_1 \) and \( m_2 \)), the right ascension and declination (RA dec), the orbital orientation (\( \iota \)) and polarization (\( \psi \)), the arrival time

---

\(^{16}\) At least to first order. Things are quite more complicated. More on this below.

\(^{17}\) It is worth checking out the Jupyter notebooks that my friend Will Farr made available at [https://github.com/farr/SelectionExample](https://github.com/farr/SelectionExample) as supplementary material of [3].

\(^{18}\) More often, one works with branching ratios, i.e. the ratio of sources produced in any pair of populations.
and phase \((t_c, \varphi_c)\) and the luminosity distance \(d_L\). Collectively, these 9 parameters are \(\vec{\theta}\).

We will assume that noise is Gaussian, stationary and additive, such that in presence of an astrophysical signal \(h(\vec{\theta})\) the recorded time-domain data is:

\[
d(t) = n(t) + h(t, \vec{\theta})
\]

(137)

the Gaussianity of the noise naturally leads to a Gaussian likelihood for the data:

\[
p(d|\vec{\theta}) \propto e^{-\langle d-h(\vec{\theta})|d-h(\vec{\theta})\rangle}
\]

(138)

where we have defined the inner product in the frequency domain:

\[
\langle a(f)|b(f)\rangle \equiv 2 \int df \frac{a(f) b^*(f) + a^*(f) b(f)}{S(f)}
\]

(139)

and \(S(f)\) is the noise auto-correlation, aka noise power spectral density (PSD) \([40, 41, 42]\): it is small where the detector is sensitive and large where it’s not so sensitive. The integration range should be wide enough to cover the frequency content of the eventual signal, and typically is limited on the low side by the seismic wall of ground-based detectors \((\sim 10 - 20 \text{ Hz})\), and on the high side by the sampling rate.

We need to calculate the fraction of detectable events \(\alpha(\vec{\lambda})\) given some hyper population:

\[
\alpha(\vec{\lambda}) = \int d\vec{\theta} p(\text{det}|\vec{\theta}) \pi(\vec{\theta}|\vec{\lambda})
\]

(140)

Given significant computational resources and patience, one would evaluate this integral by first calculating \(p(\text{det}|\vec{\theta})\) using the actual output of a search algorithm on a large set of simulated signals, and then doing appropriate reweighting according to the population prior \([11]\).

For our simple example, and for a significant fraction of the literature, a simpler approach is enough, which strongly reduces the amount of required computing time. First, we need to decide on a detection criterion which can be easily verified without having to run a full search for compact binaries. We introduce the optimal signal-to-noise ratio (SNR) \(\rho\) of a GW signal with parameters \(\vec{\theta}\) as

\[
\rho^2 \equiv \langle h(f, \vec{\theta})|h(f, \vec{\theta})\rangle
\]

(141)

where \(h(f, \vec{\theta})\) is the Fourier transform of the GW signal, and the inner product is defined above. We can call a signal detectable if it gives an SNR larger than \(\rho_{\text{thr}}\) (usually 8 or so) in a single detector. As mentioned above, reality is more complicated, and in practice a few more factors will contribute to making a signal detected by a real GW search algorithm \([43, 44, 11, 45]\), but this is a good first approximation \([11]\). To explicitly calculate \(\alpha(\vec{\lambda})\), we need to say a bit more about the actual GW signal and how it depends on \(\vec{\theta}\). As mentioned above, we will make the simplifying assumptions that no spins are present (In general this won’t affect the SNRs by more than a few percent \([47]\)). Furthermore, we will keep the limit where only the main multipolar emission mode \(- \ell = m = 2\) \([48]\) – dominates, which is reasonable for binaries of stellar mass black holes with roughly equal mass \([49, 50]\). In this scenario, there is a clean separation in the GW amplitude between extrinsic and intrinsic parameters.

\[19\] In practice, we will use the redshift \(z\) instead, converting the one to the other using the Plank 2015 cosmology \([38]\) as coded up in astropy \([39]\).
\[ h(f, \vec{\theta}) = w(RA, \text{dec}, \iota, \psi) \frac{g(m_1, m_2, z, f)}{d_L(z)} \sqrt{\frac{5}{24\pi^{4/3}}} e^{i\psi(f, \vec{\theta})} \]  

(142)

where the function \( w \) is defined as \[ w \equiv \sqrt{\left(1 + \cos^2 \iota \frac{F_+}{2} \right)^2 + (\cos \iota F_\times)^2}. \]  

(143)

\( F_+ (RA, \text{dec}, \psi) \) and \( F_\times (RA, \text{dec}, \psi) \) are the antenna patterns for the two GW polarizations: they are large in regions of the sky where the GW detectors are more sensitive \[42, 54, 55\]. The function \( w \) is quite easy to calculate, and it turns out it takes values in the range \([0, 1]\):

\[ w \in [0, 1] \]

The maximum value of \( w \) is reached for a CBC source that is exactly above the detector and with inclination equal to zero (i.e. angular momentum parallel to the line of sight).

If one refers to this lucky source as **optimally oriented**, then the SNR of any other source with the same masses and distance, but different sky position and orientation, can be re-cast in terms of the optimally oriented SNR:

\[ \rho(\vec{\theta}) = w(RA, \text{dec}, \iota, \psi) \rho_{oo}(m_1, m_2, z) \]  

(144)

where “oo” stands for optimally oriented.

Now, this is progress! Let’s see why. We can write the detection efficiency as:

\[ \alpha(\vec{\lambda}) = \int dm_1 dm_2 dz \left[ \int d(RA, \text{dec}, \iota, \psi) \rho(\text{det}|RA, \text{dec}, \iota, \psi, m_1, m_2, z) \pi(RA, \text{dec}, \iota, \psi) \right] \pi(m_1 m_2 z | \vec{\lambda}). \]

where we have made the assumption that the population priors for the sky position and orientation are independent from the ones for masses and redshift. I am using \( d(RA, \text{dec}, \iota, \psi) \) as a short way of indicating the product of differentials of all 4 variables.

Let’s focus on the integral in square brackets. It contains the probability that an event is detected, given the value of all parameters.

But, given our criterium for detectability, this is exactly 1 if \( \vec{\theta} \) is such that the SNR is above \( \rho_{\text{thr}} \), and zero otherwise.

This integral can thus be written as

\[ \int d(RA, \text{dec}, \iota, \psi) H(\rho - \rho_{\text{thr}}) \pi(RA, \text{dec}, \iota, \psi) = \]

\[ \int d(RA, \text{dec}, \iota, \psi) H(w \rho_{oo} - \rho_{\text{thr}}) \pi(RA, \text{dec}, \iota, \psi) = \]

\[ \int d(RA, \text{dec}, \iota, \psi) H \left( w - \frac{\rho_{\text{thr}}}{\rho_{oo}} \right) \pi(RA, \text{dec}, \iota, \psi) \]  

(145)

with \( H \) the Heaviside function. Note that all the dependence on the masses and distance is now contained in \( \rho_{oo} \). Solving this integral can be done numerically as follows:

---

\(^{20}\) You might have noticed that the coalescence time and phase are gone. This is because they do not contribute to the optimal SNR, hence integrating over them will just given an overall factor to \( \alpha \), the same for all events and all \( \vec{\lambda} \), which can thus be neglected.
• pick a population prior for the sky position and orientation. The standard is isotropic on both. That gives \( \pi(\text{RA,dec,} \iota, \psi) = (4\pi^4)^{-1} \)

• Generate a large number of samples from this prior, and calculate \( w \) for all of them. Those points represent a probability density function \( p(w) \).

• The integral above becomes:

\[
\int_{\rho_{thr} \rho_{oo}}^{1} dw \ p(w) = 1 - \text{CDF}_w \left( \frac{\rho_{thr}}{\rho_{oo}} \right)
\]

(146)

where \( \text{CDF} \) is the usual cumulative distribution function of a variable.

This can also be seen as follows: let’s pack RA,dec, \( \iota, \psi \) into a vector \( \vec{y} \) then Eq. (145) becomes:

\[
\int d\vec{y} \ H \left( w(\vec{y}) - k \right) \pi(\vec{y}) = \frac{1}{N_{\text{samples}}} \sum_{j=1}^{N_{\text{samples}}} H \left( w(\vec{y}) - k \right) |_{\vec{y} \sim \pi(\vec{y})}
\]

(147)

where we called \( k \equiv \frac{\rho_{thr}}{\rho_{oo}} \) which is a constant when this integral is evaluated. The term inside the sum is either 1 or 0, depending on whether the \( \vec{y} \) value being draw from the prior results in \( w(\vec{y}) \) to be larger or smaller than \( k \). That is, Eq. (145) is the fraction of points drawn from \( \pi(\vec{y}) \) (i.e. \( \pi(\text{RA,dec,} \iota, \psi) \)) which gives a \( w \) larger than some number \( k \). This is exactly what Eq. (146) shows.

One minus the CDF, is often called the complementary cumulative distribution function (CCDF) or tail function. We will use CCDF from now on:

\[
\int_{\rho_{thr} \rho_{oo}}^{1} dw \ p(w) \equiv \text{CCDF}_w \left( \frac{\rho_{thr}}{\rho_{oo}} \right)
\]

This is beautiful: we have reduced all of the dependence on masses and distance to restricting the range of a 1D integral! We typically just calculate the distribution of \( w \) once, and then calculate the CCDF, or rather use an interpolator which makes it even less time consuming, as many times as we want during the analysis.

Back to the selection function, we can now write it as

\[
\alpha(\tilde{\lambda}) = \int dm_1 dm_2 dz \ \text{CCDF}_w \left( \frac{\rho_{thr}}{\rho_{oo}} \right) \pi(m_1 m_2 | \tilde{\lambda}).
\]

(148)

The complementary cumulative distribution function is a function of the two masses and the redshift, and it can be interpreted as the fraction of sources with given masses and redshift that are detectable. In Figure 6 we show this probability as a function of the two (source-frame) masses, fixing the redshift at 0.1.

We have reduced the problem, within our simplifying hypothesis, to a 3D integral where the only computationally consuming part is the calculation of the SNR of an optimally oriented source of given masses and redshift.

In practice, this can also be made very fast using interpolation in the mass plane, and the fact that the SNR scales linearly with the distance. There is a catch though: the SNR depends on the
detector-frame masses, which are redshifted by a factor of $1 + z$ over the astrophysical masses, the ones we care about \cite{56,57}. The integral for $\alpha$, as written above, is on the source-frame masses. The simple rescaling of the SNR as $1/d_L$ would give us the SNR of a source with the same detector-frame masses at different distances, if we know its SNR at some reference distance. But all those sources would thus have different source-frame masses! This is not a problem, we just need to fold in the redshift carefully when we rescale the SNR of sources.

This completes our discussion on how to calculate the selection function. If all we want is to infer $\tilde{\lambda}$, and not the number of sources that occurred over the experiment time (or their rate per unit time) we can just use Eq. (114) and we are done. Otherwise, some more work is required. This is because $\alpha$ doesn’t tell us anything about the actual number of events, only about the fraction of detectable events. If we want to do inference on the actual number of sources, or their time-rate, we also need to calculate the expected number of detectable events $N^\uparrow$ over the experiment time $T$ so that we can calculate Eq. (93). Luckily, it’s easy because $N^\uparrow$ and $N^\uparrow$ are related by $\alpha$, which we already have:

$$N^\uparrow = N^\uparrow \alpha(\tilde{\lambda}) = \int dm_1 dm_2 dz \text{CCDF}_w \left( \frac{\rho_{\text{thr}}}{\rho_{oo}} \right) N^\uparrow \pi(m_1 m_2 | \tilde{\lambda})$$

$$= \int dm_1 dm_2 dz \text{CCDF}_w \left( \frac{\rho_{\text{thr}}}{\rho_{oo}} \right) \frac{d^3 N^\uparrow}{dm_1 dm_2 dz} (\Lambda) \quad (149)$$

where we have used Eq. (5) to relate the population prior to the differential rate per unit masses and unit redshift. Usually, the generation of compact binaries is quoted in units of mergers per unit of comoving volume per unit of (source-frame) time, and in the most generic case depends on masses, redshift, spins, as well as environmental factors (e.g. the metallicity). Let’s use the symbol $R$ for it and only explicit report masses and redshift dependence.
Detectable fraction at $z = 0$.1

Figure 6: The fraction of detectable BBH of given $m_1$ and $m_2$ (in the source frame) at $z = 0.1$. A single interferometer is used with a sensitivity representative of LIGO's second observing run, together with a threshold SNR of 8.

\[ \mathcal{R}(m_1, m_2, z) \equiv \frac{d^2N^*}{dV_c dt_s} \]  

then:

\[ N^*_\uparrow = \int dm_1 dm_2 dz \text{CCDF}_w \left( \frac{\rho_{\text{thr}}}{\rho_{\text{oo}}} \right) \frac{d^2N^*}{dm_1 dm_2 dV_c dt_s} \frac{dV_c dt_s}{dz} \frac{dt_d}{dz} \]  

\[ = T \int dm_1 dm_2 dz \text{CCDF}_w \left( \frac{\rho_{\text{thr}}}{\rho_{\text{oo}}} \right) \frac{d^2R}{dm_1 dm_2} (m_1, m_2, z, \Lambda) \frac{dV_c}{dz} \frac{1}{1+z} \]  

where we have used the standard cosmological time dilation formula $\frac{dt_d}{dz} = (1+z)^{-1}$ to relate time measured in the detector frame, $t_d$, and time measured in the source frame, at redshift $z$, $t_s$ [58]. By definition, the integration over the detector-frame time gives the total observing time $T$ as measured with our clocks.

Equation (151) is fundamentally the product of three terms: an observing time, a source rate per unit time per unit volume, and a volume. What makes it look more complex than this is the fact that the source rate in general depends of a few parameters, and the fact that only part of that rate will contribute to the detectable signals.

Before we continue with our example, I want to report two very common ways of writing this equation, which are often found in the literature. One is (e.g. [53, 28, 11]):
\[ N^* \uparrow = T \int dm_1 dm_2 dz \ p_{\text{det}}(m_1, m_2, z) \frac{d^2 R}{dm_1 dm_2} (m_1, m_2, z, \Lambda) \frac{dV_c}{dz} \frac{1}{1+z} \]  

(152)

where \( p_{\text{det}}(m_1, m_2, z) \) is used for the probability, averaged over sky positions and orientations, that a source of given source frame masses and redshift is detectable.

The other makes use of the so called the 

\[ V(m_1, m_2) \equiv \int dz \frac{1}{1+z} \frac{dV_c}{dz} p_{\text{det}}(m_1, m_2, z) \]  

(153)

to define the 

\[ VT(m_1, m_2) \]  

(154)

with \( T \) being the observation time measured in the detector frame. The sensitive time-volume, usually reported in Gpc³ yr, represents the amount of universe volume “covered” by the instrument in searching CBCs of masses \( (m_1, m_2) \), after an observing time \( T \), as measured by the instrument clocks. Then, if the intrinsic merger rate \( R \) is assumed constant in time and redshift, the number of expected detectable sources after an observing time \( T \) can be written \[46, 59, 11\]:

\[ N^* \uparrow = \int dm_1 dm_2 \frac{d^2 R}{dm_1 dm_2} (m_1, m_2, \Lambda) \ VT(m_1, m_2) \]  

(155)

We now have all the background information to approach a simple problem. We will assume that the source-frame masses of the BHs in the binaries come from independent gaussian distributions of known standard deviations (equal to 5 solar masses \( M_\odot \)) for both \( m_1 \) and \( m_2 \) and unknown means \( \mu_1 \) and \( \mu_2 \), which we want to infer. Our hyper parameter vector is thus \( \vec{\lambda} = \{ \mu_1, \mu_2 \} \), and we will use a uniform hyper prior \( \pi(\mu_1) = \pi(\mu_2) = (35 - 5)^{-1} M_\odot^{-1} \) \[21\].

Sky position and orientation of the sources are isotropic, and we assume that sources are distributed uniformly in comoving volume.

The only parameters that depend on a hyper level are thus the two component masses. We take their population prior to be:

\[ \pi(m_1 m_2 | \vec{\lambda} = \{ \mu_1, \mu_2 \}) = \mathcal{N}(m_1 - \mu_1, 5 M_\odot) \mathcal{N}(m_2 - \mu_2, 5 M_\odot) \]  

(156)

where as before \( \mathcal{N} \) is the normal distribution. Thus the population prior we will use when calculating Eq. (148) reads:

\[ \pi(m_1 m_2 z | \vec{\lambda}) = \pi(m_1 m_2 | \vec{\lambda}) \pi(z) = \frac{dV_c}{dz}(z) \mathcal{N}(m_1 - \mu_1, 5 M_\odot) \mathcal{N}(m_2 - \mu_2, 5 M_\odot) \]  

(157)

Where \( \pi(z) = dV_c/dz \) is the prior on the redshift that results from a uniform in comoving volume prior.

Next, we need to generate some detectable sources from the true population, that will be the input used in our inference. If we were applying our analysis to real GW detections, this step would

\[ \text{In real life, little is known, and the problem’s dimensionality can be high. Have a look at Ref. [11] to get a feeling of a realistic suite of analysis.} \]
be unnecessary since nature would have produced them for us! But today we are not that lucky, and we have to create our own set of detections and the resulting measurements for those parameters.

To generate a set of noisy detections we proceed as follows:

1. We draw a random value of \( \vec{\theta} \) from the true population distribution of masses, distance, etc. We take the true value of the hyper-parameters to be \( \mu_1^T = 19 \, M_\odot \) and \( \mu_2^T = 20 \, M_\odot \);

2. We calculate the true SNR of the source;

3. Now, if we want to do a realistic analysis, we cannot quite use this SNR to decide if the source is detected, because in real life what we get is a noisy version of the true signal. The inferred value of the parameters need not be centered at the true values, but might be offset by an amount that depends on how loud the signal is. We thus add to the true SNR a random number from \( \mathcal{N}(0,1) \). This gives an observed SNR:

\[
\rho_{\text{obs}} = \rho_{\text{true}} + \mathcal{N}(0,1);
\]

4. If \( \rho_{\text{obs}} > 8 \) the source is detected and its \( \vec{\theta} \) recorded;

5. We restart from point 1 and keep going till we generate 20 detectable systems.

Finally, we need to generate for each event a set of samples from the likelihood, which we will need in Equation 110. This is done as follows:

1. The measured maximum likelihood value of the generic parameter \( x \) is taken to be the true value generated above plus an offset \( \mathcal{N}(0,\sigma_x) \), where \( \sigma_x \) depends on the observed SNR (with louder events having smaller offsets). In a scientific analysis, one would use a simulation campaign to “tune” the value of \( \sigma_x \) to something representative for each parameters [61]. In our case we will just use

\[
\sigma_{m_1} = \frac{1 M_\odot}{\rho_{\text{obs}}}; \quad \sigma_{m_2} = \frac{2 M_\odot}{\rho_{\text{obs}}}
\]

2. For each event and each parameter we create a set of 150 samples from a gaussian likelihood centered at the measured maximum likelihood point, and with a standard deviation equal to \( \sigma_x \).

It should be stressed that this implies we are assuming the likelihood of each event can be written as the product of a univariate Gaussian distribution for each parameter. This is an obvious over-simplification, which we can afford because we are only after a pedagogical example. In we were working with real samples produced by analyzing real data, the likelihood would in general be more complicated and show non-trivial correlations, but the inference would proceed along the very same steps we are following.

If the population hyper parameters only affect the mass distribution, and as long as the population prior on the other parameters (e.g, redshift) matches what we use to create the population of detected signals (remember Sec. 5.1) we actually only need to generate likelihood samples for the mass parameters.

---

22 The procedure described in this page is a simplified version of Appendix A of [35]. It is simpler because we work with uniform priors so that samples from the likelihood and samples from the posterior are equivalent. See also [61].

23 Given how we have defined SNR above, one can show that if the GW noise were perfectly Gaussian and stationary, the measured SNR would indeed be distributed like this.

24 \( \sigma_x \) thus plays a double role: it sets the offset of the measured maximum likelihood value compared to the true value and it sets the width of the likelihood distribution.
7.1 Which results and why

We now have all we need! Figure 7 shows the main results. The measured maximum likelihood mass values of the individual 20 detected sources are shown as small empty circles. The blue star shows what one would obtain if they just took the mean of these points, without properly taking into account selection effects, which makes heavier systems easier to measure (cfr Figure 6). As one could have guessed, the naive mean is at larger values of mass than the truth. Finally, the contours show the proper posterior on the hyper parameters obtained using Eq. (110), which is centered at the true value \((\mu_1^T, \mu_2^T)\).

So, what did we learn? There are at least a few different way of answering this question. The most straightforward answer is: “we now have a measure of the hyper-parameters, including their uncertainty”. This is what was shown by the contours of Figure 7. But we can do more. With the samples from the hyper posterior in hand, we can visually represent draws from the population prior, Eq. (156). This is show in Figure 8, left panel: we draw 2000 samples from the 1D marginalized posterior for \(\mu_1\), \(p(\mu_1|D)\), for each of them we calculate \(p(m_1|\mu_1)\) and plot it. The right panel shows the same set of lines for the other mass.

This figure is objectively quite nice and hypnotic (at least if you are reading the eye-friendly version of these notes) but having so many lines makes it hard to interpret. What we can do instead, is to e.g. calculate the 5th and 95th percentile, together with the median, of all draws at every point in the x axis, and plot these instead.

This is show for both masses in Figure 9. The solid curve represents the median of all the lines in Figure 8 at each point, and the shaded areas show the symmetric 90% credible interval. These curves represent what we know about the distribution of the component masses in light of our 20 triggers. A related quantity one often uses is the posterior predictive distribution (PPD) [62].

The PPD of a generic parameter vector \(\theta\) is the posterior we assign to the parameters in light of
Figure 8: (left:) \( p(m_1 | \mu_1) \) for 2000 draws from the posterior distributions of \( \mu_1 \); (right:) same, but for the other mass.

Figure 9: (left:) The median (solid line) and 90\% symmetric credible interval (shaded area) of the 2000 curves in Figure 8 for \( m_1 \). The dashed line is the posterior predictive distribution; (right:) same, but for the other mass.

the data we observed:

\[
p(\vec{\theta}|D)_{\Lambda}
\]

where we have put a subscript \( \Lambda \) to stress that this is not the posterior of \( \vec{\theta} \) given the data (which would not necessarily require any knowledge about the population). The PPD tells us what we expect the (true) distribution of the parameters to be, in light of the observations we made. As such, it can be used to quantitatively predict (hence the name) what future observations for sources from our population will look like, given the ones we already made.\(^{25}\)

Mathematically, we can related the PPD to the hyper posterior samples we have acquired by using the sum rule:

\[
p(\vec{\theta}|D)_{\Lambda} = \int d\Lambda \ p(\vec{\theta}|D\Lambda)p(\Lambda|D) = \int d\Lambda \ \pi(\vec{\theta}|\Lambda)p(\Lambda|D)
\]

\(^{25}\)Please notice that this is the distribution of true parameters of the sources before selection effects. The distribution of detected sources need will look like this due to selection effects and noise.
where we have used the fact that the distribution of $\vec{\theta}$ is fully specified if $\Lambda$ is known (and it is population prior) to remove the data from the right hand side of the condition bar in $p(\theta|DA)$.

Given the $N_{\text{samples}}$ we already got from the hyper posterior, $p(\Lambda|D)$, we can approximate this as:

$$p(\vec{\theta}|D) \Lambda \simeq \frac{1}{N_{\text{samples}}} \sum_{i=1}^{N_{\text{samples}}} \pi(\vec{\theta}|\Lambda^i) \bigg|_{\Lambda^i \sim p(\Lambda|D)}$$

That is: we calculate the population prior at each value of the hyper posterior samples, and we take the mean. The marginalized 1D PPD for $m_1$ and $m_2$ in our problem are thus just the mean of the curves shown in Figure 9 and we report them in Figure 9 using dashed lines.

8 Acknowledgments

I would like to thank Sylvia Biscoveanu for a critical reading of an early draft of these notes. Thanks to Emanuele Berti, Sylvia Biscoveanu, Will Farr, Max Isi and Ken Ng for long and illuminating discussions. I acknowledges the support of the National Science Foundation and the LIGO Laboratory. LIGO was constructed by the California Institute of Technology and Massachusetts Institute of Technology with funding from the National Science Foundation and operates under cooperative agreement PHY-0757058. I acknowledge support of the MIT physics department through the Solomon Buchsbaum Research Fund. I acknowledge use of iPython [63], Matplotlib [64, 65], NumPy [66], SciPy [67], emcee [68] and SeaBorn [69]. This is LIGO Document P2000231

42
Glossary and main symbols

$\alpha(\bar{\lambda})$ The fraction of physical sources that are detectable by our experiment, according to some detection threshold. It depends on the shape population hyper parameters $\bar{\lambda}$.

$\land$ Logical “and” of two propositions. $A \land B$ is true if and only if both $A$ and $B$ are true.

**Background** An (usually uninteresting) background event that might be produced by any mechanism not related to the physics we are after. They can contaminate our sample and we usually don’t like them.

$B_t$ The number of background sources generated per unit time, i.e. $B_t = dN^b/dt$. It generally is a function of time.

$D$ The set of data actually collected by the experiment. It usually is partitioned in a set that contains triggers, and another that does not contain triggers.

$D^\uparrow$ For a given trigger, the set of data that could conceivably be produced by the experiment and for which the trigger would be detectable.

$D^\downarrow$ For a given trigger, the set of data that could conceivably be produced by the experiment and for which the trigger would not be detectable.

$D$ For a given trigger, the full set of data that could conceivably be produced by the experiment.

$d_i$ A piece of data of small duration $\delta t$ that has been found to contain a trigger at time $t_i$.

$\bar{d}_j$ A piece of data covering the time interval $[t_j, t_j + \delta t]$ that has been found not to contain any trigger.

**Detectable** A trigger is detectable if it would have a detection statistics above our pre-determined threshold. Does not necessarily implies detected (e.g. the instrument might not be taking data at that time). Everything that is detected was detectable, but not everything that was detectable is detected.

**Detected** A trigger is detected if it is above the detection threshold and we recorded the data that contains it. This is what we use for our inference. Everything that is detected was detectable and has occurred.

$\mathcal{H}$ Generically speaking, all information that is relevant to describe our experiment and data analysis. Something like this should always go on the right side of conditional bars in our probability density functions, but often, because we are lazy, we omit it.

$\mathcal{H}_\Lambda$ All information relevant for the hyper analysis, e.g. the parametrization of model, the range of validity of the hyper parameters, the population priors, etc. It will be omitted unless its absence can create ambiguity.
\mathcal{H}_{\text{PE}} \quad \text{All information relevant for the parameter estimation analysis of individual sources, e.g. the parametrization of model, the range of validity of the event parameters, the sampling prior used in MCMC, Nested sampling, or similar algorithm. In general the prior on the parameters used for sampling, } \pi(\hat{\theta}|\mathcal{H}_{\text{PE}}), \text{ will be different from what implied by the population model, } \pi(\hat{\theta}|\Lambda\mathcal{H}_{\Lambda}). \text{ It will be omitted unless its absence can create ambiguity.}

\textbf{Hyper parameters} \quad \text{The (usually unknown) set of parameters that control the properties of the individual sources in the population. We would like to infer the value of the hyper parameters, given a collection of detected sources. We use the symbol } \Lambda \text{ for the hyper parameters that fully characterize a model } \mathcal{H}_{\Lambda}.

\Lambda \quad \text{The vector of (usually unknown) hyper parameters that characterize the overall population. It might be further split in a subset of shape parameters and a scale parameter.}

\bar{\Lambda} \quad \text{The vector of (usually unknown) shape hyper parameters that characterize the overall population. It can control e.g. the ratio of black holes generated at different masses.}

M_{\odot} \quad \text{The mass of our Sun today.}

N^b \quad \text{The total number of background sources, detectable and not. It generally is a function of time.}

N^b_{\uparrow} \quad \text{The number of background sources that are above the detection threshold, and are thus detectable (though not necessarily detected!). It generally is a function of time.}

N^b_{\text{act}} \quad \text{The number of background events that were actually produced over some time interval if } N^b_{\exp} \text{ were expected. These two numbers are not the same due to the Poissonian nature of the generation mechanism, which we assume in these notes. They might not all be detectable! It might be time dependent.}

N^b_{\downarrow} \quad \text{The number of background sources that are below the detection threshold, and thus cannot be detected. It generally is a function of time.}

N^b_{\exp} \quad \text{The number of background events that are expected to be produced over some time interval. They might not all be detectable! It might be time dependent.}

N^b_{\text{obs}} \quad \text{The number of background sources that are actually observed, over the experiment time. It generally is a function of time.}

\neg \quad \text{Logical negation of the proposition that follows it. } \neg A \text{ is true if } A \text{ is false and it’s false if } A \text{ is true.}

N^s \quad \text{The total number of physical sources, detectable and not, produced in the universe. It generally is a function of time and hyper parameters } \Lambda.

N^s_{\uparrow} \quad \text{The number of physical sources that are above the detection threshold, and are thus detectable (though not necessarily detected!). It generally is a function of time and hyper parameters } \Lambda.

N_{\text{samples}} \quad \text{The number of samples that have been drawn from a distribution (usually a posterior density function or a likelihood). It might be different for different sources, though in these notes we assume it’s always to same.}
\( N_{\text{act}}^{s} \) The number of physical sources that were \textit{actually} produced over some time interval if \( N_{\text{exp}}^{s} \) were expected. These two numbers are not the same due to the Poissonian nature of the generation mechanism. They might not all be detectable! It generally is a function of time and hyper parameters \( \Lambda \).

\( N_{\downarrow}^{s} \) The number of physical sources that are below the detection threshold, and thus cannot be detected. It generally is a function of time and hyper parameters \( \Lambda \).

\( N_{\text{exp}}^{s} \) The number of physical sources that are expected to be produced over some time interval. They might not all be detectable! It generally is a function of time and hyper parameters \( \Lambda \).

\( N_{\text{obs}}^{s} \) The number of physical sources that are actually observed, over the experiment time. It generally is a function of time and hyper parameters \( \Lambda \).

\( N_{\text{tr}} \) The number of trigger that are actually observed over the experiment time. This might include both background sources and interesting physical sources. It generally is a function of time and hyper parameters \( \Lambda \) (through the sources).

\textbf{Occurrence} The occurrence of a trigger just implies it happened, not necessarily that it was detected (or detectable). Everything that is detected or detectable occurred, but not everything that has occurred is detected or detectable.

\( p(\text{det}|\vec{\theta}) \) The probability that a trigger with true parameters \( \vec{\theta} \) is detectable (not necessarily detected!!). I would have preferred to use \( p(\rho_{t}|\vec{\theta}) \) for this, but the use of “det” is so spread in the literature that I would probably only have confused people if I introduced new notation.

\textbf{Population prior} \( \pi(\vec{\theta} | \Lambda, \mathcal{H}_\Lambda) \). The expected distribution of the individual source parameters \( \vec{\theta} \) given the population hyper parameters \( \Lambda \) and the relative modeling \( \mathcal{H}_\Lambda \). \( \mathcal{H}_\Lambda \) is often dropped unless needed.

\textbf{Population} The true set of physical or astrophysical sources that we are trying to characterize by mean of a model described by some hyper parameters \( \Lambda \). In these notes we assume that only one astrophysical population exists that is producing all the sources. In real life there could be several (e.g. one for binary black holes and one for binary neutron stars), each described by its one set of hyper parameters and modeling \( \mathcal{H} \).

\( \rho \) See SNR.

\textbf{Sampling prior} \( \pi(\vec{\theta}|\mathcal{H}_{\text{PE}}) \). The expected distribution of the individual source parameters \( \vec{\theta} \) used in an MCMC or similar algorithm while sampling the posterior of the individual triggers, given the settings used for the sampling \( \mathcal{H}_{\text{PE}} \).

\textbf{SNR} The signal-to-noise ratio of an event. It depends on the true source parameters as well as on the properties of the detector. It can be used to decide if something is detectable.

\( S_\kappa \) The differential rate over a set of parameters belonging to a vector of \( \vec{\kappa} \), i.e. \( S_\kappa \equiv \frac{\text{d}^n N^{s}}{\text{d} \kappa_1 \ldots \text{d} \kappa_n} \). It might be a function of time, hyper parameters \( \Lambda \) and \( \vec{\kappa} \).

\( S_t \) The number of physical sources generated per unit time, i.e. \( S_t = \frac{\text{d} N^{s}}{\text{d} t} \). It generally is a function of time and hyper parameters \( \Lambda \).
$S_{t,V_c}$: The number of physical sources generated per unit time per unit comoving volume, i.e. $S_{t,V_c} = \frac{d^2N_s}{dtdV_c}$. It generally is a function of time and hyper parameters $\Lambda$.

**Source**: A physically interesting event, that we are trying to detect and use for our inference.

$T$: The duration of the experiment. When we talk of the number of triggers that have occurred, or are detected, or detectable, we might implicitly mean “over the duration of the experiment, $T$”.

$\vec{\theta}$: The vector of (usually unknown) parameters that characterize each and every source. Either us or someone else usually generates a finite number of fair draws from the posterior density function of $\vec{\theta}$.

$\sim$: Used to indicate a fair draw from a distribution. E.g. $x \sim p(x|\mathcal{H})$ means that $x$ is drawn from the distribution $p(x|\mathcal{H})$.

**Trigger**: A trigger recorder by our search. It is above the detection threshold, but we don’t know if it’s a real source or a background event.
References

[1] K. G. Malmquist. On some relations in stellar statistics. Meddelanden fran Lunds Astronomiska Observatorium Serie I, 100:1–52, March 1922.

[2] K. G. Malmquist. A contribution to the problem of determining the distribution in space of the stars. Meddelanden fran Lunds Astronomiska Observatorium Serie I, 106:1–12, February 1925.

[3] Ilya Mandel, Will M. Farr, and Jonathan R. Gair. Extracting distribution parameters from multiple uncertain observations with selection biases. Mon. Not. Roy. Astron. Soc., 486(1):1086–1093, 2019. doi: 10.1093/mnras/stz896.

[4] Thomas J. Loredo and Don Q. Lamb. Bayesian analysis of neutrinos observed from supernova SN-1987A. Phys. Rev. D, 65:063002, 2002. doi: 10.1103/PhysRevD.65.063002.

[5] Daniel Foreman-Mackey, David W. Hogg, and Timothy D. Morton. Exoplanet Population Inference and the Abundance of Earth Analogs from Noisy, Incomplete Catalogs. ApJ, 795:64, November 2014. doi: 10.1088/0004-637X/795/1/64.

[6] Joshua N. Winn and Daniel C. Fabrycky. The Occurrence and Architecture of Exoplanetary Systems. ARA&A, 53:409–447, August 2015. doi: 10.1146/annurev-astro-082214-122246.

[7] Thomas J. Loredo and Ira M. Wasserman. Inferring the Spatial and Energy Distribution of Gamma-Ray Burst Sources. I. Methodology. ApJS, 96:261, January 1995. doi: 10.1086/192119.

[8] Thomas J. Loredo and Ira M. Wasserman. Inferring the spatial and energy distribution of gamma-ray burst sources. 2. Isotropic models. Astrophys. J., 502:75, 1998. doi: 10.1086/305870.

[9] Sylvia Biscoveanu, Eric Thrane, and Salvatore Vitale. Constraining short gamma-ray burst jet properties with gravitational waves and gamma rays. Astrophys. J., 893:38, 2020. doi: 10.3847/1538-4357/ab7eaf.

[10] Fergus Hayes, Ik Siong Heng, John Veitch, and Daniel Williams. Comparing Short Gamma-Ray Burst Jet Structure Models. Astrophys. J., 891:124, 2020. doi: 10.3847/1538-4357/ab72fc.

[11] B.P. Abbott et al. Binary Black Hole Population Properties Inferred from the First and Second Observing Runs of Advanced LIGO and Advanced Virgo. Astrophys. J. Lett., 882(2):L24, 2019. doi: 10.3847/2041-8213/ab3800.

[12] B.P. Abbott et al. GWTC-1: A Gravitational-Wave Transient Catalog of Compact Binary Mergers Observed by LIGO and Virgo during the First and Second Observing Runs. Phys. Rev. X, 9(3):031040, 2019. doi: 10.1103/PhysRevX.9.031040.

[13] Eric Thrane and Colm Talbot. An introduction to Bayesian inference in gravitational-wave astronomy: Parameter estimation, model selection, and hierarchical models. PASA, 36:e010, March 2019. doi: 10.1017/pasa.2019.2.

[14] Thomas J. Loredo and Martin A. Hendry. Multilevel and hierarchical Bayesian modeling of cosmic populations. arXiv e-prints, art. arXiv:1911.12337, November 2019.

[15] B. P. Abbott, R. Abbott, T. D. Abbott, et al. Observation of gravitational waves from a binary black hole merger. Phys. Rev. Lett., 116:061102, 2016.
[16] J. Abadie, B. P. Abbott, R. Abbott, M. Abernathy, T. Accadia, F. Acernese, C. Adams, R. Adhikari, P. Ajith, B. Allen, and et al. TOPICAL REVIEW: Predictions for the rates of compact binary coalescences observable by ground-based gravitational-wave detectors. *Classical and Quantum Gravity*, 27(17):173001, September 2010. doi: 10.1088/0264-9381/27/17/173001.

[17] M. Dominik, K. Belczynski, C. Fryer, D. E. Holz, E. Berti, T. Bulik, I. Mandel, and R. O'Shaughnessy. Double Compact Objects. II. Cosmological Merger Rates. *ApJ*, 779:72, December 2013. doi: 10.1088/0004-637X/779/1/72.

[18] B. P. Abbott, R. Abbott, T. D. Abbott, et al. The Rate of Binary Black Hole Mergers Inferred from Advanced LIGO Observations Surrounding GW150914. 2016. [https://dcc.ligo.org/LIGO-P1500217/public/main](https://dcc.ligo.org/LIGO-P1500217/public/main).

[19] B. P. Abbott, R. Abbott, T. D. Abbott, et al. Characterization of transient noise in Advanced LIGO relevant to gravitational wave signal GW150914. 2016. [https://dcc.ligo.org/LIGO-P1500238/public/main](https://dcc.ligo.org/LIGO-P1500238/public/main).

[20] Michael Zevin et al. Gravity Spy: Integrating Advanced LIGO Detector Characterization, Machine Learning, and Citizen Science. *Class. Quant. Grav.*, 34(6):064003, 2017. doi: 10.1088/1361-6382/aa5c6a.

[21] Sebastian M. Gaebel, John Veitch, Thomas Dent, and Will M. Farr. Digging the population of compact binary mergers out of the noise. *MNRAS*, 484(3):4008–4023, April 2019. doi: 10.1093/mnras/stz225.

[22] Colm Talbot and Eric Thrane. Measuring the binary black hole mass spectrum with an astrophysically motivated parameterization. *Astrophys. J.*, 856(2):173, 2018. doi: 10.3847/1538-4357/aab34c.

[23] Michael Zevin, Chris Pankow, Carl L. Rodriguez, Laura Sampson, Eve Chase, Vassiliki Kalogera, and Frederic A. Rasio. Constraining Formation Models of Binary Black Holes with Gravitational-Wave Observations. *Astrophys. J.*, 846(1):82, 2017. doi: 10.3847/1538-4357/aa8408.

[24] Salvatore Vitale, Ryan Lynch, Riccardo Sturani, and Philip Graff. Use of gravitational waves to probe the formation channels of compact binaries. *Class. Quant. Grav.*, 34(3):03LT01, 2017. doi: 10.1088/1361-6382/aa552e.

[25] Will M. Farr, Simon Stevenson, M. Coleman Miller, Ilya Mandel, Ben Farr, and Alberto Vecchio. Distinguishing Spin-Aligned and Isotropic Black Hole Populations With Gravitational Waves. *Nature*, 548:426, 2017. doi: 10.1038/nature23453.

[26] Colm Talbot and Eric Thrane. Determining the population properties of spinning black holes. *Phys. Rev. D*, 96(2):023012, 2017. doi: 10.1103/PhysRevD.96.023012.

[27] Maya Fishbach and Daniel E. Holz. Picky Partners: The Pairing of Component Masses in Binary Black Hole Mergers. *Astrophys. J. Lett.*, 891(1):L27, 2020. doi: 10.3847/2041-8213/ab7247.

[28] Maya Fishbach, Daniel E. Holz, and Will M. Farr. Does the Black Hole Merger Rate Evolve with Redshift? *ApJ*, 863(2):L41, Aug 2018. doi: 10.3847/2041-8213/aad800.
[29] J. Veitch, V. Raymond, B. Farr, W. Farr, P. Graff, S. Vitale, B. Aylott, K. Blackburn, N. Christensen, M. Coughlin, W. Del Pozzo, F. Feroz, J. Gair, C.-J. Haster, V. Kalogera, T. Littenberg, I. Mandel, R. O’Shaughnessy, M. Pitkin, C. Rodriguez, C. Röver, T. Sidery, R. Smith, M. Van Der Sluys, A. Vecchio, W. Vousden, and L. Wade. Parameter estimation for compact binaries with ground-based gravitational-wave observations using the LALInference software library. Phys. Rev. D, 91(4):042003, February 2015. doi: 10.1103/PhysRevD.91.042003.

[30] Gregory Ashton et al. BILBY: A user-friendly Bayesian inference library for gravitational-wave astronomy. Astrophys. J. Suppl., 241(2):27, 2019. doi: 10.3847/1538-4365/ab06fc.

[31] David MacKay. Information theory, inference, and learning algorithms. Cambridge University Press, Cambridge, UK, 2003. ISBN 978-0521642989.

[32] Gregory M. Harry. Advanced LIGO: The next generation of gravitational wave detectors. Class. Quant. Grav., 27:084006, 2010. doi: 10.1088/0264-9381/27/8/084006.

[33] F. Acernese et al. Advanced Virgo: a second-generation interferometric gravitational wave detector. Class. Quant. Grav., 32(2):024001, 2015. doi: 10.1088/0264-9381/32/2/024001.

[34] LIGO Virgo Collaboration. LIGO/Virgo Public Alerts User Guide. https://emfollow.docs.ligo.org/userguide/index.html, 2020.

[35] Maya Fishbach, Will M. Farr, and Daniel E. Holz. The Most Massive Binary Black Hole Detections and the Identification of Population Outliers. Astrophys. J. Lett., 891(2):L31, 2020. doi: 10.3847/2041-8213/ab77c9.

[36] Will M. Farr, Jonathan R. Gair, Ilya Mandel, and Curt Cutler. Counting and confusion: Bayesian rate estimation with multiple populations. Phys. Rev. D, 91:023005, January 2015. doi: 10.1103/PhysRevD.91.023005.

[37] Carl L. Rodriguez, Michael Zevin, Chris Pankow, Vasilliki Kalogera, and Frederic A. Rasio. Illuminating Black Hole Binary Formation Channels with Spins in Advanced LIGO. Astrophys. J. Lett., 832(1):L2, 2016. doi: 10.3847/2041-8205/832/1/L2.

[38] Planck Collaboration, P. A. R. Ade, N. Aghanim, M. Arnaud, M. Ashdown, J. Aumont, C. Baccigalupi, A. J. Banday, R. B. Barreiro, J. G. Bartlett, and et al. Planck 2015 results. XIII. Cosmological parameters. ArXiv e-prints, February 2015.
The Astropy Project: Building an Open-science Project and Status of the v2.0 Core Package. AJ, 156:123, September 2018. doi: 10.3847/1538-3881/aabc4f.

[40] Benjamin P. Abbott et al. Sensitivity of the Advanced LIGO detectors at the beginning of gravitational wave astronomy. Phys. Rev. D, 93(11):112004, 2016. doi: 10.1103/PhysRevD.93.112004. [Addendum: Phys.Rev.D 97, 059901 (2018)].

[41] B. P. Abbott, R. Abbott, T. D. Abbott, M. R. Abernathy, F. Acernese, K. Ackley, C. Adams, T. Adams, P. Addesso, R. X. Adhikari, and et al. Prospects for Observing and Localizing Gravitational-Wave Transients with Advanced LIGO and Advanced Virgo. Living Reviews in Relativity, 19, February 2016. doi: 10.1007/lrr-2016-1.

[42] B. S. Sathyaprakash and B. F. Schutz. Physics, Astrophysics and Cosmology with Gravitational Waves. Living Reviews in Relativity, 12:2, March 2009.

[43] Surabhi Sachdev et al. The GstLAL Search Analysis Methods for Compact Binary Mergers in Advanced LIGO’s Second and Advanced Virgo’s First Observing Runs. 1 2019.

[44] Samantha A. Usman et al. The PyCBC search for gravitational waves from compact binary coalescence. Class. Quant. Grav., 33(21):215004, 2016. doi: 10.1088/0264-9381/33/21/215004.

[45] B. P. Abbott, R. Abbott, T. D. Abbott, M. R. Abernathy, F. Acernese, K. Ackley, C. Adams, T. Adams, P. Addesso, R. X. Adhikari, and et al. GW150914: First results from the search for binary black hole coalescence with Advanced LIGO. Phys. Rev. D, 93(12):122003, June 2016. doi: 10.1103/PhysRevD.93.122003.

[46] Daniel Wysocki, Jacob Lange, and Richard O’Shaughnessy. Reconstructing phenomenological distributions of compact binaries via gravitational wave observations. Phys. Rev. D, 100(4):043012, 2019. doi: 10.1103/PhysRevD.100.043012.

[47] Ken K.Y. Ng, Salvatore Vitale, Aaron Zimmerman, Katerina Chatziioannou, Davide Gerosa, and Carl-Johan Haster. Gravitational-wave astrophysics with effective-spin measurements: asymmetries and selection biases. Phys. Rev. D, 98(8):083007, 2018. doi: 10.1103/PhysRevD.98.083007.

[48] Maggiore, M. Gravitational Waves, Volume 1: Theory and Experiments. Oxford University Press, 2007.

[49] C. Cutler and É. E. Flanagan. Gravitational waves from merging compact binaries: How accurately can one extract the binary’s parameters from the inspiral waveform? Phys. Rev. D, 49:2658–2697, March 1994. doi: 10.1103/PhysRevD.49.2658.

[50] R. Abbott et al. GW190412: Observation of a Binary-Black-Hole Coalescence with Asymmetric Masses. 4 2020.
[51] R. O’Shaughnessy, V. Kalogera, and K. Belczynski. Binary Compact Object Coalescence Rates: The Role of Elliptical Galaxies. *Astrophys. J.*, 716:615–633, 2010. doi: 10.1088/0004-637X/716/1/615.

[52] Eanna E. Flanagan and Scott A. Hughes. Measuring gravitational waves from binary black hole coalescences: 1. Signal-to-noise for inspiral, merger, and ringdown. *Phys. Rev. D*, 57:4535–4565, 1998. doi: 10.1103/PhysRevD.57.4535.

[53] Michal Dominik, Emanuele Berti, Richard O’Shaughnessy, Ilya Mandel, Krzysztof Belczynski, Christopher Fryer, Daniel E. Holz, Tomasz Bulik, and Francesco Pannarale. Double Compact Objects III: Gravitational Wave Detection Rates. *Astrophys. J.*, 806(2):263, 2015. doi: 10.1088/0004-637X/806/2/263.

[54] B. F. Schutz. Networks of gravitational wave detectors and three figures of merit. *Classical and Quantum Gravity*, 28(12):125023, June 2011. doi: 10.1088/0264-9381/28/12/125023.

[55] S. W. Hawking and W. Israel. Book-Review - 300 Years of Gravitation. *Journal of the British Astronomical Association*, 99:196, August 1989.

[56] B. F. Schutz. Determining the Hubble constant from gravitational wave observations. *Nature*, 323(6086):310–311, September 1986. doi: 10.1038/323310a0.

[57] A Krolak and Bernard F Schutz. Coalescing binaries—Probe of the universe. *General Relativity and Gravitation*, 19(12):1163–1171, 1987. ISSN 1572-9532. doi: 10.1007/BF00759095. URL [http://dx.doi.org/10.1007/BF00759095](http://dx.doi.org/10.1007/BF00759095).

[58] James B. Hartle. *Gravity : an introduction to Einstein’s general relativity*. 2003.

[59] B. P. Abbott, R. Abbott, T. D. Abbott, et al. Astrophysical Implications of the Binary Black-Hole Merger GW150914. *Astrophys. J.*, 818(2):L22, 2016. doi: 10.3847/2041-8205/818/2/L22.

[60] Hsin-Yu Chen, Daniel E. Holz, John Miller, Matthew Evans, Salvatore Vitale, and Jolien Creighton. Distance measures in gravitational-wave astrophysics and cosmology. 9 2017.

[61] Ilya Mandel, Will M. Farr, Andrea Colonna, Simon Stevenson, Peter Tiño, and John Veitch. Model-independent inference on compact-binary observations. *MNRAS*, 465(3):3254–3260, March 2017. doi: 10.1093/mnras/stw2883.

[62] Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2nd ed. edition, 2004.

[63] Fernando P´erez and Brian E. Granger. IPython : A System for. *IEEE Journals & Magazines*, 9(3):21–29, 2007. ISSN 15219615. doi: 10.1109/MCSE.2007.53.

[64] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007. doi: 10.1109/MCSE.2007.55.

[65] John D. Hunter. Matplotlib: A 2D graphics environment. *Computing in Science and Engineering*, 9(3):99–104, 2007. ISSN 15219615. doi: 10.1109/MCSE.2007.55.

[66] Stéfan Van Der Walt, S. Chris Colbert, and Gaël Varoquaux. The NumPy array: A structure for efficient numerical computation. *Computing in Science and Engineering*, 13(2):22–30, 2011. ISSN 15219615. doi: 10.1109/MCSE.2011.37.
[67] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, and van der Walt. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, 17(3):261–272, 2020. ISSN 15487105. doi: 10.1038/s41592-019-0686-2.

[68] Daniel Foreman-Mackey, David W. Hogg, Dustin Lang, and Jonathan Goodman. emcee: The MCMC Hammer. *PASP*, 125(925):306, March 2013. doi: 10.1086/670067.

[69] Michael Waskom, Olga Botvinnik, Drew O’Kane, Paul Hobson, Saulius Lukauskas, David C Gemperline, Tom Augspurger, Yaroslav Halchenko, John B. Cole, Jordi Warmenhoven, Julian de Ruiter, Cameron Pye, Stephan Hoyer, Jake Vanderplas, Santi Villalba, Gero Kunter, Eric Quintero, Pete Bachant, Marcel Martin, Kyle Meyer, Alistair Miles, Yoav Ram, Tal Yarkoni, Mike Lee Williams, Constantine Evans, Clark Fitzgerald, Brian, Chris Fonnesbeck, Antony Lee, and Adel Qalieh. mwaskom/seaborn: v0.8.1 (september 2017), September 2017. URL https://doi.org/10.5281/zenodo.883859