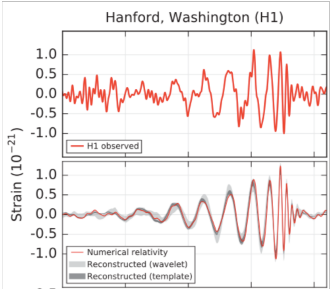
GW data analysis: parameter estimation

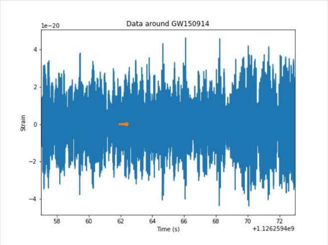
9.2.21

General schedule

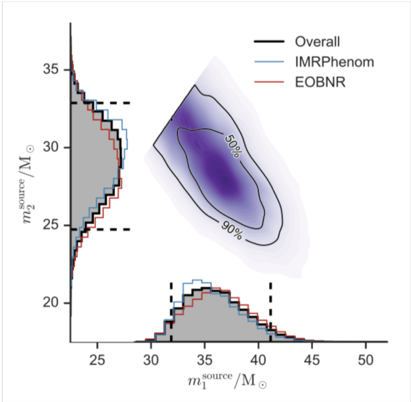
- History
- Introduction to general relativity
- Detection principles
- Detectors
- Binary black-hole system
- Bursts and continuous waves
- Rates and populations & cosmology
- Stochastic GW background. Tests of general relativity using GWs
- Data analysis: signal processing
- Data analysis: parameter estimation
    - Frequentist vs Bayesian
    - Bayesian inference
    - Samplers, hyper-parameters, marginalization
    - Fisher information matrix

After detection





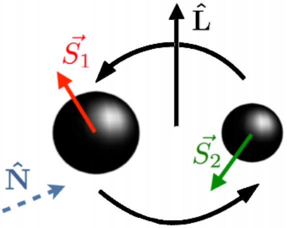
(detection of something other than just noise in the data)

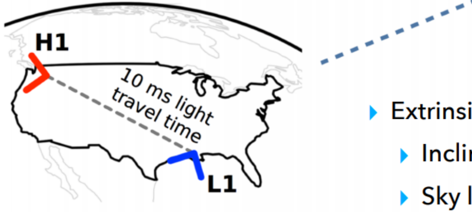


(establishing confidence that the signal is astrophysical and has certain parameters)

Binary system waveform: 15+ parameters

- Intrinsic:
  - masses
  - spins
  - tidal deformability





Credit: LIGO/Virgo

- Extrinsic:
  - Inclination, distance, polarisation
  - Sky location
  - Time, reference phase

Searching for the best fitted waveform → calculating the likelihood at sufficiently dense grid in 15-dimensional space → ”the curse of dimensionality”

Statistical approaches: frequentist vs Bayesian

- ⋆ Probability is the limit of the relative frequency of an event after many trials N:

  $$P = n/N$$

  where n is the number of desired outcomes,
- ⋆ only data from the current experiment when evaluating outcomes,
- ⋆ Establishing significance: p-value - evidence against a null hypothesis (the smaller the p-value, the stronger the evidence the null hypothesis should be rejected). Essentially the probability of a false positive based on the data in the experiment,
- ⋆ Null hypothesis: a default hypothesis that the given observation is not extraordinary,
- ⋆ P-values are (objective) probability statements about the data sample not about the hypothesis itself.

Notable figures in 20th century: Roland Fisher (1890 – 1962), Jerzy Neyman (1894 – 1981), Egon Pearson (1895 – 1980).

Statistical approaches: frequentist vs Bayesian

Thomas Bayes (1701-1761): ”probability is orderly opinion, and that inference from data is nothing other than the revision of such opinion in the light of relevant new information”

- probability expresses a degree of belief in an event,
- as opposed to the frequentist approach, bayesian method expresses the chance of an event happening (based on available data),
- conditional concept of probability: uses prior knowledge and current data to predict the posterior.

Statistical approaches: frequentist vs Bayesian

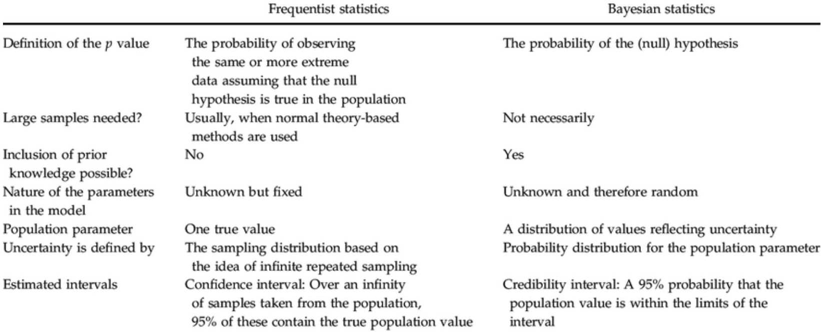
Frequentist:

- ⋆ never uses/gives the probability of a hypothesis (no prior or posterior),
- ⋆ depends on the likelihood $P(data|hypothesis)$ for both observed and unobserved data,
- ⋆ does not require a prior ("events occur with frequency"),
- ⋆ dominated statistical practice during the 20th century,
- ⋆ computationally feasible.

Bayesian:

- ⋆ uses probabilistic language for both hypotheses and data,
- ⋆ depends on the prior and likelihood of observed data,
- ⋆ requires one to know or construct a "subjective prior",
- ⋆ dominated statistical practice before the 20th century,
- ⋆ may be computationally intensive (due to integration over many parameters).

Statistical approaches: frequentist vs Bayesian



	Frequentist statistics	Bayesian statistics
Definition of the *p* value	The probability of observing the same or more extreme data assuming that the null hypothesis is true in the population	The probability of the (null) hypothesis
Large samples needed?	Usually, when normal theory-based methods are used	Not necessarily
Inclusion of prior knowledge possible?	No	Yes
Nature of the parameters in the model	Unknown but fixed	Unknown and therefore random
Population parameter	One true value	A distribution of values reflecting uncertainty
Uncertainty is defined by	The sampling distribution based on the idea of infinite repeated sampling	Probability distribution for the population parameter
Estimated intervals	Confidence interval: Over an infinity of samples taken from the population, 95% of these contain the true population value	Credibility interval: A 95% probability that the population value is within the limits of the interval

Bayes' theorem

$$\underbrace{p(\theta|d,M)}_{\textit{posterior}} = \frac{\overbrace{p(d|\theta,M)}^{\textit{likelihood}}\ \overbrace{p(\theta|M)}^{\textit{prior}}}{\underbrace{p(d|M)}_{\textit{evidence}}}$$

with

- ⋆ d: the data (e.g. time series from LIGO-Virgo detectors),
- ⋆ M: the model (e.g. the binary inspiral waveform)
- ⋆ θ: parameters of the model (e.g. the 15 parameters of the waveform, location in the sky etc.)

Sometimes written as

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

Bayes' theorem: posterior distribution p

We would like to obtain the posterior distribution,

$$p(\theta|d)$$

i.e. the probability density function for the continuous variables θ given the data d. Probability that the true value of θ is between $(\theta', \theta' + d\theta)$ is given by $p(\theta'|d)d\theta'$. The posterior is normalized:

$$\int p(\theta|d)d\theta = 1.$$

We want to learn the $p(\theta|d)$ not only for the values of θ but also to construct credible intervals for these values → estimate the errors of the measurement.

Bayes' theorem: likelihood $\mathcal{L}$

$\mathcal{L}(d|\theta)$ is the likelihood function of the data *given* the parameters θ.

- $\star$ Can be chosen according to the nature of the observation,
- $\star$ Is related to our assumed/known model of the noise,
- $\star$ In GW astronomy, we typically assume Gaussian noise likelihood function:

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

with $\mu(\theta)$ a template for the GW waveform given θ, and σ is the detector noise. Likelihood is typically not normalized w.r.t θ

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

(however, it is normalized w.r.t the data *d*, and describes the chance of getting data *d*. $\mathcal{L}$ is a probability density function with units of inverse data: integrated over all possible *d* gives 1).

Bayes' theorem: prior distribution π

Like $\mathcal{L}$, the prior is a choice: it incorporates our belief in θ before observation.

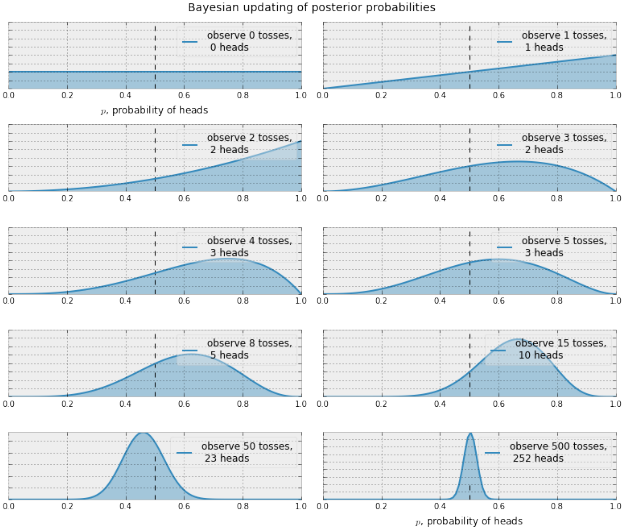
- $\star$ sometimes easy to chose e.g. for sky localisation (isotropic),
- $\star$ sometimes it depends on unknown astrophysics, e.g. mass of the primary component in the BBH system $\pi(m_1)$.

In case of no knowledge on certain θ_i, we usually assume an uniform or log-uniform distribution (the later if we don't know the order of magnitude of a quantity).

Of course, for the next measurement, previous posterior becomes prior ($\rightarrow$ building on previous knowledge).

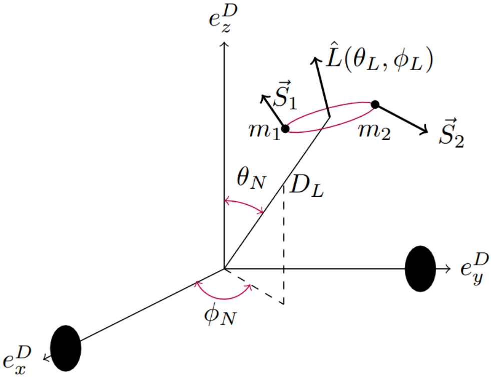
Example: coin flip

Updading posterior probability of p_H (coin flip results in heads)



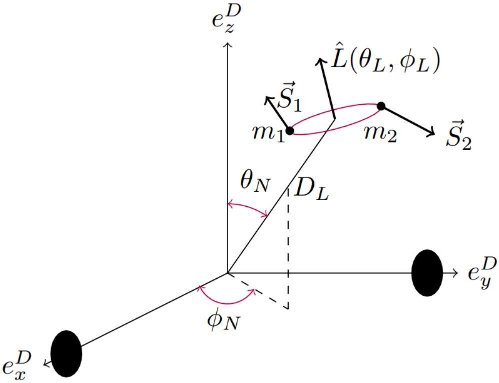
”Bayesian methods for hackers”, `https://camdavidsonpilon.github.io/Probabilistic-Programming-and-Bayesian-Methods-for-Hackers`

Binary system waveform: prior on extrinsic parameters



D_L	Uniform in volume
θ_N ϕ_N	Uniform in the sky
θ_L ϕ_L	Uniform in direction

Binary system waveform: prior on intrinsic parameters



m_1 m_2 Uniform in some range

$\vec{S}_2$ $\vec{S}_1$ Uniform in direction and magnitude in $[0, m_i^2]$

λ_1 λ_2 Uniform in (0,5000)

(In practice, sometimes non-trivial choices to be made).

Bayes’ theorem: marginalizing the posterior

Parameters θ is usually a large set of parameters. How to study a specific one?

Marginalizing (integrating ”away”) the parameters we are not interested in (the ”nuisance parameters”) to get a marginalized posterior

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

$\mathcal{L}(d|\theta_i)$ is the marginalized likelihood:

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

Bayes' theorem: marginalizing the posterior

Example:

- ⋆ Marginalization over variable θ_a to obtain a posterior on θ_b:
- → means calculating the best guess for θ_b given uncertainty in θ_a.

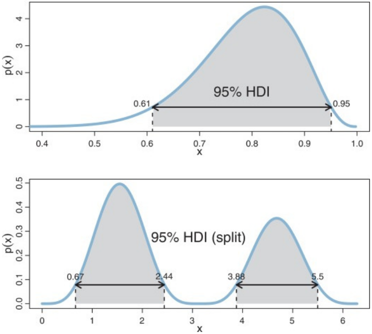
- ⋆ If θ_a and θ_b are covariant, marginalization over θ_a introduces its uncertainty into the posterior for θ_b.
- → The marginalized posterior $p(\theta_b|d)$ is broader than the *conditional posterior* $p(\theta_b|d,\theta_a)$ (=a slice through the $p(\theta_b|d)$ posterior at a fixed value of θ_a).

In the GW context of a binary system: covariance between the luminosity distance D_L and the inclination angle θ_{JN}.

Credible intervals

Credible interval: region of parameter space containing fraction of posterior probability (in frequentist approach: *confidence* intervals),

- ⋆ highest posterior density interval (HPDI):



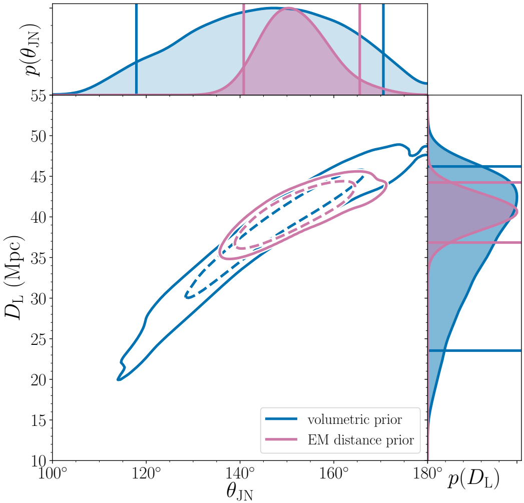
- ⋆ Relation to σ levels, e.g. 2σ credible region includes 95% of posterior probability,
- ⋆ Symmetric credible intervals: with a cumulative distribution function

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

the $X\%$ symmetric credible region is

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

GW170817 joint posterior for distance-inclination angle



Joint posterior for luminosity distance and inclination angle for GW170817.

Blue contours: 90% credible region using GW data alone, purple contours obtained with EM observation.

Bayes’ theorem: evidence $\mathcal{Z}$

The ”evidence” $\mathcal{Z}$ is usually treated as a normalization factor

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

but it also plays an important role in model selection and can be viewed as completely marginalized likelihood.

Model selection: signal vs noise

Evidence $\mathcal{Z}$ is useful to select a better model (=the one which is statistically preferred by the data), and quantify by how much it is better.

Example: "signal model" (described by $M(\theta)$) vs "noise model" (no parameters).

Let's define signal evidence $\mathcal{Z}_S$ and a noise evidence $\mathcal{Z}_N$ ("null likelihood" $\mathcal{L}(d|0)$):

$$\mathcal{Z}_S \equiv \int d\theta \mathcal{L}(d|\theta)\,\pi(\theta), \qquad \mathcal{Z}_N \equiv \mathcal{L}(d|0) = \frac{1}{2\pi\sigma^2}\exp\left(-\frac{1}{2}\frac{|d|^2}{\sigma^2}\right).$$

The ratio of the evidence for two different models is the Bayes factor.

$$\mathrm{BF}_N^S \equiv \frac{\mathcal{Z}_S}{\mathcal{Z}_N}, \quad \text{usually used: } \log\left(\mathrm{BF}_N^S\right) \equiv \log(\mathcal{Z}_S) - \log(\mathcal{Z}_N).$$

(strong evidence when $|\log \mathrm{BF}| = 8$).

Model selection: same model, different priors

$\star$ Let's compare two identical models employing different priors, e.g. BH waveform with uniform spin prior vs zero-spin prior. $\rightarrow$ Bayes factor should select which version is preferred by the data:

$$\mathcal{Z}_{\text{spin}} = \int d\theta \mathcal{L}(d|\theta)\, \pi(\theta) \qquad \mathcal{Z}_{\text{no spin}} = \int d\theta \mathcal{L}(d|\theta)\, \pi_{\text{no spin}}(\theta).$$

The spin/no spin Bayes factor is

$$\text{BF}^{\text{spin}}_{\text{no spin}} = \frac{\mathcal{Z}_{\text{spin}}}{\mathcal{Z}_{\text{no spin}}}.$$

$\star$ different models $M_A(\theta)$ and $M_B(\nu)$ with priors $\pi(\theta)$ and $\pi(\nu)$:

$$\mathcal{Z}_A = \int d\theta \mathcal{L}(d|\theta, M_A)\, \pi(\theta), \qquad \mathcal{Z}_B = \int d\nu \mathcal{L}(d|\nu, M_B)\, \pi(\nu).$$

The A/B Bayes factor is of course

$$\text{BF}^A_B = \frac{\mathcal{Z}_A}{\mathcal{Z}_B}.$$

Question: what if the number of ν parameters is different from θ parameters?

Bayes' odds

Formally, the correct metric to compare two models is not the Bayes factor, but rather the Bayes' odds

$$\mathcal{O}_B^A \equiv \frac{\mathcal{Z}_A}{\mathcal{Z}_B}\frac{\pi_A}{\pi_B}.$$

=the product of the Bayes factor with the prior odds π_A/π_B (which describes our prior belief about the relative likelihood of hypotheses A and B).

Evidence and ”Occam’s razor” factor

- ⋆ Likelihood $\mathcal{L}$ describes how well the model $M(\theta)$ fits the data d,
- ⋆ marginalization describes the size of the parameter space volume,
- ⋆ We want the best fit (highest likelihood) with the smallest prior volume (smallest model).

- ⋆ It’s possible that model with a decent fit and small prior volume yields a greater evidence than a model with an excellent fit and a huge prior volume. In these cases, the Bayes factor penalizes the more complicated model.

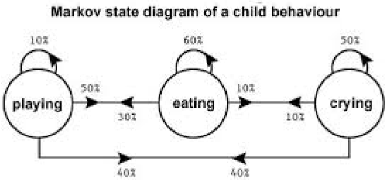
Solving practical problems: samplers

Searching for the best fitted waveform:

- ⋆ Since we (more or less) know how to produce precise-enough GW waveforms (”aproximants”) given parameters θ, it easy to predict how *d* looks given θ (a forward problem).
- ⋆ Calculating the posterior $p(\theta|d)$, the probability of parameters θ given the data *d* is an inverse problem.
- ⋆ Solution: stochastic sampling
  - ⋆ Markov-chain Monte Carlo (MCMC) (Metropolis et al., 1953; Hastings, 1970)
  - ⋆ Nested sampling (Skilling, 2004)
- ⋆ Result: a list of posterior samples $\{\theta\}$ drawn from the posterior distribution such that the number of samples on the interval $(\theta, \theta + \Delta\theta) \propto p(\theta)$

Samplers: MCMC

- ⋆ A Markov chain is a stochastic model describing a sequence of possible events in which the probability of each event depends only on the state attained in the previous event:



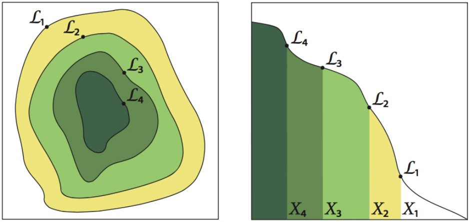
- ⋆ Particles (walkers) randomly walk through the parameter space, where the probability of moving to a new location is governed by the proposal density function, evaluating the likelihood,
- ⋆ Suitable proposal density is e.g. a Gaussian centered on the current location,
- ⋆ Burn-in period before the walkers ”forget” their starting positions.
- ⋆ Adjacent samples in a chain are correlated (chain thinning by the integrated autocorrelation time to obtain the correct posterior distribution).

Samplers: nested sampling

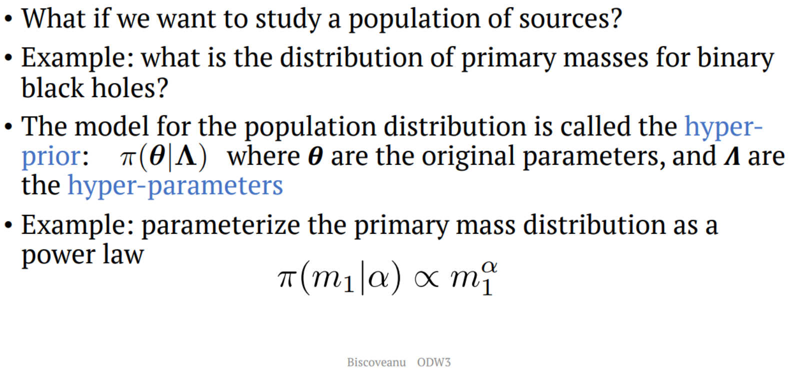
While MCMC methods probe the posterior distribution $p(\theta|d)$, nested sampling calculates the evidence $\mathcal{Z}$ → posterior samples are actually by-products, because *Likelihood* × *Prior* = *Evidence* × *Posterior*

1. parameter space populated with ”live points” drawn from the prior distribution,
2. At each iteration, the lowest likelihood point is removed, new point with higher likelihood drawn,
3. evidence is evaluated by assigning each removed point a prior volume and then computing the sum of the likelihood multiplied by the prior volume for each sample,

→ Points are moving to higher likelihood values → estimate of evidence volume by assuming that the entire remaining prior volume has a likelihood equal to the highest likelihood ”live point”.



Hyper-parameters, hierarchical modeling



- What if we want to study a population of sources?
- Example: what is the distribution of primary masses for binary black holes?
- The model for the population distribution is called the hyper-prior: $\pi(\boldsymbol{\theta}|\boldsymbol{\Lambda})$ where $\boldsymbol{\theta}$ are the original parameters, and $\boldsymbol{\Lambda}$ are the hyper-parameters
- Example: parameterize the primary mass distribution as a power law

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

Biscoveanu ODW3

Hyper-parameters, hierarchical modeling

The new likelihood is the original likelihood *marginalized* over the original parameters:

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

Hyper-prior

Original likelihood (doesn't depend on hyper-parameters)

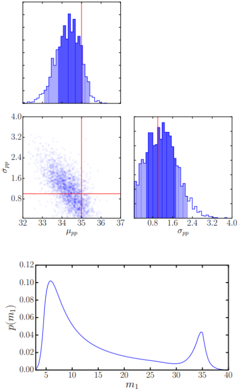
$$= \int d\boldsymbol{\theta}\,\mathcal{L}(d|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{\Lambda})$$

Original evidence

$$= \int d\boldsymbol{\theta}\,\frac{p(\boldsymbol{\theta}|d)\mathcal{Z}_{\theta}}{\pi_0(\boldsymbol{\theta})}\pi(\boldsymbol{\theta}|\boldsymbol{\Lambda})$$

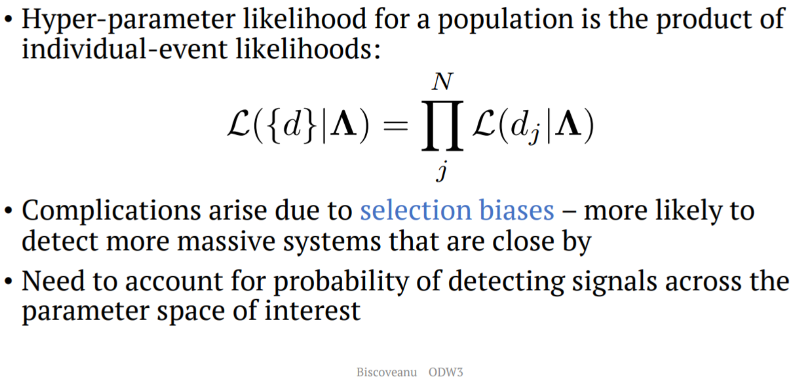
Biscoveanu ODW3

Original prior



(Talbot & Thrane, 2018)

Hyper-parameters, hierarchical modeling



- Hyper-parameter likelihood for a population is the product of individual-event likelihoods:

$$\mathcal{L}(\{d\}|\mathbf{\Lambda}) = \prod_{j}^{N} \mathcal{L}(d_j|\mathbf{\Lambda})$$

- Complications arise due to selection biases – more likely to detect more massive systems that are close by
- Need to account for probability of detecting signals across the parameter space of interest

Biscoveanu ODW3

Additional errors: calibration

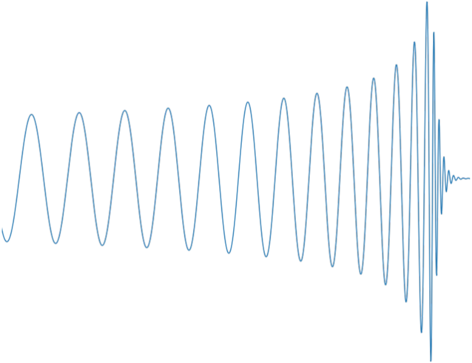
Detector’s calibration is yet another source of uncertainty:

$$p(d|\theta) \propto \exp\left[-\frac{1}{2}\sum_k \overset{\text{Noise-weighting}}{\boxed{\langle h_k(\theta)-d_k|h_k(\theta)-d_k\rangle}}\right]$$

$$h_k(\theta) \rightarrow \underbrace{\boxed{h_k(\theta)}}_{\text{Waveform}}\overset{\text{Calibration}}{\boxed{[1+\delta A_k]\exp\left[i\delta\phi_k\right]}}$$

Additional errors: calibration

$h' \rightarrow$



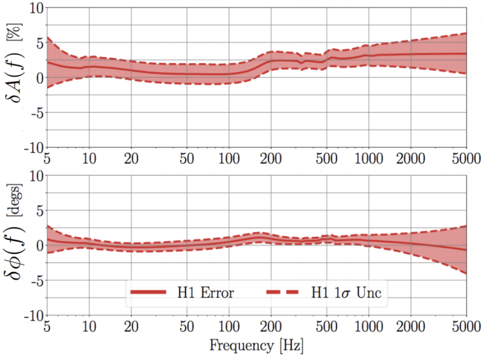
$+\delta h'$

$$h' \rightarrow h'(1 + \delta A)e^{i\delta\phi}$$

Marginalize over
an amplitude and
a phase uncertainty

Additional errors: calibration

$$h' \rightarrow h'(1+\delta A)e^{i\delta\phi}$$



$$\delta A(f) = p_s(f; \{f_i, \delta A_i\})$$

$$\delta\phi(f) = p_s(f; \{f_i, \delta\phi_i\})$$

Interpolate with cubic splines and marginalize over the calibration error

Cahillane+ (PRD:96, 102001)

Farr+ LIGO Document T1400682-v1

Fisher information matrix

An amount of information that observable data carries about an unknown parameter θ, specifically about the measurement errors of θ.

Fisher information matrix (FIM):

$$F_{ij} = -\left\langle \frac{\partial^2 \ln \mathcal{L}(d|\theta)}{\partial\theta_i \partial\theta_j} \right\rangle$$

with $\mathcal{L}(d|\theta) \propto \exp(-\frac{|d-\mu(\theta)|^2}{2\sigma^2})$

- ⋆ FIM quantifies information on θ_i including the correlations between θ_i and θ_j,
- ⋆ The inverse of FIM gives a lower bound on the covariance matrix for the parameters (asymptotically the covariance matrix):

$$F_{ij}^{-1} = \mathtt{Cov}_{ij}$$

- ⋆ In the high SNR regime, diagonal elements $\mathtt{Cov}_{ii}$ provide estimates for errors σ_i on parameters θ_i,
- ⋆ (see however `arXiv:gr-qc/0703086` for general case and complications).

Literature / resources

- ⋆ E. Thrane, C. Talbot, ”An introduction to Bayesian inference in gravitational-wave astronomy: parameter estimation, model selection, and hierarchical models”, `arXiv:1809.02293`
- ⋆ ”Probabilistic Programming & Bayesian Methods for Hackers”, `https://camdavidsonpilon.github.io/Probabilistic-Programming-and-Bayesian-Methods-for-Hackers`
- ⋆ J. Orloff, J. Bloom, ”Comparison of frequentist and Bayesian inference”, `https://ocw.mit.edu/courses/mathematics/18-05-introduction-to-probability-and-statistics-spring-2014/readings/MIT18_05S14_Reading20.pdf`
- ⋆ D. Wittman, ”Fisher Matrix for Beginners” `http://wittman.physics.ucdavis.edu/Fisher-matrix-guide.pdf`
- ⋆ A. Ly et al., ”A Tutorial on Fisher Information” `https://arxiv.org/pdf/1705.01064.pdf`
- ⋆ J. Skilling, ”Nested sampling for general Bayesian computation” `https://projecteuclid.org/euclid.ba/1340370944`
- ⋆ `Bilby`, a user-friendly Bayesian inference library `https://lscsoft.docs.ligo.org/bilby`
- ⋆ TensorFlow Probability `https://www.tensorflow.org/probability`