Statistical Power

Understanding the Universe with Bayesian Analysis

Natalie Williams & Hauke Koehn, Summer Semester 2025

Admin

- Please register on PULS
  - This course belongs to module PHY 756 Methods in Modern Astrophysics
- Lectures and seminars swap each week, Thursdays 12:15-13:45
- This is a new course (ran once before but not by me), so any feedback to help refine it is appreciated
- There is a moodle page for this course where you can find lecture slides and material
- This course will have problem sets which you are expected to work on them and actively participate during the class and seminar
- To pass this course you must achieve at least 50% on the problem sheets

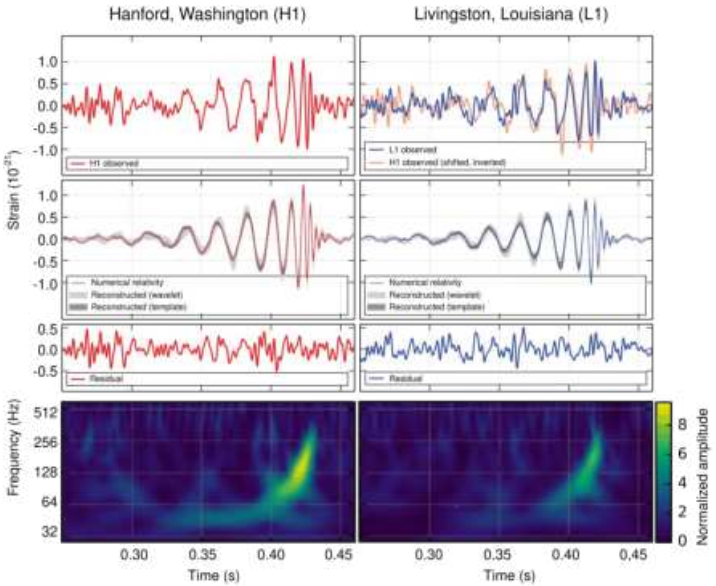
Course Outline

- Introduction
  - Frequentist vs Bayesian Statistics, Random Variables
- Frequentist Statistics
- Bayesian Statistics
- Examples in Astrophysics
- Computational Methods
- Applications in Astrophysics

About 4 hands-on problem sets will be given in seminars which must be completed to pass the course

Literature: *"Making Sense of Data"* from the International Max Planck Research School at the Max Planck Institute for Gravitational Physics

Why should I care about Bayesian Statistics?



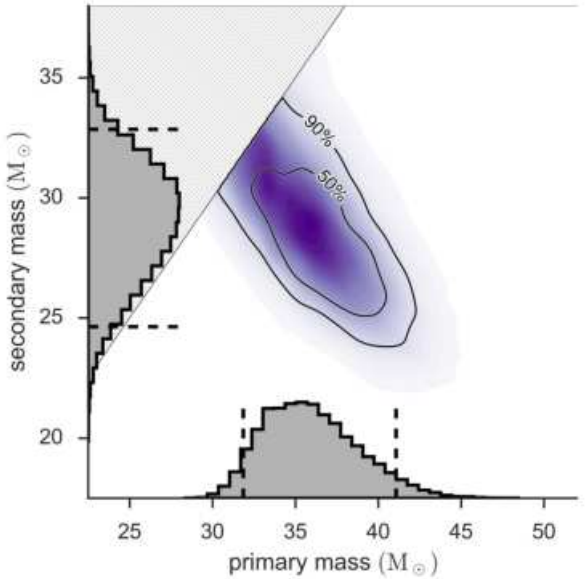
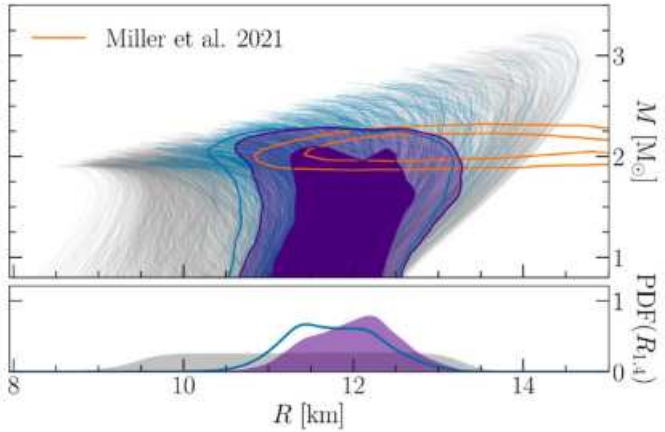


Image: Ligo-Virgo Collaborations

Inferring astrophysical source properties

Why should I care about Bayesian Statistics?

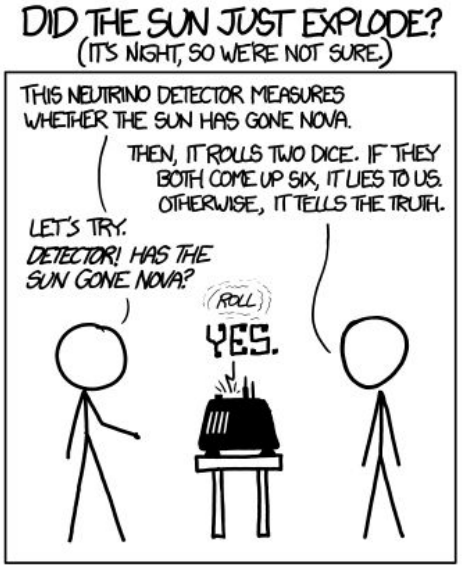


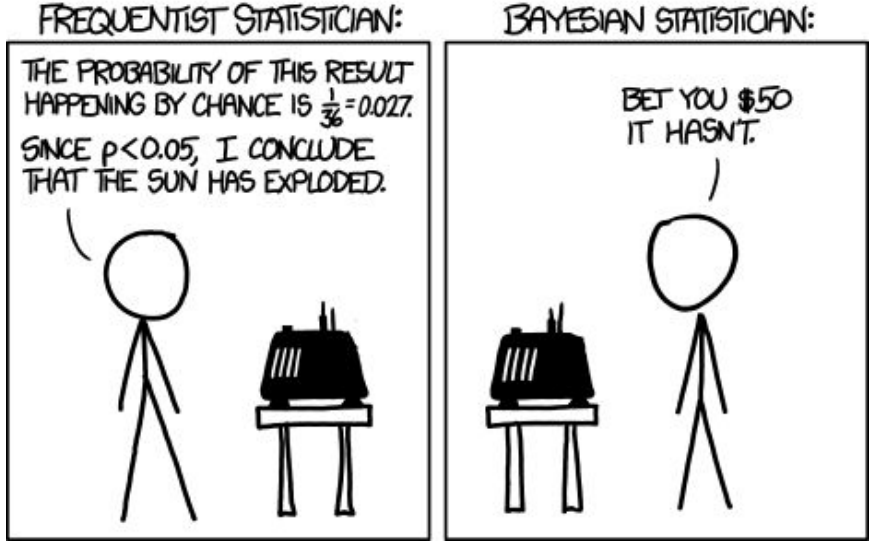
Pang et al., Astrophys.J. 922 (2021) 1, 14

Inferring astrophysical source properties

1. Introduction

1.1 Bayesians vs Frequentists





What *is* probability?

Frequentist

"Probability is defined by the limit of an outcome's relative frequency as the number of trials approaches infinity."

Bayesian

"Probability is defined by the degree of belief you have for an outcome"

Right…what's the difference?

Flipping a coin

Assume we have a fair coin - this is our *null hypothesis*. What is the probability that this coin will land on heads when flipped?

Frequentist : *"Over many, many coin flips, 50% will land on heads"* ✅

Bayesian: *"I believe there is a 50% chance that any coin flip will land on heads"* ✅

Aren't these the same statement?

Ok, now I'm going to flip it, take a look at the result and cover it with my hand so you can't see. **What is the probability *now* of this coin having landed on heads?**

Do the statements from the Frequentist and the Bayesian still apply?

Frequentist ❌ Bayesian ✅

Flipping a coin

Frequentist : *“This question is meaningless, the coin has already been flipped! If it’s heads the probability is 100%, and if it’s tails it’s 0%. There’s one correct answer - I just don’t know what that answer is. Probability has nothing to do with it!“*

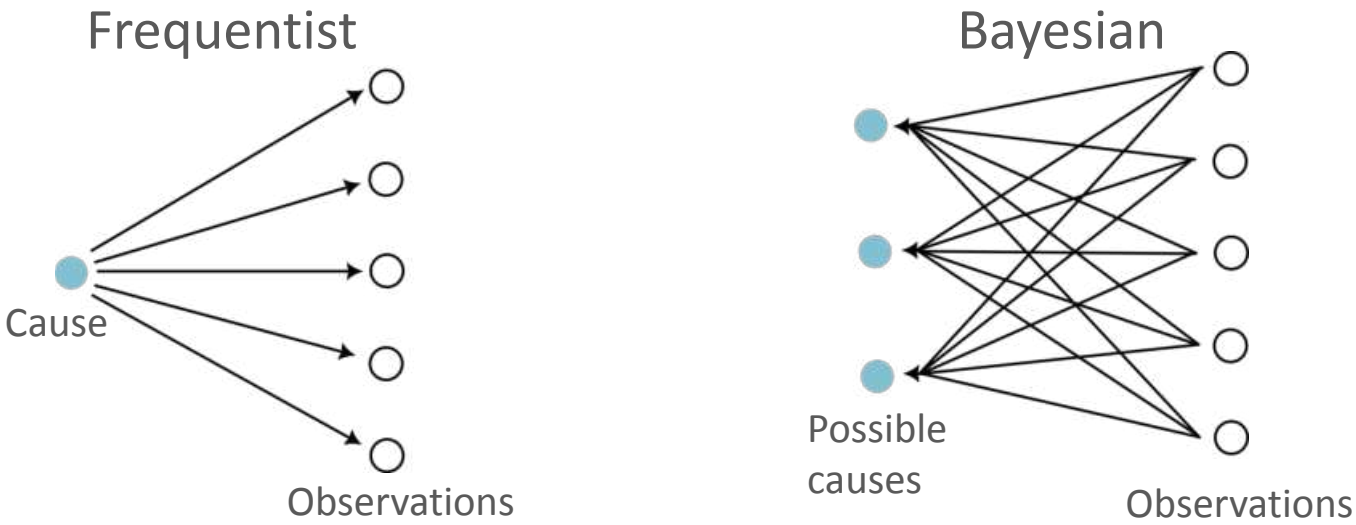
Objective

Bayesian: *“From my perspective, there is a 50% chance that this coin has landed on heads.“*

Subjective

Either there is an unknown, but objective right answer for us to find *or* there is no single right answer but a variety of possible answers with varying probabilities.

Parameters as random variables



Note: This doesn't imply there are multiple actual outcomes — the coin has landed one way. Rather, we consider different possible underlying causes (e.g., whether the coin is fair), each assigned a probability. This allows us to represent uncertainty and incorporate prior beliefs about the coin's fairness.

Revealing the coin

Frequentist : *“The result was heads/tails this time.“*

Data driven

Bayesian: *“I am now 100% confident that this coin has landed on heads/tails.“*

Model driven

The Frequentist adds this data as a point in their dataset, and can compute a statistic such as a p-value to compare this result to our null hypothesis, and either reject or fail to reject the hypothesis.

The Bayesian used the observed data to update their *prior beliefs* into a probability statement about the null (and/or any alternative) hypothesis.

Approach	Frequentist	Bayesian
What is the assumption?	The parameters I want to estimate are fixed	There's a probability distribution around the parameters I want to estimate
What does it ask?	How likely is my data given my hypothesis?	How likely is my hypothesis given my data?
What do I need?	A stopping criterion A predetermined experimental design	A prior A data set
What is the output?	A point estimate (p-value)	A probability of the hypothesis
Advantages	Simple and easy to use Widely accepted No prior needed	You measure the evidence which incorporates Occam's razor
Disadvantages	Significance depends on sample size Only gives a yes or no answer	Priors can be subjective and effect the result Requires more advances statistics
When to use it	When you have a large amount of data	When you have limited data When you have priors When you have enough computing power

~~Choose a side!~~

Bayesian and Frequentist statistics are both valid forms or statistics, just used in different circumstances. Often they can be complementary to one another.

Bayesian statistics for astrophysics

Bayesian statistics is particularly useful for astrophysics such as gravitational wave astronomy and multi-messenger astronomy because:

- Single events are analysed that cannot be repeated
- Systems have unknown source parameters which we can consider as random variables
- Information from various data streams and experiments can be used to update our beliefs

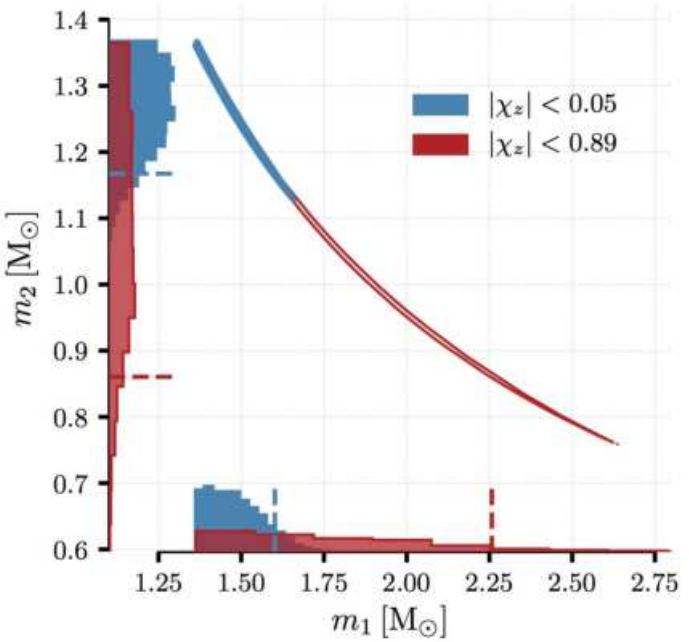


Image: Ligo-Virgo Collaborations

Bayesian statistics elsewhere

Medicine and Healthcare: *"What is the probability that a patient has a certain disease, given the test results, the disease prevalence, and the accuracy of the test?"*

Genetics and Bioinformatics: *"Based on observed genetic sequences, what's the most probable evolutionary relationship between species X and species Y?"*

Machine Learning and AI: *"How likely is it that a new email is spam, given the presence of certain keywords in the message and prior classifications of similar emails?"*

Economics and Finance: *"How likely is it that the stock market will rise or fall tomorrow, given recent trends, economic indicators, and historical data?"*

Political forecasting: *"What is the probability that party A will win a majority in the next election, considering current voting preferences and past election results?"*

1.2 Random Variables

We consider systems which are in reality not (or effectively not) deterministic due to things such as measurement uncertainty and sensitivity to initial conditions.

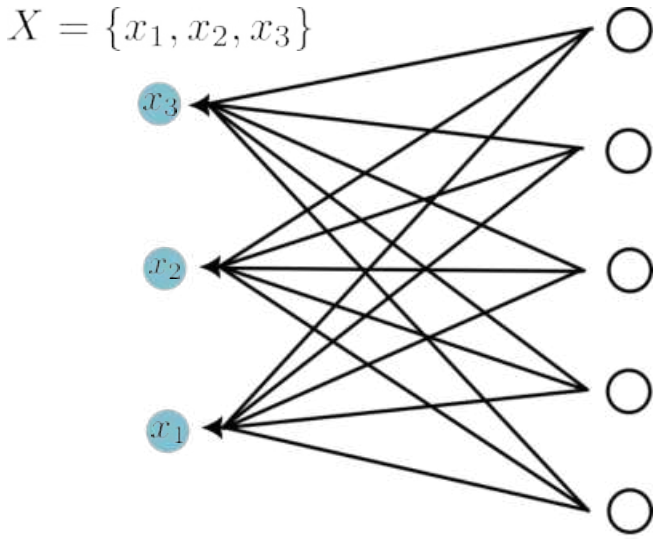
This in-deterministic nature is encoded within the concept of *random variables*.

Random Variables & Probability Distribution: A *random variable*, X, is a quantity that, when observed, can take one of a (possibly infinite) number of values. Before making a measurement, the value of the random variable cannot be predicted with certainty but only with a given probability. The relative frequency of the outcomes over many experiments can be described by a *probability distribution*. The value that X takes in a particular observation (or experiment), x_i, is called a *realization* of the random variable.

1.2.1 Discrete Random Variables

A discrete random variable X can take on any of a (possibly infinite, but countable) set of possible values which together compromises a *sample space*. The probability that X takes any particular value is represented by a *probability mass function* (pmf), which is a set of numbers p_i with properties

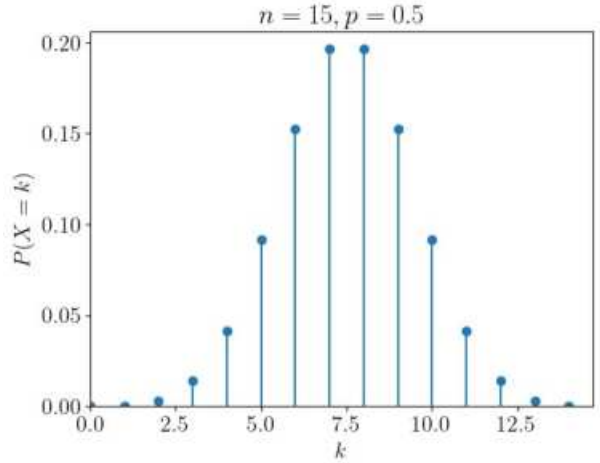
$$0 \leq p_i \leq 1 \quad \text{and} \quad \sum_i p_i = 1$$



Example - Binomial Distribution

The *Binomial distribution* is the distribution of the number of success in n trials for which the probability of success in one trial is p. We use notation $X \sim B(n,p)$ and

$$P(X = k) = p_k = \begin{cases} \binom{n}{k} p^k (1-p)^{n-k} & \text{if} \quad k \in 1, \ldots, n \\ 0 \quad \text{otherwise} \end{cases}$$



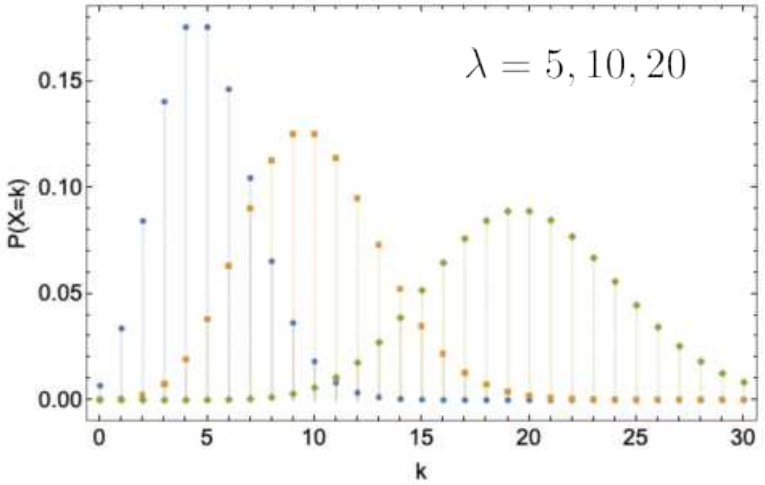
For n=1 this is the *Bernoulli distribution* and the binomial distribution is the sum of n Bernoulli distributions

Applications: counting problems (e.g. distribution of events in categories or time)

Example - Poisson Distribution

The *Poisson distribution* determines the number of occurrences of *independent events* in a time interval which occur at *rate* λ.

$$P(X = k) = p_k = \begin{cases} \lambda^k e^{-\lambda}/k! & \text{if} \quad k \in 0, 1, \ldots \\ 0 & \text{otherwise} \end{cases}$$



The Poisson distribution is the limiting distribution of $B(n,p)$ for $n \to \infty$ and $p \to 0$, but $np = \lambda$

Applications: distribution of number of events in a population (e.g. gravitational wave events)

1.2.1 Continuous Random Variables

A continuous random variable X can take on any value within some continuous range or set of ranges which together compromises a *sample space* $\boldsymbol{X}$. The probability that X takes any particular value is represented by a *probability density function* (pdf), $p(x)$.

The probability the X takes a value in the range x to $x + dx$ is $p(x)dx$. The pdf

$$\int_{x \in \mathcal{X}} p(x)\mathrm{d}x = 1 \qquad 0 \leq p(x) \leq 1 \text{ for all } x \in \mathcal{X}$$

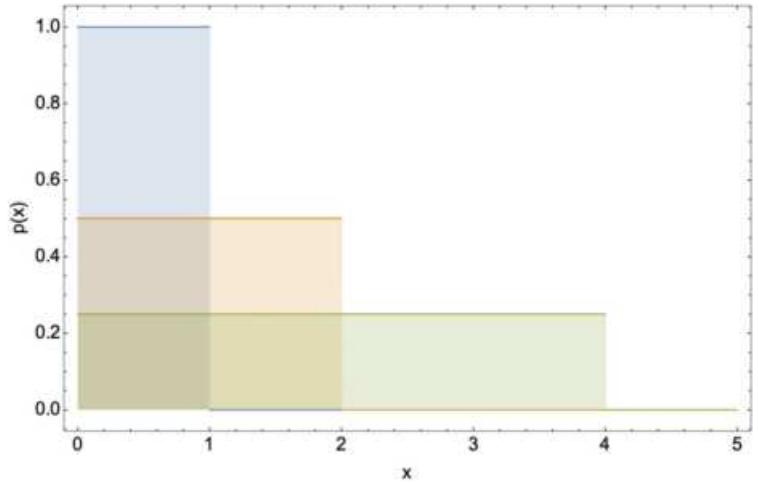
For single valued random variables with non-disjoint sample spaces, continuous random variables may also be characterised by the *cumulative distribution function* (CDF)

$$P(X \leq x) = \int_{-\infty}^{x} p(x)dx$$

Example - Uniform distribution

If X is uniform on an interval (a,b), denoted $X \sim U[a,b]$, then the pdf is given by

$$p(x) = \begin{cases} \frac{1}{b-a} & \text{if} \quad x \in [a,b] \\ 0 & \text{otherwise} \end{cases}$$



Applications: used as an “uninformative prior in parameter estimation”

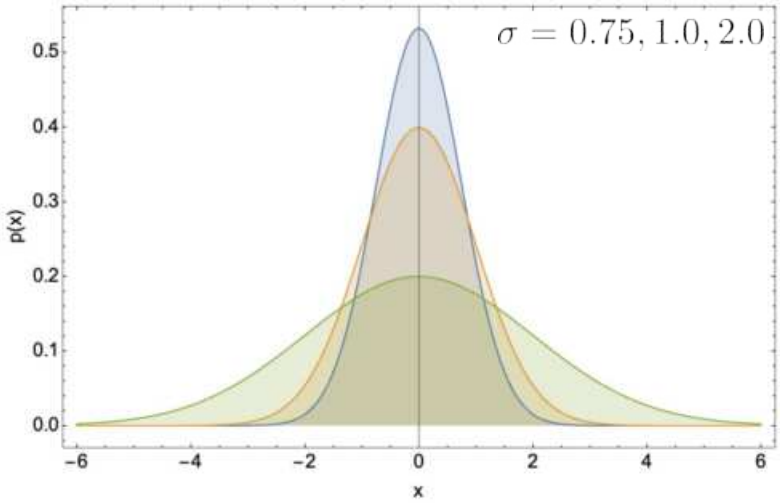
Example - Normal distribution

X is Normal with *mean* μ and *variance* σ^2 denoted as $X \sim N(\mu, \sigma^2)$ if the pdf follows

$$p(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

If μ=0 and σ^2 = 1 X follows a *standard normal distribution.*

A *chi-squared distribution* is the sum of the squares of k standard normal distributions

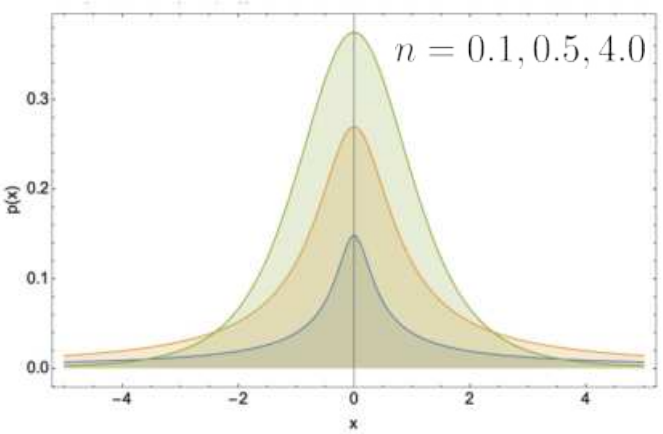


Applications: distribution of noise fluctuations
standard distribution for measurement uncertainty

Example - Student's t-Distribution

X follows a Student's t-distribution with n>0 degrees of freedom $X \sim t_n$ if it has pdf

$$p(x) = \frac{\Gamma((n+1)/2)}{\sqrt{n\pi}\,\Gamma(n/2)} \left(1 + \frac{x^2}{n}\right)^{\frac{n+1}{2}}$$



The Student t-distribution is the distribution of the ratio between a standard Normal distribution to the square root of an independent chi-square distribution, normalized by the degrees of freedom. If

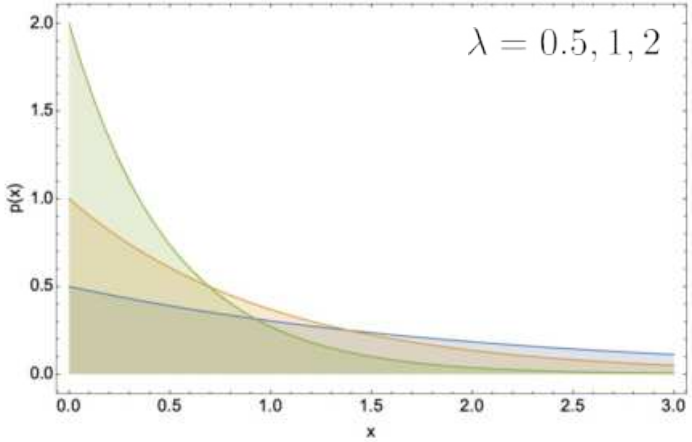
$$X \sim N(0,1) \quad \text{and} \quad Y \sim \chi^2_n, \quad \text{then} \quad X/\sqrt{Y/n} \sim t_n$$

Applications: arises when marginalising over uncertainty in power-spectral density information

Example - Exponential Distribution

X is exponential with rate λ if $X \sim E(\lambda)$ has pdf of the form

$$p(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if} \quad x > 0 \\ 0 & \text{otherwise} \end{cases}$$



The exponential distribution is the distribution of the time that elapses between successive events of a Poisson process.

Applications: distribution of time lag between gravitational wave events

1.2.3 Properties of Random Variables

The pdf (or pmf) contains all the information of a random variable. However it's often more useful to quantify properties of the pdf to interpret the distribution. These properties all rely on the notion of an *expectation value*, defined for function $T(X)$ as

$$\mathbb{E}(T(X)) = \sum_{i=1}^{\infty} p_i t(x_i)$$

Discrete

$$\mathbb{E}(T(X)) = \int_{-\infty}^{\infty} p(x)t(x)\mathrm{d}x$$

Continuous

Mean:

The mean is simply the expectation value

$$\mu = \mathbb{E}(X)$$

Median:

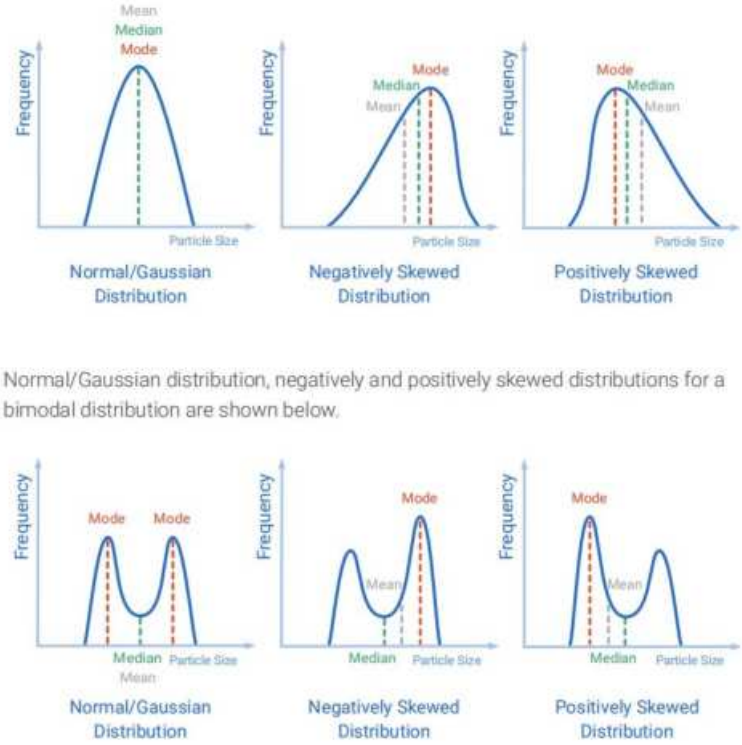
The median is the central value of the distribution

$$\sum_{i:x_i<m} p_i < 0.5 \quad \text{and} \quad \sum_{i:x_i\leq} p_i \geq 0.5. \qquad \int_{-\infty}^{m} p(x)\mathrm{d}x = \int_{m}^{\infty} p(x)\mathrm{d}x = \frac{1}{2}$$

Mode:

The most probably value of the random variable in the distribution (may not be unique!)

$$M = \mathrm{argmax}_{i\in\mathcal{X}}(p_i) \text{ or } M = \mathrm{argmax}_{x\in\mathcal{X}} p(x)$$



Variance:

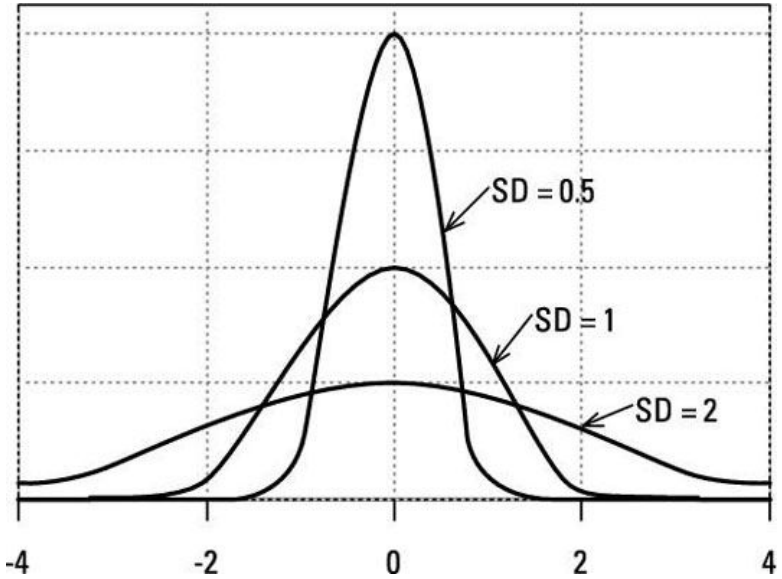
The variance is the expectation value of the squared distance from the mean

$$\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}(X))^2]$$

Standard deviation:

Simply the square root of the variance

$$\sigma = \sqrt{\sigma^2} = \sqrt{\mathrm{Var}(X)}$$

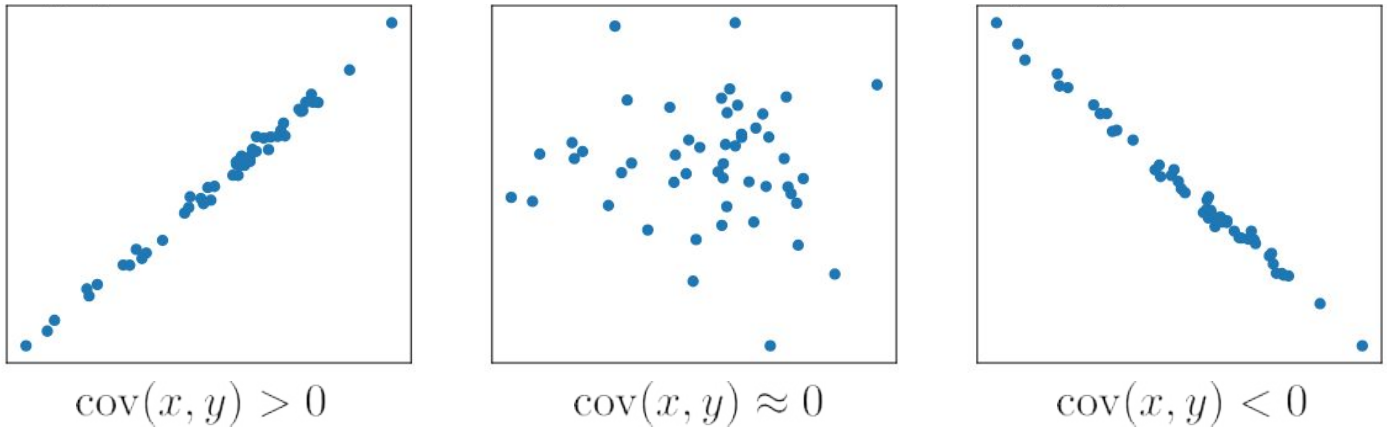


Q: Why do we use both of these quantities when they're so closely related?

Covariance:

When considering two random variables, X and Y, the covariance is the expectation value of the product of their distance from their respective means

$$\operatorname{cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))]$$



$\operatorname{cov}(x, y) > 0$ $\operatorname{cov}(x, y) \approx 0$ $\operatorname{cov}(x, y) < 0$

Skewness:

With a given mean and variance the skewness is

$$\gamma_q = \mathbb{E}\left[\left(\frac{x-\mu}{\sigma}\right)^3\right]$$

Kurtosis:

The kurtosis is a measure of tailedness of a distribution and is defined as

$$\operatorname{Kurt}(X) = \mathbb{E}\left[\left(\frac{x-\mu}{\sigma}\right)^4\right]$$

Higher Moments:

The nth moment about a reference value of a probability is

$$\mathbb{E}\left[(x-c)^n\right]$$

The value of c is usually the mean, what is the exception presented here?

1.2.4 Moment generating functions

A neat trick for computing these quantities for a distribution is the *moment generating function* defined as

$$M_X(t) = \mathbb{E}\left[\mathrm{e}^{tX}\right] \quad t \in \mathbb{R}.$$

This is useful as the nth derivative of this function with respect to t evaluated at t=0 gives the nth moment.

For example this is the moment generating function for a normal distribution

$$M_X(t) = e^{(\mu t + \frac{1}{2}\sigma^2 t^2)}$$

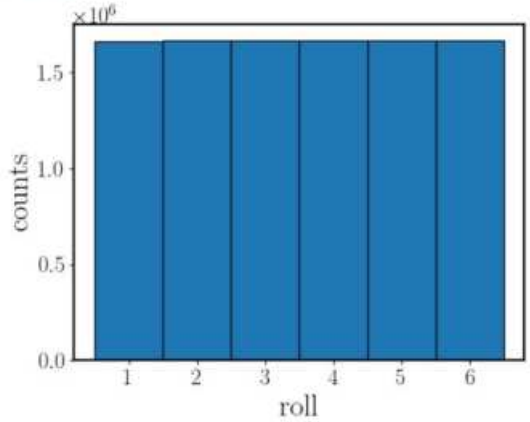
Examples:

1. Rolling a dice
2. A uniform distribution in (0.5, 6.5) with $t(x)=x$
3. A uniform distribution in (0.5, 6.5) with $t(x)=x^2$
4. A normal distribution N(3,2) with $t(x)=x^2$

Rolling a dice

```
[37]: dice = [1,2,3,4,5,6]
      probs = [1/6,1/6,1/6,1/6,1/6,1/6]

[121]: rolls = np.random.choice(dice, size = 10000000, p=probs)
       plt.hist(rolls, bins=np.arange(0.5, 7.5))
       plt.xlabel('roll')
       plt.ylabel('counts')
       plt.xticks(dice)
       plt.show()
```



1.2.5 Independance

In terms of the pdf (or pmf), the random variables are said to be independent if they joint distribution can be separated such as

$$p(x_1, ..., x_N) = p_{X_1}(x_1) \cdot p_{X_2}(x_2) \cdot ... \cdot p_{X_N}(x_N)$$

Independence implies that the covariance is 0, but the opposite is not necessarily true. A set of variables $\{X_i\}$ are called *independently identically distributed* if they are all independent and have the same probability distribution.

1.2.6 Linear Combinations of random variables

Suppose $X_1, \ldots, X_n$ are random variables and consider

$$Y = \sum_{i=1}^{N} a_i X_i$$

For any set of variables

$$\mathbb{E}(Y) = \sum_{i=1}^{N} a_i \mathbb{E}(X_i), \qquad \mathrm{Var}(Y) = \sum_{i=1}^{N} a_i^2 \mathrm{Var}(X_i) + \sum_{i \neq j} a_i a_j \mathrm{cov}(X_i, X_j).$$

If these variables are independent this reduces to

$$\mathrm{Var}(Y) = \sum_{i=1}^{N} a_i^2 \mathrm{Var}(X_i)$$

A commonly used linear combination of random variables is the *sample mean* of a set of independently identically distributed random variables

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} X_i \qquad \text{where} \qquad \mathbb{E}(\hat{\mu}) = \mathbb{E}(X_1), \qquad \mathrm{Var}(\hat{\mu}) = \frac{1}{n} \mathrm{Var}(X_1), \qquad M_{\hat{\mu}}(t) = \left(M_{X_1}\left(\frac{t}{N} \right) \right)$$

1.2.5 Law of large numbers

Suppose that $X_1, \ldots, X_n$ is a sequence of independently identically distributed random variables, each having finite mean μ and variance σ^2. We denote the sum of the random variables

$$S_n = \sum_{i=1}^{n} X_i, \quad \text{which implies } \mathbb{E}(S_n) = n\mu, \quad \mathrm{Var}(S_n) = n\sigma^2.$$

The law of large numbers tells us that the sample mean becomes increasingly concentrated around the mean of the random variable as the number of samples tends to infinity.

$$P\left(\left|\frac{S_n}{n} - \mu\right| > \epsilon\right) \to 0, \text{ as } \quad n \to \infty$$

Weak law of large numbers (ε>0)

The sample mean converges in probability

$$P\left(\frac{S_n}{n} \to \mu\right) = 1.$$

Strong law of large numbers

The sample mean converges almost surely

1.2.5 Central limit theorem

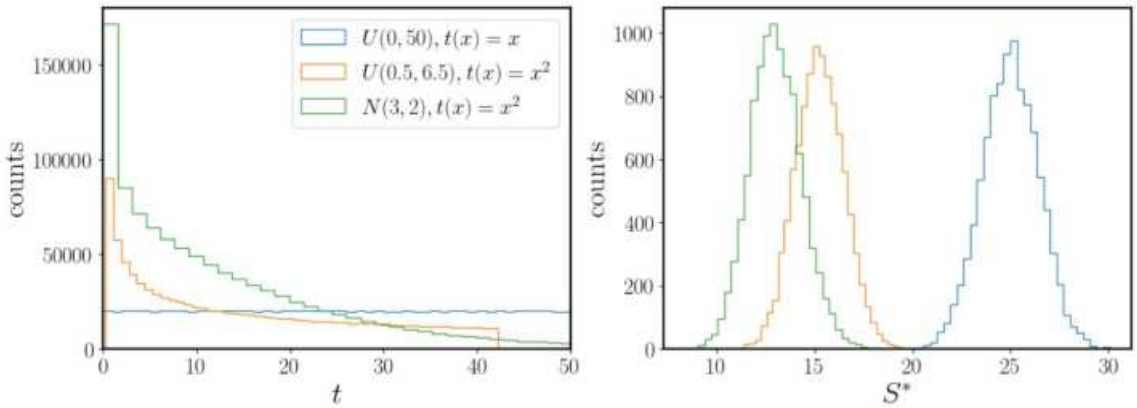
It is often assumed that data generating processes are Normally distributed. Why?

This is a consequence of the *Central Limit Theorem*, which states that the standardised sample mean S_n^* is approximately normal in the limit $n \to \infty$

$$S_n^* = \frac{S_n - n\mu}{\sigma\sqrt{n}}$$

1.2.5 Central limit theorem

This means that regardless* of the form of our pdf (or pmf), the distribution of sample means will approach a normal distribution with increased sample means



*with a few exceptions…

2. Frequentist Statistics

Recap - Basic principles of frequentist statistics

In frequentist statistics probability is defined as…

the limit of an outcomes frequency over (infinitely) many trials

In frequentist statistics variables are treated as…

fixed but unknown

The output of frequentist statistics is

the confidence of a hypothesis given the data

Frequentist statistics is useful in situations with…

a large amount of data

2.1 Statistic, Estimator, and Likelihood

Statistic: A statistic is a random variable or random vector $T = t(X)$ which is a function of X, but does not depend on the parameters of the distribution, $\boldsymbol{\theta}$.

Its realized value is $t = t(\mathbf{x})$, i.e., a statistic is a function of observed data only, not the unknown parameters. A statistic might also be used to provide an upper or lower limit for a confidence interval on the value of a parameter or to evaluate the validity of a hypothesis in hypothesis testing.

Estimator: An estimator is a statistic used to estimate the value of a parameter.

Typically the random vector would be a set of independent identically distributed random variables, $X_1,\dots, X_n$ with pdf $p(x|\Theta)$. A function $\Theta'(X_1,\dots, X_n)$ of $X_1,\dots, X_n$ used to infer the parameter values is called an estimator of Θ'. Here, Θ' is a random variable with a sampling distribution in this latter context. The value of the estimator at the observed data Θ' is called an estimate of Θ'.

2.1 Statistic, Estimator, and Likelihood

Likelihood: If an event E has a probability which is a specified function of parameters $\boldsymbol{\Theta}$, then the likelihood of E is $P(E|\boldsymbol{\Theta})$, i.e. the probability of an event under the given data.

This means that the likelihood denoted as $L(\boldsymbol{\Theta};\mathbf{x})$ is functionally the same as the pdf of the generating process, but the likelihood is regarded as a function of the parameters Θ with fixed observed data $\mathbf{x}$*, while the pdf is regarded as a function of the observed data,* $\mathbf{x}$, with fixed parameters Θ*.*

$$L(\vec{\theta};\mathbf{x}) = p(\mathbf{x};\vec{\theta})$$

Often we work with the log-likelihood for simplicity

$$l(\vec{\theta};\mathbf{x}) = \ln[L(\vec{\theta};\mathbf{x})] = \ln[p(\mathbf{x};\vec{\theta})]$$

For a set of independent identically distributed samples $\mathbf{x}=\{x_1, x_2,\ldots,x_n\}$ the joint likelihood is

$$L(\vec{\theta},\mathbf{x}) = \prod_{i=1}^{n} p(x_i;\vec{\theta}) \Rightarrow l(\vec{\theta},\mathbf{x}) = \sum_{i=1}^{n} l(x_i;\vec{\theta})$$

Example - Poisson distribution

Recall $$p(X = k; \lambda) = \begin{cases} \lambda^k e^{-\lambda}/k! & \text{if} \quad k \in 0, 1, ... \\ 0 & \text{otherwise} \end{cases}$$

Then the joint likelihood function is

$$L(\lambda; k_1, k_2, ..., k_n) = \lambda^{k_1} e^{-\lambda}/k_1! \cdot \lambda^{k_2} e^{-\lambda}/k_2! \cdot ... \cdot \lambda^{k_n} e^{-\lambda}/k_n!$$

$$\Longrightarrow \quad L(\lambda; \mathbf{k}) = \lambda^{\sum_{i=1}^{n} k_i} e^{-n\lambda} / \prod_{i=1}^{n} k_i!$$

And the log likelihood is $$l(\lambda; \mathbf{x}) = -n\lambda + \left(\sum_{i=1}^{n} k_i \right) \ln \lambda - \ln \left(\prod_{i=1}^{n} k_i! \right)$$

Estimators

Estimator: An estimator is a statistic used to estimate the value of a parameter.

Maximum likelihood estimator:

The value of parameters which maximises the likelihood $\hat{\theta} = \text{argmax}_{\theta \in \vec{\theta}} L(\theta; \mathbf{x})$

Sample Mean:

Used to estimate mean $\hat{\mu} = \frac{1}{n} \sum_{j=1}^{n} x_j$

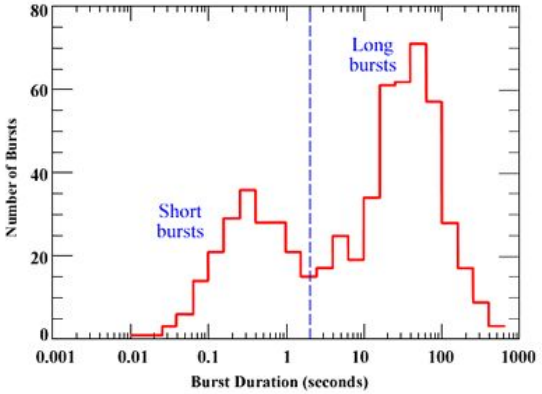
Sample Variance:

Used to estimate variance $\hat{\sigma}^2 = \frac{1}{n-1} \sum_{j=1}^{n} (x_j - \hat{\mu})$

Example - Maximum likelihood estimator

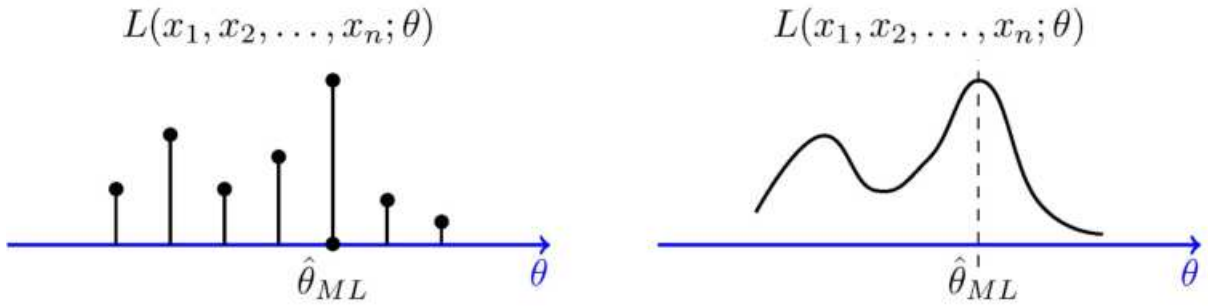
You have access to Fermi data for 1 week, and observe 6 gamma ray bursts (GRBs). You classify 1 of them as short GRBs (prompt emission <2 seconds), and the others as long GRBs. Use the maximum likelihood estimator to suggest the fraction of short GRBs.

What distribution seems wise to use in this case?



On average about 240 GRBs are seen per year and a bit more than ⅙ of them are sGRBs - this demonstrates for small sample sizes the maximum likelihood estimator can be biased

Example - Maximum likelihood estimator



Note that the maximum likelihood estimator is not the quantity you are interested in, in particular, if the distribution is not very symmetric the ML differs a lot from the mean or median value - this can be quantified with the *bias*.

The bias

The *bias* of an estimator of a parameter measures the difference between the mean value and the value of the parameter being estimated. An estimator $\hat{\theta}$ is *unbiased* if

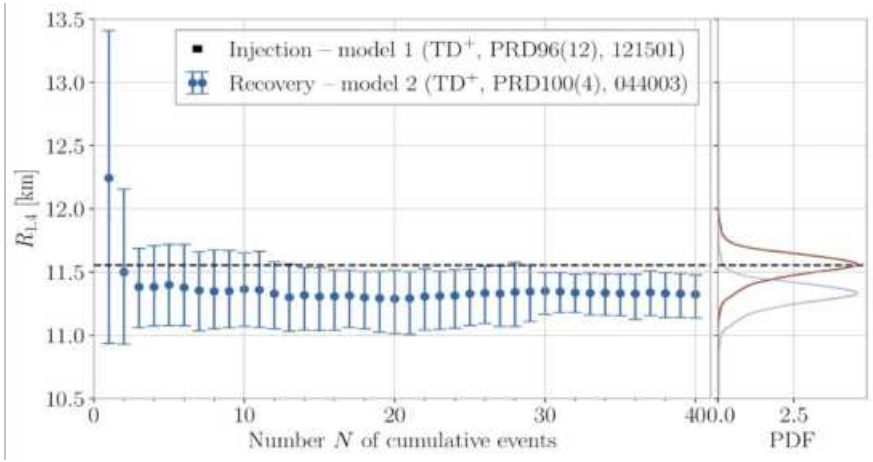
$$\mathbb{E}(\widehat{\theta}) = \theta.$$

Otherwise the estimator is biased with bias function

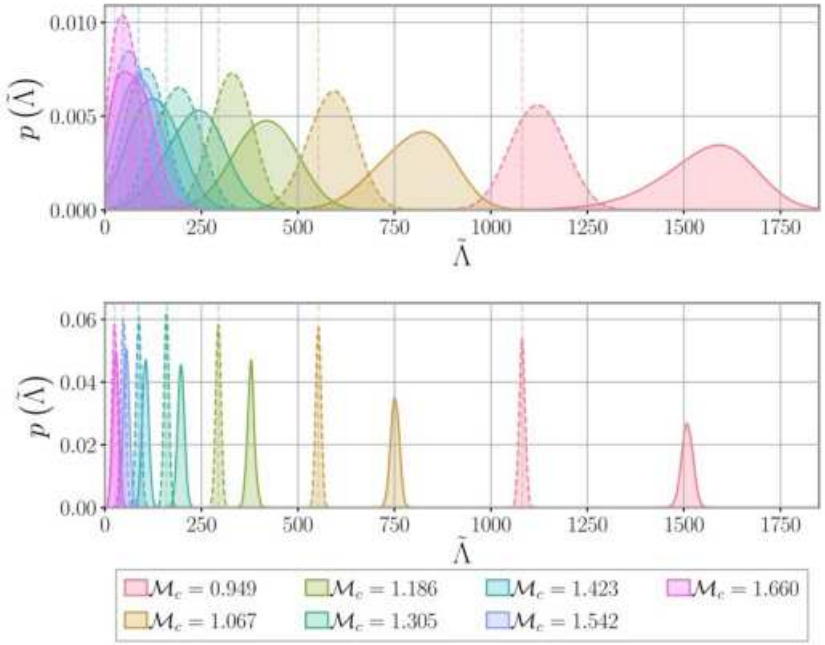
$$\text{bias}(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta$$

Is the estimator for the binomial distribution we used for the GRB calculation biased? What about the our observation, do we observe a bias?

Examples - GW model uncertainty



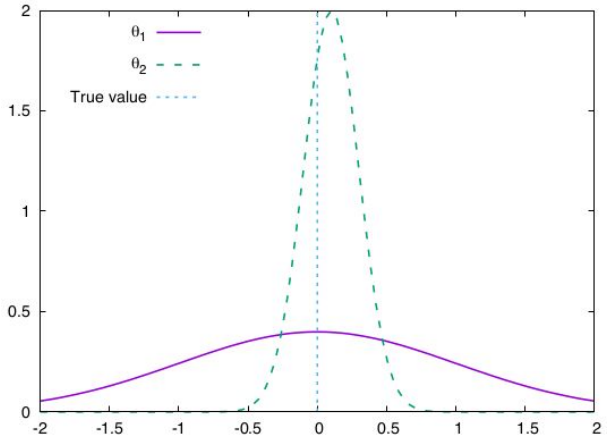
Study by PhD student N. Kunert



Pratten, Schmidt, Williams 2022,
Phys. Rev. Lett. 129, 081102

The bias

An unbiased estimator is not necessarily a deal breaker!



Which estimator would you rather use here?

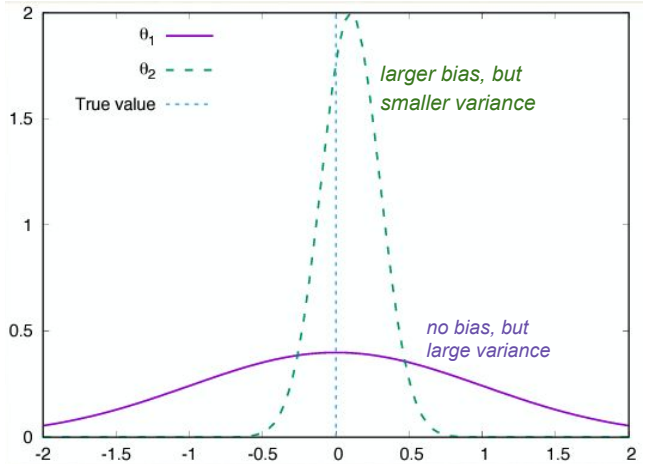
We call the estimator *asymptotically unbiased* if $\mathbb{E}(\hat{\theta}) \to \theta \quad \text{for } n \to \infty$

The mean square error

Another possibility to quantify the unbiasedness is the estimate of the mean square error

$$\mathrm{MSE} = \mathbb{E}[(\theta - \theta)^2] = \mathrm{var}(\hat{\theta}) + \mathrm{bias}(\hat{\theta})^2$$

It will depend on the exact problem to either use an estimator with a smaller bias or an estimator with a smaller mean square error.



Consistency

If an estimator is *consistent*, it becomes increasingly concentrated around the true value, when the number of estimations increases

$$\mathbb{P}(|\hat{\theta} - \theta| > \epsilon) \to 0 \qquad \text{for } n \to \infty; \text{ for any } \epsilon > 0$$

$$\textit{If } \text{var } (\widehat{\theta}) \to 0 \textit{ and } \text{bias}(\widehat{\theta}) \to 0 \textit{ as } n \to \infty, \textit{ then } \widehat{\theta} \textit{ is (weakly) consistent.}$$

The *efficiency* of an estimator is the ratio of the minimum possible variance to the variance of the estimator.

2.2 Sufficient Statistics

Sufficient Statistics: A statistic is *sufficient* if it contains the same information as the full dataset

A statistic S (recall this just means a function of observed data only) is sufficient for all $\vec{\theta}$ if the distribution of $\boldsymbol{X}$ given S, $p_{\mathbf{X}|S}(\mathbf{X}|s,\vec{\theta})$, does not depend on

In simple terms, if I know statistic S, I don't need to know the original data to make inferences about $\vec{\theta}$, all the information is contained within S

A full set of observations $\boldsymbol{X}$ is always sufficient, however often there are sufficient statistics with lower dimensionality

Sufficient statistics that achieve the smallest reduction in size of the data are called *minimally sufficient*.

2.2 Sufficient Statistics

Neyman Factorisation Theorem: Let $\boldsymbol{X}=(X_1,\ldots,X_n)\sim p(\mathbf{x}|\theta)$. Then the statistic $s=s(X_1,\ldots,X_n)$ is sufficient for θ if there exists functions h of $\mathbf{x}$ and g of (s,θ) such that

$$p(\mathbf{x} \mid \vec{\theta}) = L(\vec{\theta}; \mathbf{x}) = g(s(\mathbf{x}), \vec{\theta})h(\mathbf{x})$$

Proof:

$$\begin{aligned}
L(\vec{\theta}; \mathbf{x}) = p_{\mathbf{X}}(\mathbf{x}|\,\vec{\theta}) &= \mathbb{P}(\mathbf{X} = \mathbf{x}|\,\vec{\theta}) \\
&= \mathbb{P}(\mathbf{X} = \mathbf{x}\,\&\,S = s(\mathbf{x}) \mid \vec{\theta}) \\
&= \mathbb{P}(\mathbf{X} = \mathbf{x}|S = s(\mathbf{x}), \vec{\theta})\,\mathbb{P}(S = s(\mathbf{x})|\,\vec{\theta}) \\
&= \mathbb{P}(\mathbf{X} = \mathbf{x}|S = s(\mathbf{x}))\,\mathbb{P}(S = s(\mathbf{x})|\,\vec{\theta}) \quad [\text{since } S \text{ is sufficient}] \\
&= h(\mathbf{x})g(s(\mathbf{x}), \vec{\theta}).
\end{aligned}$$

2.2 Sufficient Statistics

Neyman Factorisation Theorem: Let $\boldsymbol{X}=(X_1,\ldots,X_n)\sim p(\mathbf{x}|\theta)$. Then the statistic $s=s(X_1,\ldots,X_n)$ is sufficient for θ if there exists functions h of $\mathbf{x}$ and g of (s,θ) such that

$$p(\mathbf{x} \mid \vec{\theta}) = L(\vec{\theta}; \mathbf{x}) = g(s(\mathbf{x}), \vec{\theta})h(\mathbf{x})$$

Example in gravitational wave astronomy:
The usual likelihood for observed gravitational-wave data takes the form

$$\mathcal{L}_{\rm GW} \propto \exp\left(-\frac{1}{2}\langle d-h(\vec{\theta})|d-h(\vec{\theta})\rangle\right) \qquad \langle a|b\rangle = 4\Re\int_{f_{\rm low}}^{f_{\rm high}} \frac{\tilde{a}(f)\tilde{b}^*(f)}{S_n(f)}df$$

For many waveform families it is possible to find a reduced basis that can be used to reconstruct all the waveforms in the family

$$h(t;\vec{\theta}) = \sum_{i=1}^{M} a_i(\vec{\theta})h_i(t)$$

The overlaps of the basis waveforms with the data, $\langle h_i|d\rangle$, are sufficient statistics for inferring the waveform parameters

2.3 Cramer Rao-bound

The Cramer-Rao bound determines the smallest achievable variance of an estimator. Let $X_1, \ldots, X_n$ denote a random sample from $p(\mathbf{x}|\theta)$ and suppose that $\hat{\theta}$ is an estimator for θ. With the following regularity conditions

1. $\forall \theta_1, \theta_2 \in \Theta$ such that $\theta_1 \neq \theta_2$, $p(x \mid \theta_1) \neq p(x \mid \theta_2)$ [identifiability]. *Each parameter gives a unique distribution*
2. $\forall \theta \in \Theta$, $p(x \mid \theta)$ have common support. *The set of possible x values must not depend on θ*
3. Θ is an open set. *Ensures we can differentiate on θ*
4. $\exists \partial p(x \mid \theta)/\partial\theta$. *Ensures that the pdf is differentiable wrt θ*
5. $\mathbb{E}\left(\partial \log p(\mathbf{X}|\theta)/\partial\theta\right)^2 < \infty$. *Ensures that the variance of an estimator is finite*

$$\operatorname{var}(\widehat{\theta}) \geq \frac{\left(1 + \frac{\partial b}{\partial \theta}\right)^2}{I_\theta} \quad \text{with} \quad I_\theta = \mathbb{E}\left[\left(\frac{\partial \ell}{\partial \theta}\right)^2\right]$$

The Fisher information

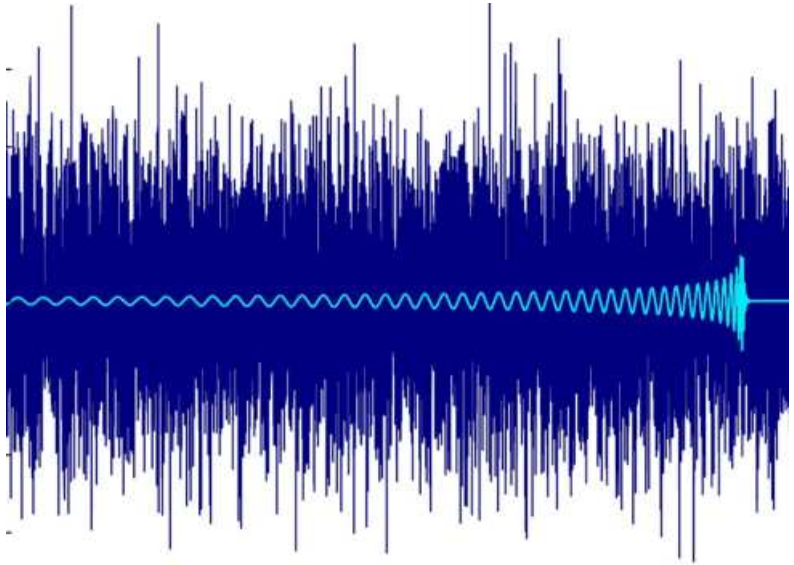
Under the same conditions of the Cramer-Rao bound

$$I_\theta = -\mathbb{E}\left[\frac{\partial^2 l}{\partial \theta^2}\right]$$

This result follows from the integration by parts, and dropping a boundary term by assuming that the probability tends to zero asymptotically.

The Fisher Information: The fisher information quantifies how sensitive the likelihood function is to changes in the parameter. The greater the Fisher information, the more easily we can estimate the parameter from the data, because small changes in the parameter will produce large changes in the likelihood.

Application in Gravitational Wave Astronomy: Fisher Matrix Approximation



H Gabbard *et al/ Phys. Rev. Lett.*

Signal: $s(t) = n(t) + h(t|\vec{\lambda}),$

data *noise* *signal*

Frequency dependent noise: $\langle \tilde{n}^*(f)\tilde{n}(f')\rangle = S_h(f)\delta(f-f').$

Power spectral density

Likelihood: $\mathcal{L}(s|\vec{\lambda}) = p(n(t) = s(t) - h(t|\vec{\lambda})) \propto \exp\left[-\frac{1}{2}(s-h(\vec{\lambda})|s-h(\vec{\lambda}))\right]$

Application in Gravitational Wave Astronomy: Fisher Matrix Approximation

We denote the Fisher Matrix as:

$$\mathrm{cov}(\hat{\lambda}_i, \hat{\lambda}_j) \geq [\mathbf{\Gamma}_\lambda]^{-1}_{ij} \qquad (\mathbf{\Gamma}_\lambda)_{ij} = \mathbb{E}\left[\frac{\partial l}{\partial \lambda_i}\frac{\partial l}{\partial \lambda_j}\right]$$

$$(\mathbf{\Gamma}_\lambda)_{ij} = \mathbb{E}\left[\frac{\partial l}{\partial \lambda_i}\frac{\partial l}{\partial \lambda_j}\right] = \left\langle \left(\frac{\partial h}{\partial \lambda_i}\middle|\mathbf{n}\right)\left(\frac{\partial h}{\partial \lambda_i}\middle|\mathbf{n}\right)\right\rangle = \left(\frac{\partial h}{\partial \lambda_i}\middle|\frac{\partial h}{\partial \lambda_j}\right)$$

The Fisher Matrix gives a lower bound on the variance of any unbiased estimator of the parameters of the signal, and hence it provides a guide to how accurately the parameters can be measured.

Application in Gravitational Wave Astronomy: Fisher Matrix Approximation

We have the true parameter and some small variation from it

Leading order expansion

$$\vec{\lambda} = \vec{\lambda}_0 + \Delta\vec{\lambda}, \qquad \Longrightarrow \qquad h(t;\vec{\lambda}) = h(t;\vec{\lambda}_0) + \partial_i h(t;\vec{\lambda}_0)\Delta\lambda^i$$

This is the linear signal approximation

$$\mathcal{L}(s|\vec{\lambda}) \propto \exp\left[-\frac{1}{2}(n - \partial_i h(t|\vec{\lambda})\Delta\lambda^i | n - \partial_j h(t|\vec{\lambda})\Delta\lambda^j)\right]$$

$$= \exp\left\{-\frac{1}{2}\left[(n|n) - 2(n|\partial_i h(t|\vec{\lambda}))\Delta\lambda^i + (\partial_i h(t|\vec{\lambda})|\partial_j h(t|\vec{\lambda}))\Delta\lambda^i\Delta\lambda^j\right]\right\}$$

Application in Gravitational Wave Astronomy: Fisher Matrix Approximation

Following this through leads to the maximum likelihood estimator

$$\widehat{\Delta\lambda}^{i} = (\Gamma^{-1})_{ik}(n|\partial_k h(t|\vec{\lambda}))$$

with

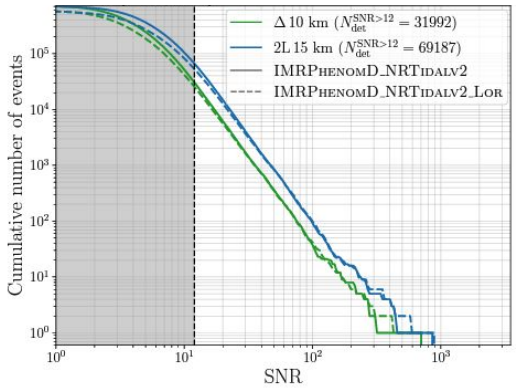
$$\mathbb{E}\left(\widehat{\Delta\lambda}^{i}\right) = 0, \qquad \mathrm{cov}\left(\widehat{\Delta\lambda}^{i}, \widehat{\Delta\lambda}^{j}\right) = \Gamma^{-1}_{ij}$$

Mean *Variance*

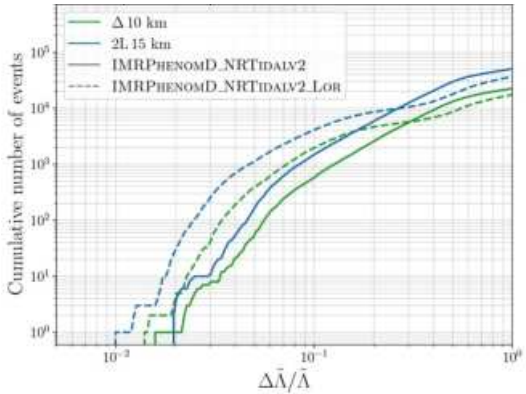
which shows that the Fisher Matrix is an *approximation* for the uncertainty of the parameter estimation.

Application in Gravitational Wave Astronomy: Fisher Matrix Approximation

Fisher matrix studies are useful when considering a large number of detected events with high signal-to-noise ratio e.g. to understand capabilities of future gravitational wave detectors such as the Einstein Telescope



very large number of events



Estimate of the error on the measured tidal deformability of neutron stars

2.4 Hypothesis testing: key concepts

Having observed data $\mathbf{x}$ we want to know if it is consistent with some preconceived assumptions.

Hypothesis testing is usually formulated as a test of reference of a reference null hypothesis against an alternative hypothesis.

If a hypothesis is completely specified it is *simple*, otherwise it is *composite*.

Examples:

- The average number of gravitational wave events observed on different days of the week is 5 **Simple**
- A trigger in a gravitational wave detector is due to noise **Composite**
- The number of gravitational wave events per year is Poisson distributed **Composite**

Which examples are simple and composite and why?

2.4 Hypothesis testing: key concepts

The outcome of a hypothesis is a decision to reject or accept the null hypothesis

Decision is based on a *test statistic* $t(\mathbf{x})$ and if values of the statistics are within an acceptance region or a *critical region*.

There are two kinds of error that are possible:

1. *Type I error*: Reject hypothesis although true
    i. Considered to be very problematic
    ii. The significance level of a test is the probability of the Type I error
2. *Type II error*: Fail to reject hypothesis when being false
    i. The power of a test is the probability of a Type II error

2.4 Hypothesis testing: key concepts

Test statistics have to fulfil:

1. Values of t are ordered with respect to the evidence for a departure from the null hypothesis
2. The distribution of $T=t(\mathbf{x})$ under the null hypothesis is known, or at least approximately

For any observation $\mathbf{x}$ we can measure consistency with the null hypothesis with *p-value*

$$p = \mathbb{P}(T \geq t(\mathbf{x}) | \, H_0),$$

Alternative hypothesis can be specified or unspecified, if unspecified then you just perform pure significance tests, e.g. looking for clustering of GW data on the sky.

2.4 Hypothesis testing: critical regions

Define (for a defined probability) a region in which the statistic has to lie in for the case the null hypothesis to be rejected

For any α in the interval $(0, 1)$, a subset R_α of X is a **critical region of size** α if

$$\mathbb{P}(\mathbf{X} \in R_\alpha | H_0) = \alpha$$

where

(i) points in R_α are regarded as not consistent with H_0 at level α;

(ii) points in R_α are "significant at level α";

(iii) if $\mathbf{x} \in R_\alpha$, then H_0 is "rejected" in a test of size α.

A significance test is defined by a set of critical regions $\{R_\alpha : 0 < \alpha < 1\}$ satisfying

$$R_{\alpha_1} \subset R_{\alpha_2} \text{ if } \alpha_1 < \alpha_2.$$

2.4 Hypothesis testing: confidence intervals

Define (for a defined probability) a region in which the statistic has to lie in for the case the null hypothesis to be rejected

For any α in the interval $(0, 1)$, a subset R_α of X is a **critical region of size** α if

$$\mathbb{P}(\mathbf{X} \in R_\alpha | H_0) = \alpha$$

where

(i) points in R_α are regarded as not consistent with H_0 at level α;

(ii) points in R_α are "significant at level α";

(iii) if $\mathbf{x} \in R_\alpha$, then H_0 is "rejected" in a test of size α.

A significance test is defined by a set of critical regions $\{R_\alpha : 0 < \alpha < 1\}$ satisfying

$$R_{\alpha_1} \subset R_{\alpha_2} \text{ if } \alpha_1 < \alpha_2.$$

3. Basics of Bayesian Statistics

Recap - Basic principles of Bayesian statistics

In Bayesian statistics probability is defined as…

the degree of belief you have for a certain outcome

In Bayesian statistics variables are treated as…

random variables with probability distributions

The output of Bayesian statistics is

the probabilities of a hypothesis

Bayesian statistics is useful in situations with…

limited or noisy data, and when you want to fold in other information

Recap - Basic principles of Bayesian statistics

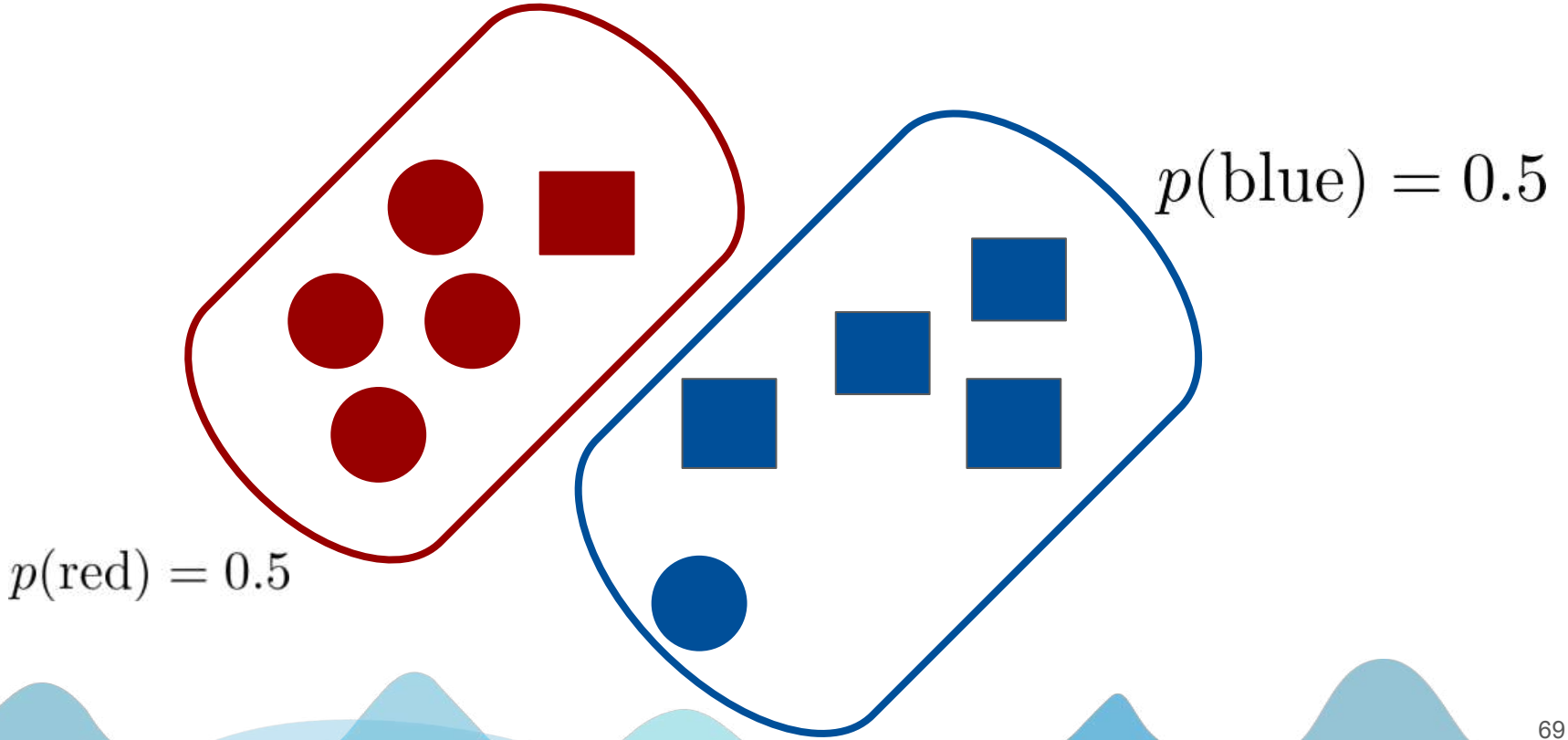
Frequentist : *“What is the probability of the observed data for a given hypothesis?“*

Bayesian: *“What is the probability of different hypotheses for given observed data?“*

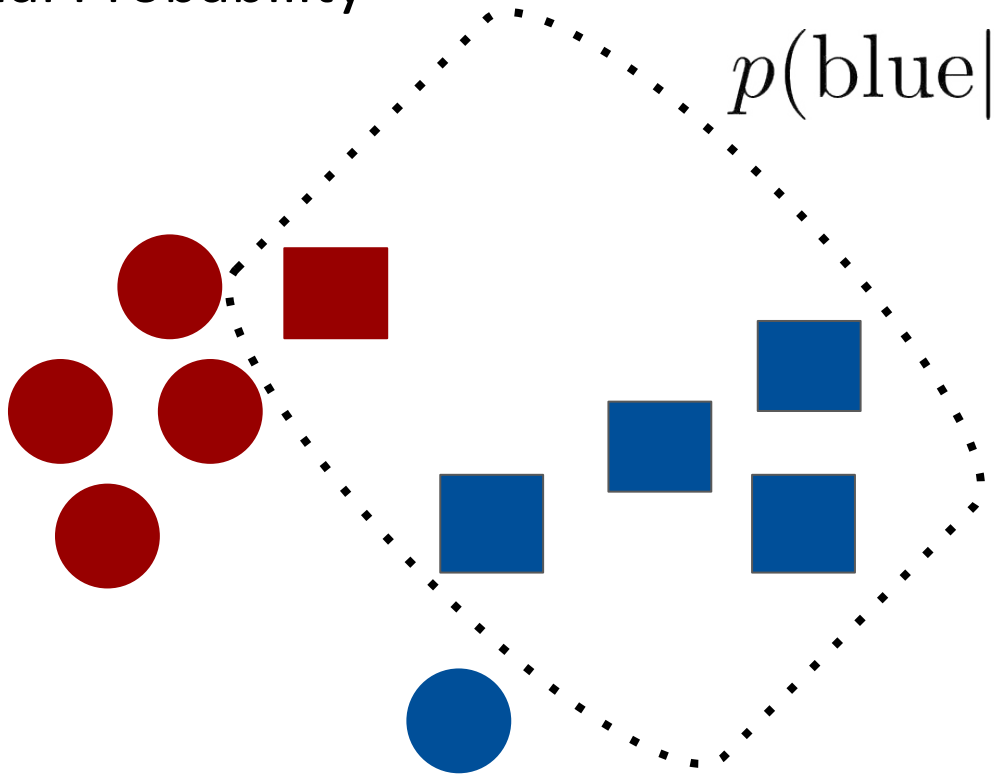
Bayesian statistics allows us to incorporate our own beliefs about the data, and update our beliefs as we include the data.

In gravitational wave data analysis we can’t repeat the same observation many times - each event has different properties and parameters, therefore Bayesian statistics is a better fit when determining source parameters

3.1 Conditional Probability



3.1 Conditional Probability



3.2 Bayes Theorem

$$p(A|B) = \frac{p(A \cap B)}{p(B)}$$

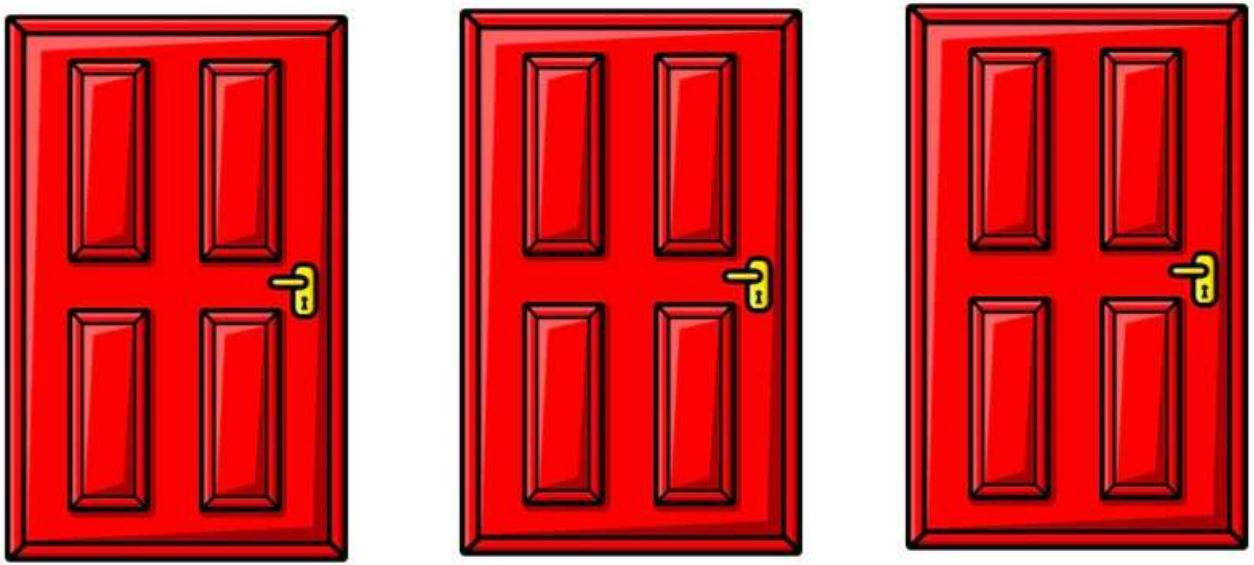
$$p(A \cap B) = p(A|B)p(B) = p(B|A)p(A)$$

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}$$

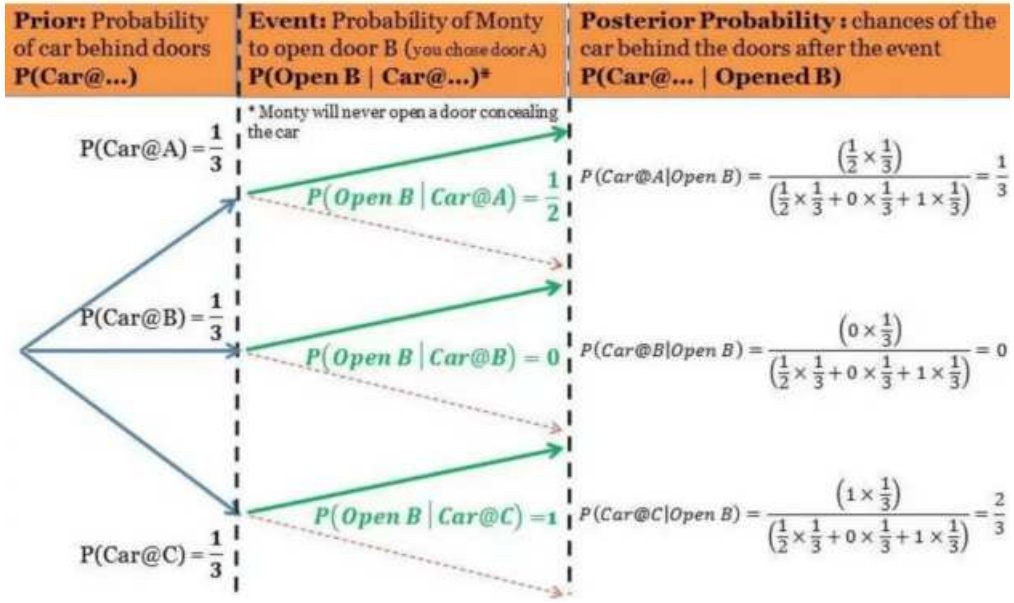
This is the cornerstone of Bayesian Statistics

Bayes Theorem

Bayes Theorem Example -The Monty Hall Problem



Bayes Theorem Example -The Monty Hall Problem



$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}$$

What if there are n doors?

$$p(A|B) = \frac{n-1}{n}$$

Bayes Theorem for Data Analysis

$$p(\vec{\theta}|\mathbf{x}) = \frac{p(\mathbf{x}|\vec{\theta})p(\vec{\theta})}{p(\mathbf{x})}$$

$$\mathcal{P}(\vec{\theta}) = \frac{\mathcal{L}(\vec{\theta})\pi(\vec{\theta})}{Z}$$

$\mathcal{P}(\vec{\theta})$: Posterior

$\mathcal{L}(\vec{\theta})$: Likelihood

$\pi(\vec{\theta})$: Prior

Z : Evidence

3.3 The Prior

$$\mathcal{P}(\vec{\theta}) = \frac{\mathcal{L}(\vec{\theta})\pi(\vec{\theta})}{Z}$$

1. Informative/expert priors

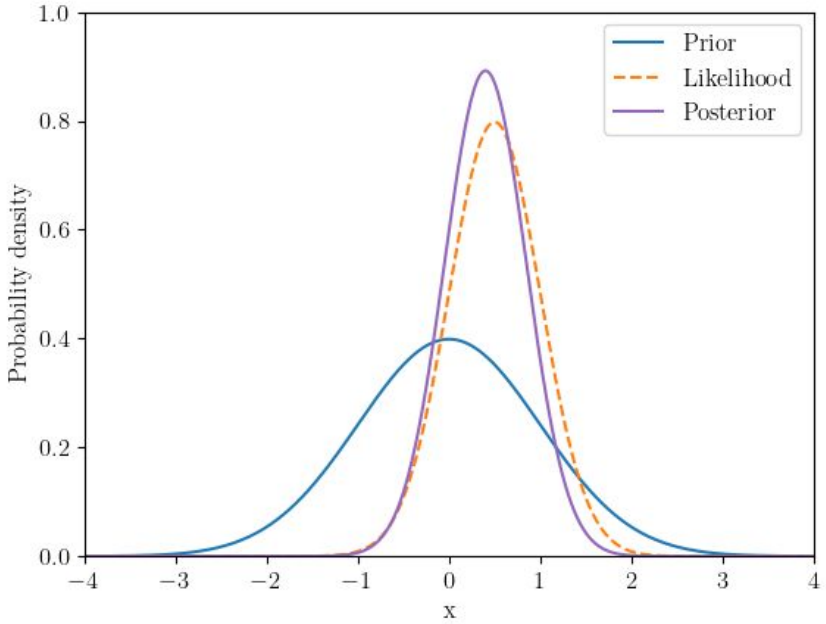
If previous information or measurements are available, it is reasonable to use this information within the prior

If no information is available “experts” may be able to give a reasonable prior

You can combine expert priors to make a *mixture prior*

$$p(\vec{\theta}) = \sum_{j=1}^{J} \omega_j p_j(\vec{\theta})$$

If you have no previous information typically an *uninformative prior* (often uniform) is usually used.



3.3 The Prior

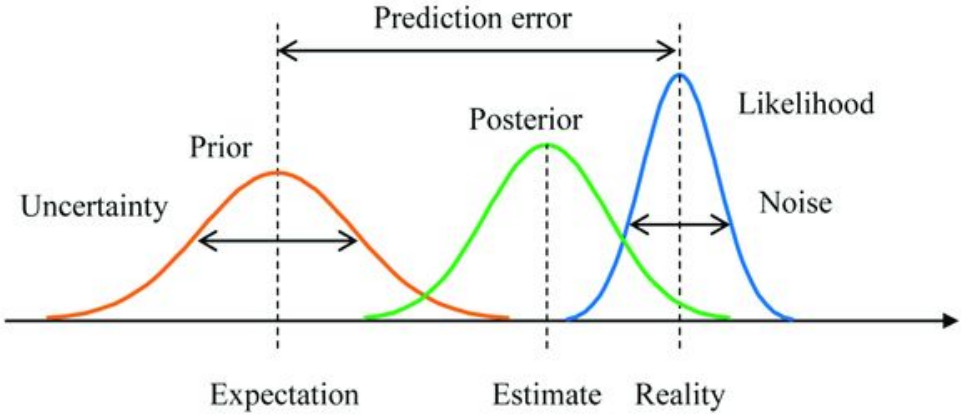
$$\mathcal{P}(\vec{\theta}) = \frac{\mathcal{L}(\vec{\theta})\pi(\vec{\theta})}{Z}$$

1. Informative/expert priors

A common criticism of Bayesian statistics is that results can be sensitive to the prior.

However, this is the desired behaviour, we want to include all the information we have access to.

On the other hand, badly chosen priors can lead to misleading or biased results.



Prior Selection: Example

$$\mathcal{P}(\vec{\theta}) = \frac{\mathcal{L}(\vec{\theta})\pi(\vec{\theta})}{Z}$$

Suppose we have a simple model $x \sim \mathcal{N}(\mu, 1)$. What prior should we use?

Uniform - Uninformative

$$\pi(\mu) \sim U[0, 1]$$

Suppose we observe one data point at $x = 1$
The posterior is then

$$p(\mu|x) \propto p(x|\mu)p(\mu) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(1-\mu)^2\right) (1)$$

The probability that $\mu > 0$ is

$$\int_0^\infty p(\mu|x) = 0.841$$

Gaussian - Informative

$$\pi(\mu) \sim \mathcal{N}\left(0, \frac{1}{10}\right)$$

We observe the same data point at $x = 1$
The posterior is then

$$p(\mu|x) \sim \mathcal{N}\left(\frac{1}{11}, \frac{1}{11}\right)$$

The probability that $\mu > 0$ is

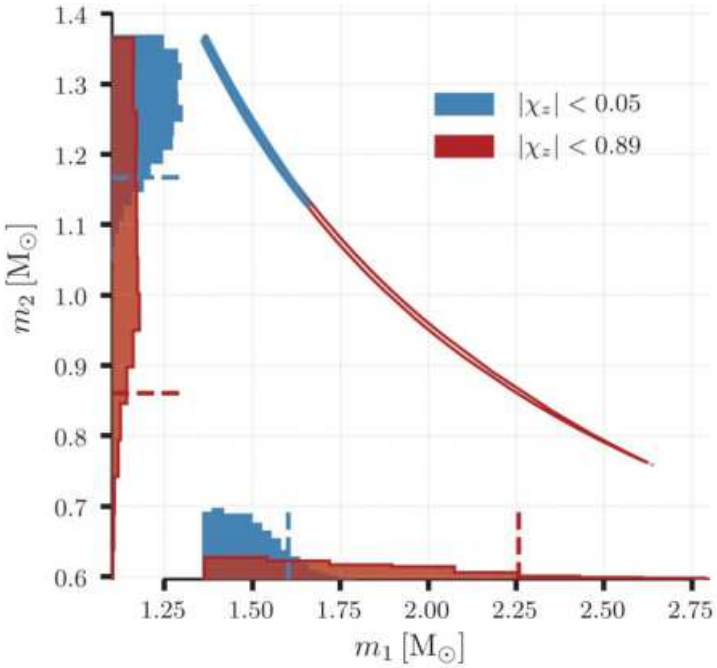
$$\int_0^\infty p(\mu|x) = 0.618$$

Very different results!

Prior Selection: Example

$$\mathcal{P}(\vec{\theta}) = \frac{\mathcal{L}(\vec{\theta})\pi(\vec{\theta})}{Z}$$

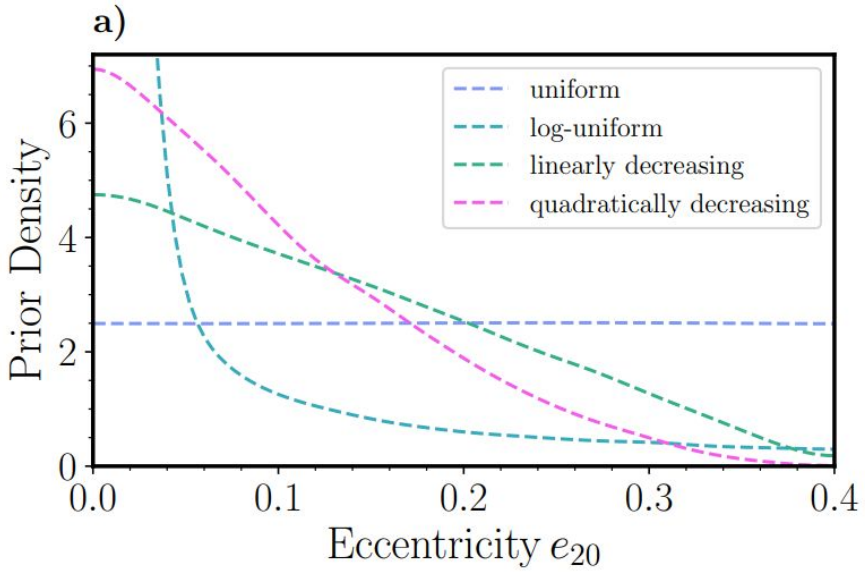
GW170817 binary neutron star merger - what were the component masses?

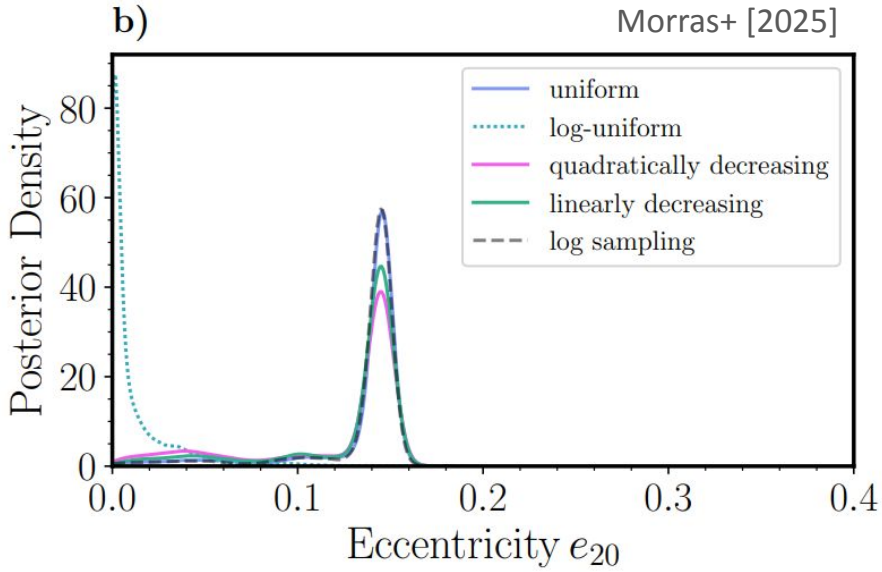


High spin prior or low spin prior? Both are uniform, but with different ranges.

Prior Selection: Example

$$\mathcal{P}(\vec{\theta}) = \frac{\mathcal{L}(\vec{\theta})\pi(\vec{\theta})}{Z}$$





Morras+ [2025]

The functional form of prior choices can influence a measurement.

3.3 The Prior

$$\mathcal{P}(\vec{\theta}) = \frac{\mathcal{L}(\vec{\theta})\pi(\vec{\theta})}{Z}$$

2. Conjugate Priors

The prior is chosen such that the prior and posterior are in the same family.

Often used in the past, when computational power was not powerful.

Common examples:

- Beta-Binomial model
- Poisson-Gamma model
- Normal-Normal/Normal-Gamma model

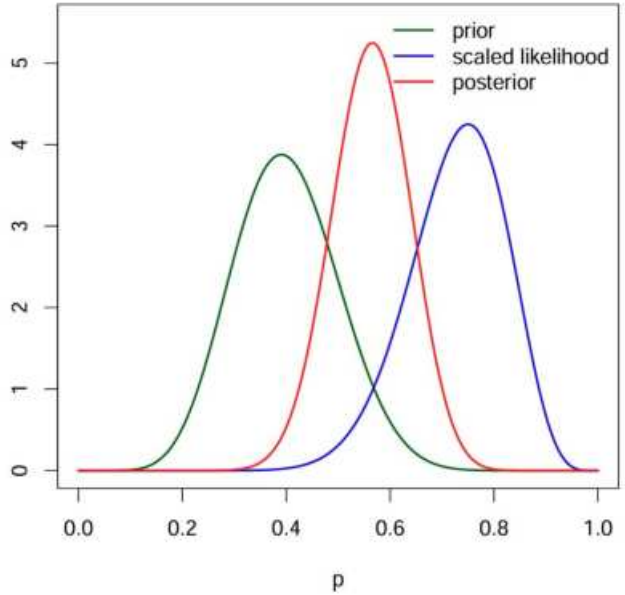
3.3 The Prior

$$\mathcal{P}(\vec{\theta}) = \frac{\mathcal{L}(\vec{\theta})\pi(\vec{\theta})}{Z}$$

3. Conjugate Priors with expert opinion

Example: Consider a drug to be given for relief of chronic pain. Experience with similar compounds has suggested that response rates, p, between 0.2 and 0.6 could be feasible. We plan to observe the response rate in n patients and want to infer a posterior on p. Propose a suitable conjugate prior for p based on the available information.

A response rate between 0.2 and 0.6 could be used to set a uniform prior in that range. However, this is not conjugate to the binomial distribution that determines the observed data. Therefore, it would be better to use a conjugate prior. A $U[0.2, 0.6]$ distribution has mean 0.4 and standard deviation of 0.1. We can find a Beta distribution that has the same mean and standard deviation. Rearranging the equations given earlier we deduce $\text{Beta}(a = 9.2, b = 13.8)$ has the desired mean and variance. This prior is conjugate and reflects the expert opinion as regards the expected response rate for the drug. Suppose now we observe $n = 20$ patients and $x = 15$ respond positively. The posterior is then $\text{Beta}(9.2 + 15, 13.8 + 5) = \text{Beta}(24.2, 18.8)$. The prior, (scaled) likelihood and posterior are illustrated in Figure 8.



Example taken from J. Gair lecture notes.

3.3 The Prior

$$\mathcal{P}(\vec{\theta}) = \frac{\mathcal{L}(\vec{\theta})\pi(\vec{\theta})}{Z}$$

4. Jeffrey prior

Often we use uniform priors for uninformative priors, however they are not invariant under parameterisation.

If we know nothing about θ we also don't know θ^2 or θ^3

Jefferies (1961) proposed a class of prior that are invariant under reparameterisation.

$$p(\vec{\theta}) \propto \sqrt{\det[I(\vec{\theta})]}, \qquad \text{where } I(\vec{\theta})_{ij} = \mathbb{E}\left[\frac{\partial l}{\partial \theta_i}\frac{\partial l}{\partial \theta_j}\right]$$

3.3 The Prior

$$\mathcal{P}(\vec{\theta}) = \frac{\mathcal{L}(\vec{\theta})\pi(\vec{\theta})}{Z}$$

4. Jeffrey prior

Jefferies (1961) proposed a class of prior that are invariant under reparameterisation.

$$p(\vec{\theta}) \propto \sqrt{\det[I(\vec{\theta})]}, \qquad \text{where } I(\vec{\theta})_{ij} = \mathbb{E}\left[\frac{\partial l}{\partial \theta_i}\frac{\partial l}{\partial \theta_j}\right]$$

Example: Poisson distribution For a single observation, x, from the Poisson(λ) distribution with pmf

$$p(x|\lambda) = \frac{\lambda^x \mathrm{e}^{-\lambda}}{x!}$$

we have

$$\frac{\partial \log p}{\partial \lambda} = \frac{x}{\lambda} - 1, \quad \frac{\partial^2 \log p}{\partial \lambda^2} = -\frac{x}{\lambda^2} \quad \Rightarrow \quad I(\lambda) \equiv \mathbb{E}\left[-\frac{\partial^2 \log p}{\partial \lambda^2}\right] = \frac{1}{\lambda}.$$

The Jeffreys prior for the Poisson distribution is therefore $p(\lambda) \propto 1/\sqrt{\lambda}$. This is an example of an **improper** prior, since it cannot be normalised to integrate to 1 unless the range of rates is restricted.

3.4 The Likelihood

$$\mathcal{P}(\vec{\theta}) = \frac{\mathcal{L}(\vec{\theta})\pi(\vec{\theta})}{Z}$$

The likelihood is problem specific and quantifies how well your prediction agrees with the data.

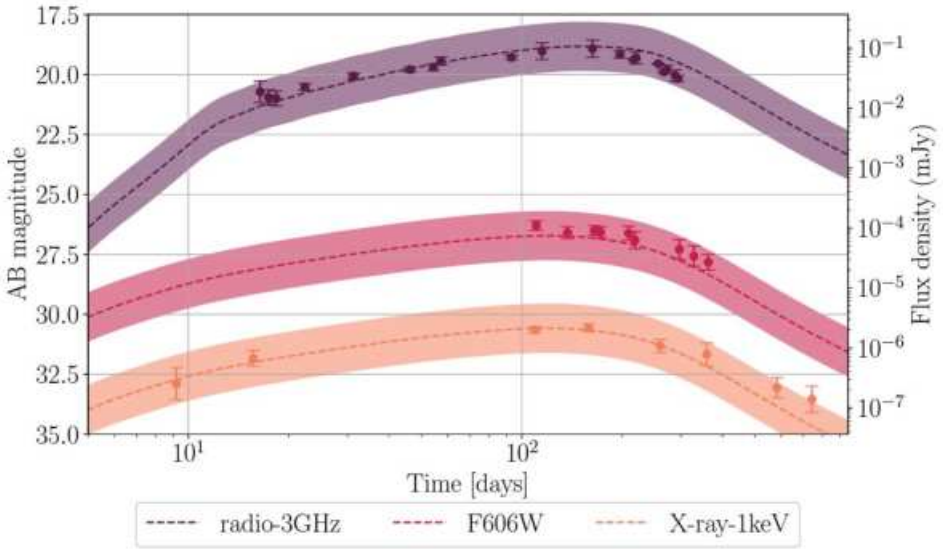
It is necessary (and useful!) that the likelihood fulfills some regularity conditions:

- Existence of a global maximum (e.g. for maximum likelihood estimation)
- Continuous on a compact support (the function is defined within a finite and bounded range)

3.4 The Likelihood

$$\mathcal{P}(\vec{\theta}) = \frac{\mathcal{L}(\vec{\theta})\pi(\vec{\theta})}{Z}$$

Example: Observation of the GRB afterglow of a binary neutron star merger



$$\mathcal{L}_{\rm EM} \propto \exp\left(-\frac{1}{2}\sum_{ij}\left(\frac{m_i^j - m_i^{j,\rm est}(\vec{\theta})}{\sigma_i^j}\right)^2\right)$$

data *prediction* *uncertainty*

3.4 The Likelihood

$$\mathcal{P}(\vec{\theta}) = \frac{\mathcal{L}(\vec{\theta})\pi(\vec{\theta})}{Z}$$

Example: Observation of Gravitational Waves

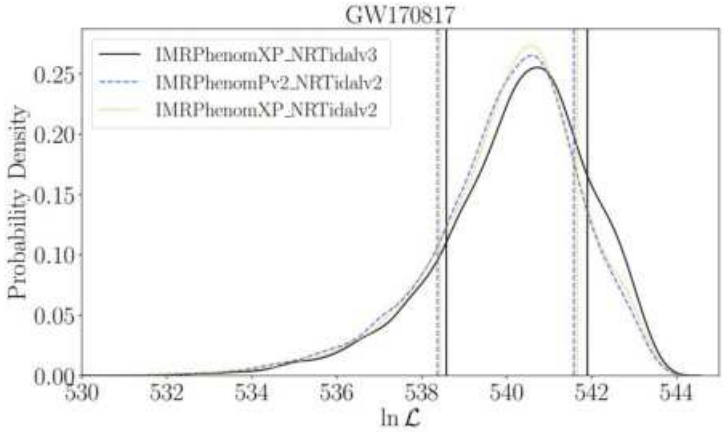
By assuming stationary Gaussian noise, the GW likelihood give data *d* and model *h* is given as

$$\mathcal{L}_{\rm GW} \propto \exp\left(-\frac{1}{2}\langle d - h(\vec{\theta})|d - h(\vec{\theta})\rangle\right)$$

with

$$\langle a|b\rangle = 4\Re\int_{f_{\rm low}}^{f_{\rm high}} \frac{\tilde{a}(f)\tilde{b}^*(f)}{S_n(f)}df$$

Noise weighted inner product



Log likelihood from GW170817

3.4 The Likelihood

$$\mathcal{P}(\vec{\theta}) = \frac{\mathcal{L}(\vec{\theta})\pi(\vec{\theta})}{Z}$$

Example: NICER Observations

- Integer detection of photons
- Each measurement is independent

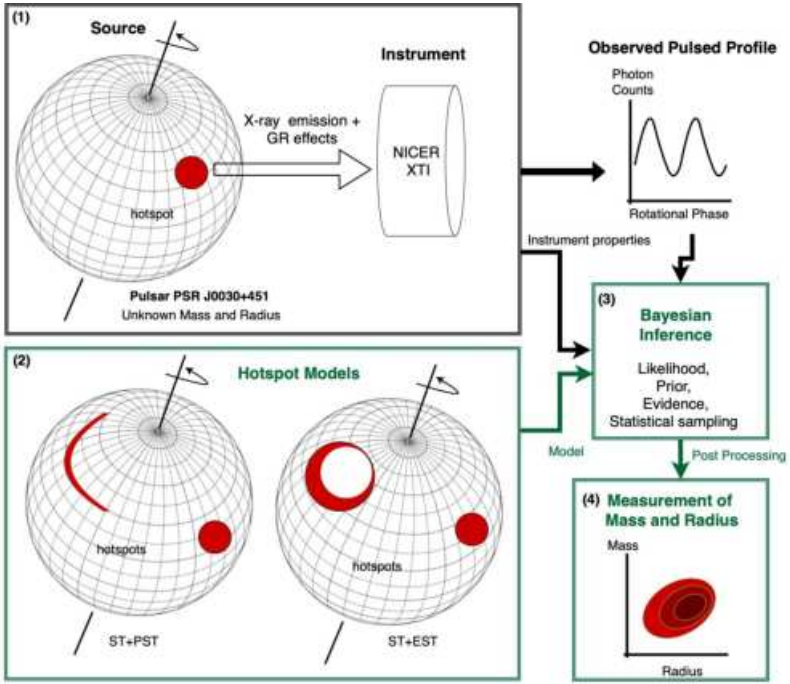
$$\ln \mathcal{L} = \sum_{\phi_i} \sum_{E_j} (d_{ij} \ln m_{ij} - m_{ij})$$

data *model prediction*

instruments

energy bins

Poisson distribution



3.5 The Evidence

$$\mathcal{P}(\vec{\theta}) = \frac{\mathcal{L}(\vec{\theta})\pi(\vec{\theta})}{Z}$$

The evidence is a normalisation constant that can be ignored for the purposes of parameter estimation

However, it is required for comparing hypotheses and model selection.

The evidence is difficult to compute via MCMC, but possible via nested sampling

$$\mathcal{Z} = p(\mathbf{x}) = \int p(\mathbf{x} \mid \vec{\theta}) p(\vec{\theta}) \mathrm{d}\vec{\theta}$$

Overall the evidence is the probability of the data occurring with the given model

3.6 The Posterior

$$\mathcal{P}(\vec{\theta}) = \frac{\mathcal{L}(\vec{\theta})\pi(\vec{\theta})}{Z}$$

Often we are only interested in one parameter, then we *marginalise* over all other parameters

$$p_{\text{marg}}(\theta_1|\mathbf{x}) = \int p(\vec{\theta}|\mathbf{x}) \mathrm{d}\theta_2 \ldots \mathrm{d}\theta_m$$

From here we can computer informative point estimates such at the posterior mean, median and mode for this parameter

$$\mu = \int_{-\infty}^{\infty} \theta_1 p_{\text{marg}}(\theta_1|\mathbf{x}) \mathrm{d}\theta_1$$

$$\int_{-\infty}^{m} p_{\text{marg}}(\theta_1|\mathbf{x}) \mathrm{d}\theta_1 = 0.5 = \int_{m}^{\infty} p_{\text{marg}}(\theta_1|\mathbf{x}) \mathrm{d}\theta_1$$

$$M = \text{argmax } p_{\text{marg}}(\theta_1|\mathbf{x})$$

3.6 The Posterior

$$\mathcal{P}(\vec{\theta}) = \frac{\mathcal{L}(\vec{\theta})\pi(\vec{\theta})}{Z}$$

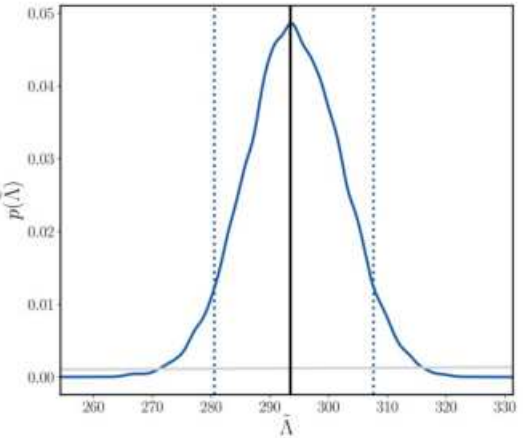
Credible Intervals: An interval (a,b) is a 100(1-α)% posterior credible interval for θ_1 if

$$\int_a^b p_{\rm marg}(\theta_1|\mathbf{x})\mathrm{d}\theta_1 = (1-\alpha), \quad 0 \leq \alpha \leq 1.$$

Symmetric posterior credible interval

An interval (a,b) is a symmetric 100(1-α)% posterior credible interval if

$$\int_{-\infty}^a p_{\rm marg}(\theta_1|\mathbf{x})\mathrm{d}\theta_1 = \frac{\alpha}{2} = \int_b^{\infty} p_{\rm marg}(\theta_1|\mathbf{x})\mathrm{d}\theta_1.$$



3.6 The Posterior

$$\mathcal{P}(\vec{\theta}) = \frac{\mathcal{L}(\vec{\theta})\pi(\vec{\theta})}{Z}$$

Credible Intervals: An interval (a,b) is a 100(1-α)% posterior credible interval for θ_1 if

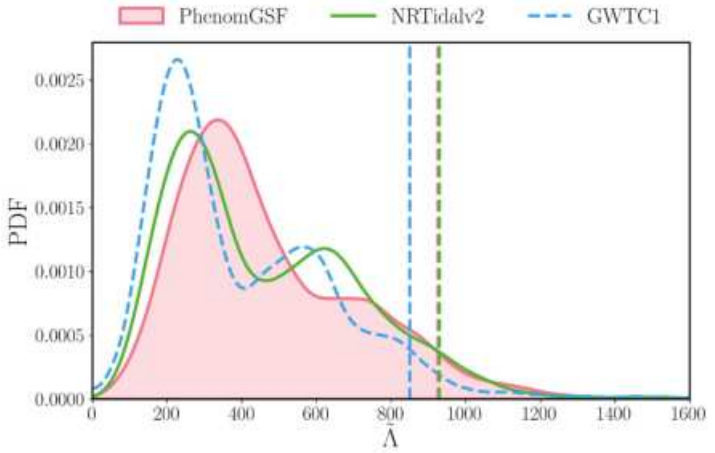
$$\int_a^b p_{\rm marg}(\theta_1|\mathbf{x})\mathrm{d}\theta_1 = (1-\alpha), \quad 0 \leq \alpha \leq 1.$$

Highest posterior density credible interval

An interval (a,b) is a 100(1-α)% highest posterior density interval (HPD) if

1. $[a,b]$ is a $100(1-\alpha)\%$ credible interval for θ_1;
2. for all $\theta \in [a,b]$ and $\theta' \notin [a,b]$ we have $p_{\rm marg}(\theta|\mathbf{x}) \geq p_{\rm marg}(\theta'|\mathbf{x})$.

Why might you use one credible interval over the other?



3.6 The Posterior

$$\mathcal{P}(\vec{\theta}) = \frac{\mathcal{L}(\vec{\theta})\pi(\vec{\theta})}{Z}$$

Representing results through posterior samples

- Using single numbers to represent the posterior leads to loss of information
- Usually the posterior cannot be given in closed form
- Often the posterior is approximated by a large number of samples from the posterior
- Posterior integrals can then be approximated through a discrete sum

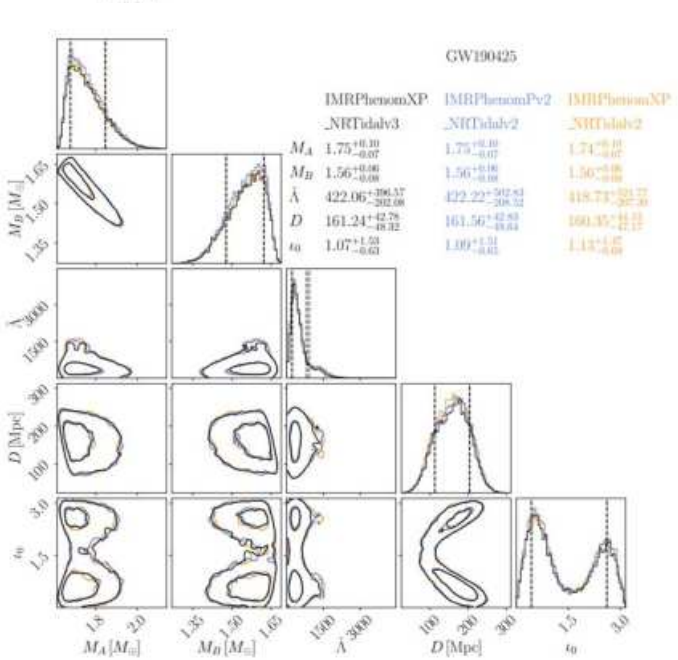
3.6 The Posterior

$$\mathcal{P}(\vec{\theta}) = \frac{\mathcal{L}(\vec{\theta})\pi(\vec{\theta})}{Z}$$

Corner Plots

- Corner plots are often used to visualise one and two dimensional posteriors across a parameter space
- This allows more information to be shown

What can we infer from this corner plot that we wouldn't be able to infer with point estimates?



4. Bayesian Statistics

4.2 Hypothesis testing

1. Assign prior probabilities to hypotheses H_0 and H_1
2. Use Bayes theorem to update the priors after observing the data, and calculate the posterior probabilities $p(H_0|\mathbf{x})$ and $p(H_1|\mathbf{x})$
3. Compute a Bayes factor or an odds ratio to compare the two hypotheses
4. Use this choose the hypothesis with the most support

Hypothesis testing uses the evidence (probability of the data given our model):

$$\mathcal{Z} = p(\mathbf{x}|M) = \int p(\mathbf{x}|\vec{\theta}, M)p(\vec{\theta}|M)d\vec{\theta}$$

The evidence incorporates Occam's razor (over complicated models are penalised)

4.2 Model Comparison

For a proper model comparison we can use the evidence (probability of the data given the model) and obtain the posterior *odds ratio*:

$$O_{12} = \frac{p(\mathbf{x}|\mathbf{M_1})}{p(\mathbf{x}|\mathbf{M_2})} \frac{p(M_1)}{p(M_2)}$$

Odds ratio — *Bayes factor* — *Prior odds ratio*

Often we just consider the Bayes factor when comparing hypotheses with identical priors

Bayes Factor	Interpretation
< 3	No evidence of M_1 over M_2
> 3	Positive evidence for M_1
> 20	Strong evidence for M_1
> 150	Very strong evidence for M_1

4.2 Model Comparison

For a proper model comparison we can use the evidence (probability of the data given the model) and obtain the posterior *odds ratio*:

$$O_{12} = \frac{p(\mathbf{x}|\mathbf{M_1})}{p(\mathbf{x}|\mathbf{M_2})} \frac{p(M_1)}{p(M_2)}$$

Odds ratio — *Bayes factor* — *Prior odds ratio*

Often we just consider the Bayes factor when comparing hypotheses with identical priors

Calculating the Bayes factor is challenging. It can be approximated by

$$\frac{1}{\mathcal{Z}} = \int \frac{1}{p(\mathbf{x} \mid \vec{\theta})} \frac{p(\mathbf{x} \mid \vec{\theta}) p(\vec{\theta})}{\mathcal{Z}} \mathrm{d}\vec{\theta} \approx \frac{1}{M} \sum_{i=1}^{M} \frac{1}{p(\mathbf{x} \mid \vec{\theta}_i)}$$

Numerically quite unstable

4.2 Model Comparison

It is important to be careful with the interpretation of Bayes factors

- You are only making statements for individual models (what if both models are wrong?)
- If your model has issues describing the system this might not mean another model might suffer the same issue
- You only get a probability estimate, you can never be sure about the real outcome
- Bayes factors can be sensitive to the prior choices

4.3 Compare Distributions

Kullback-Leibler (KL) divergence

How much information do you gain from the data?

$$D_{\mathrm{KL}}(\mathcal{P}(\vec{\theta}) \,||\pi(\vec{\theta})) = \int d\vec{\theta}\mathcal{P}(\vec{\theta}) \log\left(\frac{\mathcal{P}(\vec{\theta})}{\pi(\vec{\theta})}\right)$$

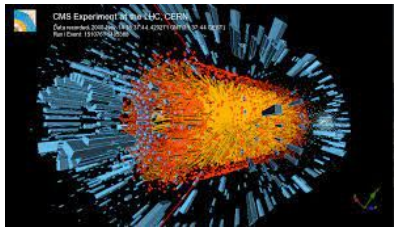
- If the posterior is identical to the prior the KL divergence is zero
- The larger the KL divergence, the more information gained
  - As the posterior approaches a Dirac delta function, the KL divergence approaches infinity

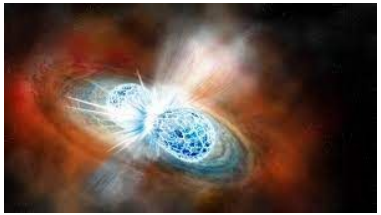
4.3 Compare Distributions

Kullback-Leibler (KL) divergence - Example for matter at extreme densities

	D_{KL}	Astro only	HIC only	Astro+HIC
ρ	$1.0n_{\mathrm{sat}}$	0.079	0.109	0.270
	$1.5n_{\mathrm{sat}}$	0.075	0.108	0.266
	$2.0n_{\mathrm{sat}}$	0.112	0.019	0.174
	$2.5n_{\mathrm{sat}}$	0.244	0.006	0.274
M_{NS}	$1.0M_{\odot}$	0.090	0.054	0.128
	$1.4M_{\odot}$	0.185	0.022	0.210
	$1.6M_{\odot}$	0.225	0.015	0.251
	$2.0M_{\odot}$	0.228	0.008	0.222

[Huth+ (2022)]





4.3 Compare Distributions

Jensen-Shannon (JS) divergence

The Jensen-Shannon divergence is a symmetrised and smoothed version of the KL divergence

$$D_{JS}(P\|Q) = \frac{1}{2}D_{KL}(P\|M) + \frac{1}{2}D_{KL}(Q\|M)$$

$$\text{where } M = \frac{1}{2}(P+Q)$$

This allows for a measure which is always symmetric, finite and bounded

$$0 \leq D_{JS}(P\|Q) \leq \log 2$$

4.4 Predictive Checking

Prior predictive distribution

The likelihood weighted by the assigned prior distribution and represents our expectation of the distribution of data sets that we will observe **before** seeing any data.

Notice we integrate out unknown parameters

$$p(\mathbf{x}) = \int_{\vec{\theta} \in \Theta} p(\mathbf{x}|\vec{\theta}) p(\vec{\theta}) \mathrm{d}\vec{\theta}.$$

Posterior predictive distribution

The likelihood weighted by the posterior probability **after** observing data $\mathbf{x}$ and represents our expectation about the distribution of future data sets $\mathbf{y}$.

$$p(\mathbf{y}|\mathbf{x}) = \int_{\vec{\theta} \in \Theta} p(\mathbf{y}|\vec{\theta}) p(\vec{\theta}|\mathbf{x}) \mathrm{d}\vec{\theta}.$$

4.4 Predictive Checking

Prior predictive distribution

Do our assumptions lead to realistic data?

Notice we integrate out unknown parameters

$$p(\mathbf{x}) = \int_{\vec{\theta} \in \Theta} p(\mathbf{x}|\vec{\theta}) p(\vec{\theta}) \mathrm{d}\vec{\theta}.$$

Posterior predictive distribution

Does our model produce similar data to that observed?

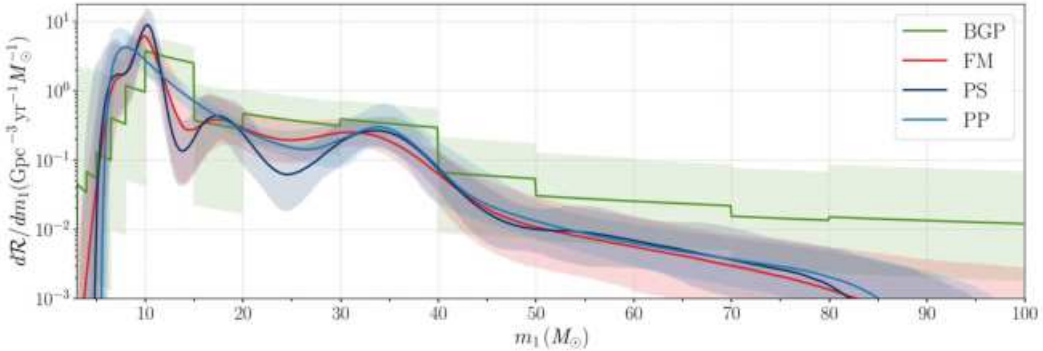
$$p(\mathbf{y}|\mathbf{x}) = \int_{\vec{\theta} \in \Theta} p(\mathbf{y}|\vec{\theta}) p(\vec{\theta}|\mathbf{x}) \mathrm{d}\vec{\theta}.$$

Practically we draw new samples from our posterior and compare to the observed data.

For high observed data dimensions, this comparison can be difficult, so we typical use summary statistics such as mean, skewness, kurtosis etc.

4.5 Hierarchical models

In reality the parameters of interest are not fully independent, but connected through some population model of astrophysical interest. Therefore you may consider your priors again as random variables that come from their own *hyper-prior*, which is characterised by *hyper-parameters.*



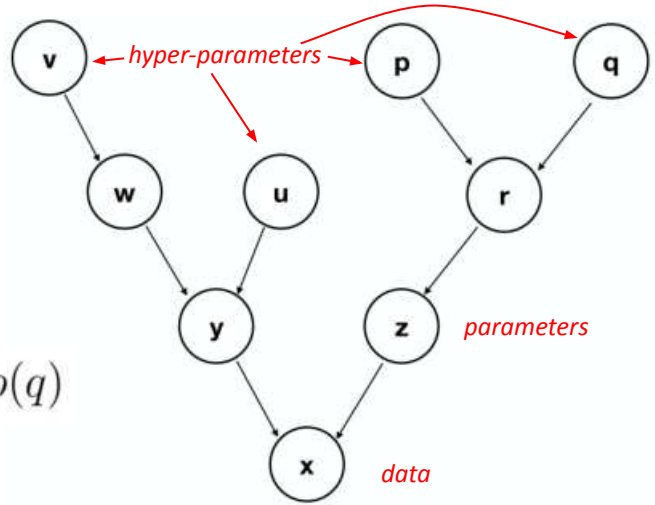
Merger rate of binary black holes as a function of the primary mass

$$p(m_1|\theta) = [(1 - \lambda_m)A(\theta)m_1^{-\alpha}\Theta(m_{\max} - m_1) + \lambda_m B(\theta)\exp\left(-\frac{(m_1 - \mu_m)^2}{2\sigma_m^2}\right)]S(m_1, m_{\min}, \delta m)$$

Example - Power law + Peak (PP)

4.5 Hierarchical models

In reality the parameters of interest are not fully independent, but connected through some population model of astrophysical interest. Therefore you may consider your priors again as random variables that come from their own *hyper-prior*, which is characterised by *hyper-parameters.*



$$p(p,q,r,s,t,u,v,w,x,y,z) =$$

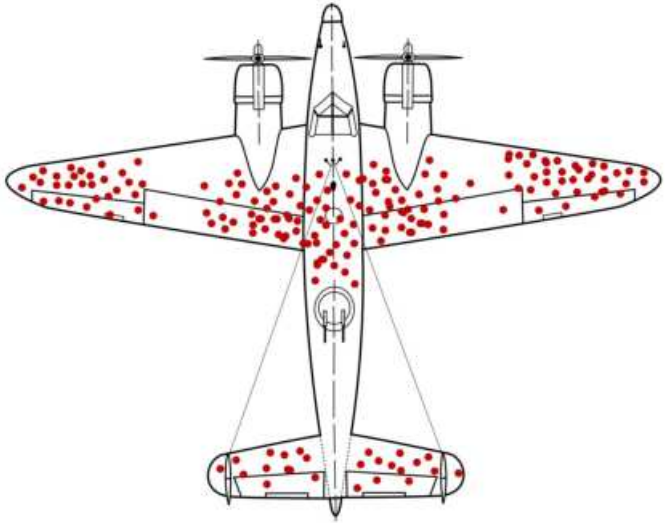
$$p(x|y,z)p(y|u,w)p(w|v)p(u)p(v)p(z|r)p(r|p,q)p(p)p(q)$$

4.5 Selection effects

The US in WW2 observed that their planes coming back from missions were damaged more in certain areas than others. They decided to reinforce planes but to keep planes light needed to be selective where.

Where should be planes be reinforced?

This is a biased dataset - these were the planes that *came back* from missions. This suggests that the planes shot in the areas with no data were destroyed and did not come back to be part of the dataset - *reenforce the areas without damage.*



This is an example of selection effects/selection bias

4.5 Selection effects in astrophysics

No instrument is infinitely sensitive, hence not all sources are seen and there is an inherent bias in our data towards closer and louder/brighter events. This could be corrected for by

1. Correcting the likelihood so it represents the likelihood of detected events

$$p(\mathbf{x}|\vec{\theta}, \text{obs}) = \frac{1}{p_s(\vec{\theta})} p(\mathbf{x}|\vec{\theta}), \quad \text{where } p_s(\vec{\theta}) = \int_{\mathbf{x}>\text{threshold}} p(\mathbf{x}|\vec{\theta}) \mathrm{d}\mathbf{x}.$$

Assuming the number of detections has no information about the parameters of interest.

4.5 Selection effects in astrophysics

2. Separate likelihoods for detected and non-detected events

$$p\left(\left\{\vec{\theta}_i\right\},\left\{\vec{\theta}_j\right\},\{\mathbf{x}_i\},\{\mathbf{x}_j\} \mid \vec{\lambda}\right) \propto \left[\prod_{i=1}^{N\text{obs}} p\left(\mathbf{x}_i \mid \vec{\theta}_i\right) \frac{\mathrm{d}N}{\mathrm{d}\vec{\theta}_i}\left(\vec{\lambda}\right)\right] \times \left[\prod_{j=1}^{N_{\text{nobs}}} p\left(\mathbf{x}_j \mid \vec{\theta}_j\right) \frac{\mathrm{d}N}{\mathrm{d}\vec{\theta}_j}\left(\vec{\lambda}\right)\right] \exp\left[-N\left(\vec{\lambda}\right)\right]$$

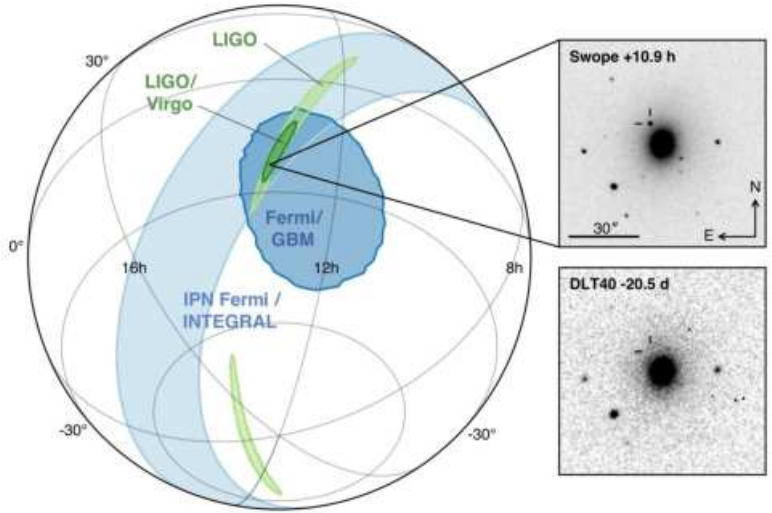
Marginalising over the unobserved data and number of unobserved events we obtain

$$N_{\text{det}}(\vec{\lambda}) = N \int_{\mathbf{x}>\text{threshold}} \int p(\mathbf{x}|\vec{\theta}) p(\vec{\theta}|\vec{\lambda}) \mathrm{d}\vec{\theta}\mathrm{d}\mathbf{x} = N p_s(\vec{\lambda}).$$

Which for a scale invariant prior on overall merger rate N, leads us back to method 1.

4.5 Example - Inferring the Hubble constant with GWs

Binary neutron star merger GW170817 was observed in GWs and with electromagnetic counterpart (GRB + kilonova). The counterpart measurement allowed a redshift measurement. The GWs allowed a luminosity distance measurement which can be used together to make a measurement on the rate of universe expansion - *the Hubble constant:*



Luminosity distance

$$v_r = H_0 d + v_p$$

Recessional velocity — *Hubble constant* — *Peculiar velocity*

4.5 Example - Inferring the Hubble constant with GWs

Observation of GW data

Distance, inclination

$$p(x_{\rm GW} \mid d, \cos\iota) = \int p(x_{\rm GW} \mid d, \cos\iota, \vec{\lambda})\, p(\vec{\lambda}) \mathrm{d}\vec{\lambda}$$

Marginalise over irrelevant parameters

Observation of recessional velocity from EM data

$$p\left(v_r \mid d, v_p, H_0\right) = N\left[v_p + H_0 d, \sigma_{v_r}^2\right]\left(v_r\right)$$

The measured smoothed peculiar velocity

$$p\left(\langle v_p \rangle \mid v_p\right) = N\left[v_p, \sigma_{v_p}^2\right]\left(\langle v_p \rangle\right)$$

4.5 Example - Inferring the Hubble constant with GWs

The combined likelihood of all data streams

$$p(x_{\rm GW}, v_r, \langle v_p \rangle \mid d, \cos\iota, v_p, H_0) = \frac{1}{\mathcal{N}_s(H_0)} p(x_{\rm GW} \mid d, \cos\iota)\, p(v_r \mid d, v_p, H_0)\, p(\langle v_p \rangle \mid v_p).$$

Selection effect parameter

Independent priors

$$p(d, \cos\iota, v_p, H_0) = p(d)p(\cos\iota)p(v_p)p(H_0).$$

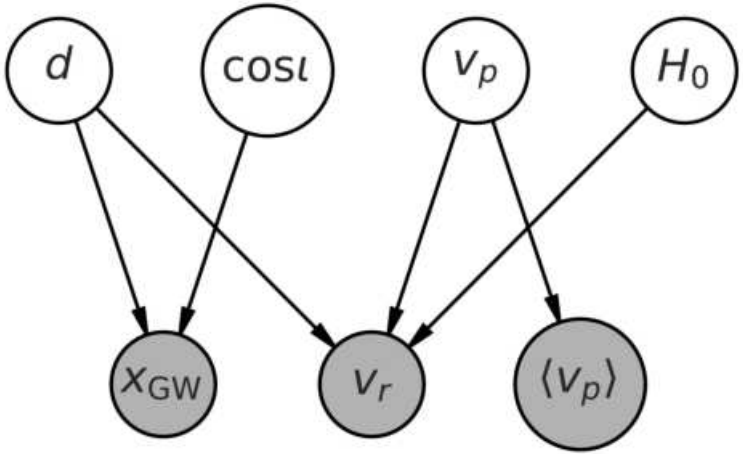
4.5 Example - Inferring the Hubble constant with GWs

The posterior is then

$$p(H_0, d, \cos\iota, v_p \mid x_{\rm GW}, v_r, \langle v_p \rangle)$$

$$\propto \frac{p(H_0)}{\mathcal{N}_{\rm s}(H_0)}\, p(x_{\rm GW} \mid d, \cos\iota)\, p(v_r \mid d, v_p, H_0)$$

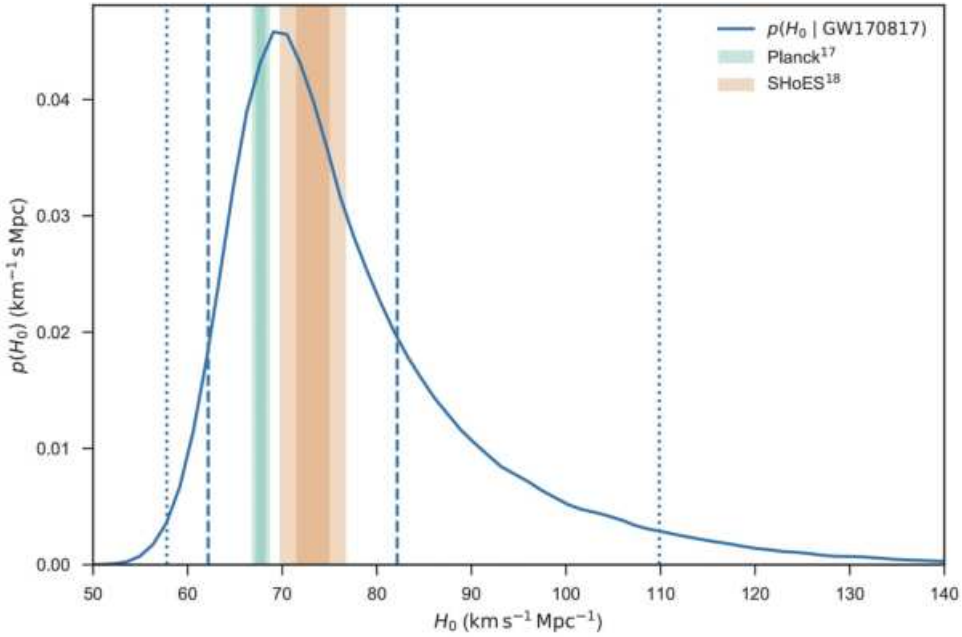
$$\times\ p(\langle v_p \rangle \mid v_p)\, p(d)\, p(v_p)\, p(\cos\iota)$$



4.5 Example - Inferring the Hubble constant with GWs

Then marginalising over distance, inclination and peculiar velocity give a posterior on the Hubble constant

$$p(H_0 \mid x_{\rm GW}, v_r, \langle v_p \rangle) \propto \frac{p(H_0)}{\mathcal{N}_s(H_0)} \int \mathrm{d}d\,\mathrm{d}v_p\,\mathrm{d}\cos\iota \times p(x_{\rm GW} \mid d, \cos\iota)\, p(v_r \mid d, v_p, H_0) \times p(\langle v_p \rangle \mid v_p)\, p(d)\, p(v_p)\, p(\cos\iota)$$



5. Computational Methods

What does this mean in practice?

Ok, I see the theory, but how do we estimate the posterior in real life?

$$\mathcal{P}(\vec{\theta}) = \frac{\mathcal{L}(\vec{\theta})\pi(\vec{\theta})}{Z}$$

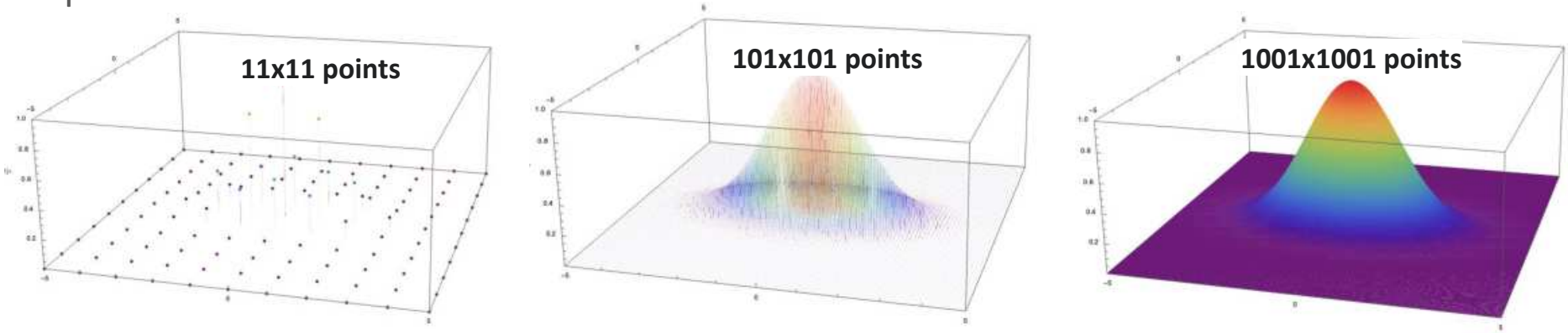
…with great difficulty

In most cases the posterior cannot be written down analytically and we must use other tools to approximate the posterior distribution. We can do this by calculating the posterior at fixed points

$$\int f(\boldsymbol{\theta})\; p(\boldsymbol{\theta}|\mathbf{x})\; \mathrm{d}\boldsymbol{\theta} \approx \frac{1}{M}\sum_{i=1}^{M} f(\boldsymbol{\theta}_i)$$

5.1 Numerical Integration

Sounds simple enough, let's make a uniform grid and compute the posterior at each point



How does this method scale with number or points and dimensions?

Instead it is more practical to use a sampling method, which can be *direct* or *stochastic*

5.2 Direct Sampling Methods

Method of Inversion

If we assume the pdf has a cdf F for random variable X then $F(X)$ follows a $U[0,1]$ distribution. Thus if we can analytically compute the inverse of the CDF we can sample from a uniform distribution. Ie if

$$F(x) = P(X \leq x)$$

And it has inverse F^{-1} then we:

1. Generate $u \sim U[0,1]$
2. Compute $x = F^{-1}(u)$

5.2 Direct Sampling Methods

Method of Inversion

Example: exponential distribution with parameter r Suppose we want to draw $X \sim \text{Exp}(r)$. The pdf of the exponential distribution is

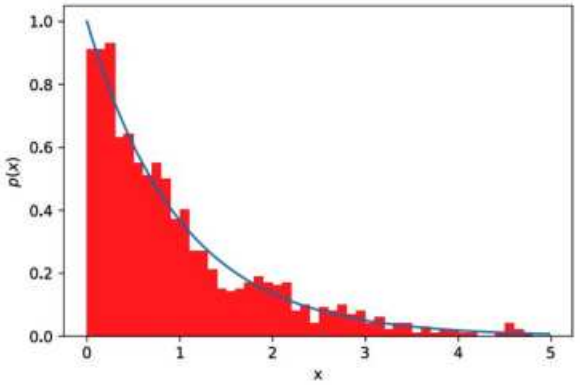
$$p(x|r) = r\exp(-rx)$$

which has cumulative density function

$$F(X) = \int_0^X r\exp(-rx)\mathrm{d}x = 1 - \exp(-rX).$$

The inverse can be found as

$$u = F(x) \quad \Rightarrow \quad x = F^{-1}(u) = -\frac{1}{r}\ln(1-u).$$



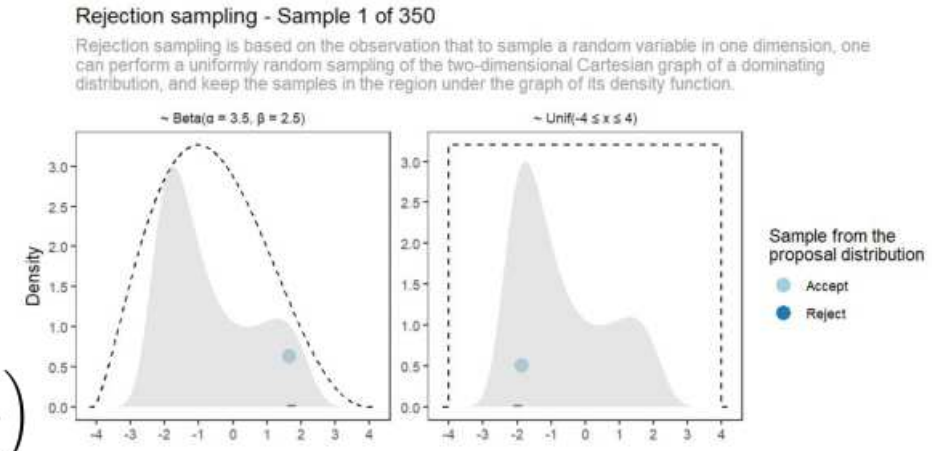
5.2 Direct Sampling Methods

Rejection Sampling

Samples are drawn from a distribution that can be directly sampled and contains the target distribution, and then discards a subset that do not match the desired distribution. The algorithm is

1. Draw $\theta \sim g(\theta)$
2. Draw $y \sim U[0, 1]$
3. If $y \leq p(\theta)/(Mg(\theta))$ accept θ as a sample from $p(\theta)$. Otherwise return to 1

With $Mg(\theta) \geq p(\theta)\ \forall\theta$ where $M = \sup_{\theta}\left(\frac{p(\theta)}{g(\theta)}\right)$



5.2 Direct Sampling Methods

Importance Sampling

What if we don't have a known distribution close to the target distribution? Also why waste computational power to throw away so many samples? Importance sampling uses all samples, but weights them with respect to importance:

$$w_i = \frac{p(\theta)}{g(\theta)}$$

Integrals over the target distribution become

$$\int f(\theta)p(\theta)\mathrm{d}\theta \approx \frac{1}{M}\sum_{i=1}^{M} w_i f(\theta_i)$$

Importance sampling is unbiased, however can have a large variance

5.2 Direct Sampling Methods

Sampling importance resampling

We may also use our importance samples to draw new samples from our target distribution. Given importance samples $\{\theta_1, \ldots, \theta_M\}$ with corresponding weights, you draw new samples $\{\phi_1, \ldots, \phi_M\}$ and replace the old ones. Integrals over the target distribution can then be approximated by

$$\int f(\theta)p(\theta)\mathrm{d}\theta \approx \frac{1}{M}\sum_{i=1}^{M} f(\phi_i)$$

This means that the low impotence samples are mostly discarded, however if the chosen distribution is not a good choice for the target distribution the few samples with weight can dominate and lead to high variance.

5.3 Markov chain Monte Carlo Methods

Direct sampling struggle to be extended to higher dimensional problems. Instead often *stochastic* methods are instead used - such as Markov chain Monte Carlo (MCMC) methods. This constructs a Markov chain (sequence of events where the probability of each events **only** depends on the previous event).

A Markov chain is a sequence where each value θ_{n+1} depends only on the previous value θ_n . It evolves via a transition kernel $K(\theta_{n+1}|\theta_n)$, which defines the chain. If the chain is *aperiodic* and *irreducible*, it will converge to a stationary distribution, regardless of the starting point. In Bayesian inference, we design the chain so that this stationary distribution is the posterior.

5.3 Markov chain Monte Carlo Methods

A Markov chain is a reversible Markov chain if it satisfies the **detailed balance** for a distribution $\pi(\theta)$

$$\pi(\theta)\mathcal{K}(\phi|\theta) = \pi(\phi)\mathcal{K}(\theta|\phi) \; \forall \, \phi, \theta$$

For us, the posterior

transition probability

The main approaches constructing Markov chains which satisfy detail balance:

- Gibbs Sampling
- Metropolis-Hastings Algorithm

5.3 Markov chain Monte Carlo Methods

Gibbs Sampling

Gibbs sampling works by sampling sequentially from full conditional distributions on each parameter given the current state of the other parameters:

- Sample θ_1^{t+1} from $p(\theta_1|\theta_2^t, \theta_3^t, \ldots, \theta_p^t, \mathbf{x})$.
- Sample θ_2^{t+1} from $p(\theta_2|\theta_1^{t+1}, \theta_3^t, \ldots, \theta_p^t, \mathbf{x})$.
- .....
- Sample θ_i^{t+1} from $p(\theta_i|\theta_j^{t+1}$ for $j < i$ and θ_j^t for $j > i, \mathbf{x})$.
- .....
- Sample θ_p^{t+1} from $p(\theta_p|\theta_1^{t+1}, \ldots, \theta_{p-1}^{t+1}, \mathbf{x})$.

5.3 Markov chain Monte Carlo Methods

Gibbs Sampling

This set of sequential updates is repeated at each iteration of the algorithm to generate a set of samples from the target distribution.

The transitional kernel in Gibbs sampling is

$$\mathcal{K}_G(\boldsymbol{\theta}^{t+1}|\boldsymbol{\theta}^t) = \prod_{i=1}^{k} p(\theta_i|\theta_j^{t+1} \text{ for } j < i \text{ and } \theta_j^t \text{ for } j > i, \mathbf{x})$$

which satisfies detailed balance.

5.3 Markov chain Monte Carlo Methods

Example: Gibbs Sampling

Problem - Estimating the mean and variance of of data assumed from a Normal distribution

Model - $x_i \sim N(\mu, \sigma^2)$ for data $x = (x_1, ..., x_n)$

Likelihood - $p(x|\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2} \right)$

Priors - $\mu|\sigma^2 \sim \mathcal{N}(\mu_0, \sigma^2/\kappa_0)$ $\sigma^2 \sim \text{InverseGamma}(\alpha_0, \beta_0)$

5.3 Markov chain Monte Carlo Methods

Example: Gibbs Sampling

Gibbs sampling

Step 1: Sample $\mu^{(t+1)}|\sigma^{2(t)}, x$

conditional posterior: $\mu|\sigma^2, x \sim \mathcal{N}(\mu_n, \sigma^2/\kappa_n)$

thus sample: $\mu^{(t+1)} \sim \mathcal{N}\left(\mu_n, \sigma^{t(t)}/\kappa_n\right)$

$$\kappa_n = \kappa_0 + n$$
$$\mu_n = \frac{\kappa_0\mu_0 + n\bar{x}}{\kappa_0 + n}$$

Step 2: Sample $\sigma^{2(t+1)}|\mu^{(t+1)}, x$

conditional posterior: $\sigma^2|\mu, x \sim \text{InverseGamma}(\alpha_n, \beta_n)$

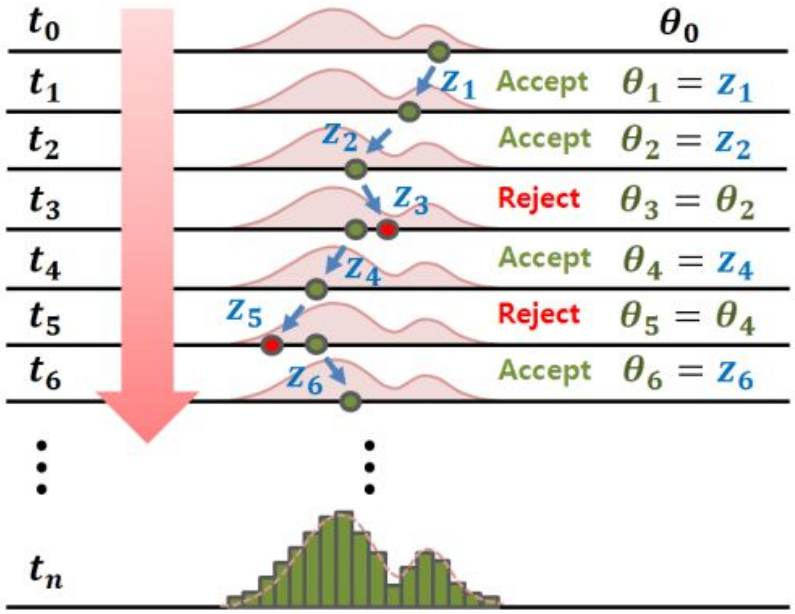
thus sample: $\sigma^{2(t+1)} \sim \text{InverseGamma}(\alpha_n, \beta_n)$

$$\alpha_n = \alpha_0 + n/2$$
$$\beta_n = \beta_0 + \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2 + \frac{\kappa_0 n}{2(\kappa_0 + n)}(\bar{x} - \mu_0)^2$$

5.3 Markov chain Monte Carlo Methods

Metropolis-Hastings algorithm

In the Metropolis-Hastings algorithm all parameters are updates simultaneously. This is achieved through a *proposal distribution* $q(\vec{\phi}, \vec{\theta})$ to propose a new point $\vec{\phi}$



5.3 Markov chain Monte Carlo Methods

Metropolis-Hastings algorithm

Algorithm:

1. Initialise $\vec{\theta}_0$ by drawing from a distribution of starting values (often from the prior)
2. At step t:
   a. Propose a new point $\vec{\phi} \sim q(\vec{\phi}|\vec{\theta}^{t-1})$
   b. Compute the *acceptance probability*

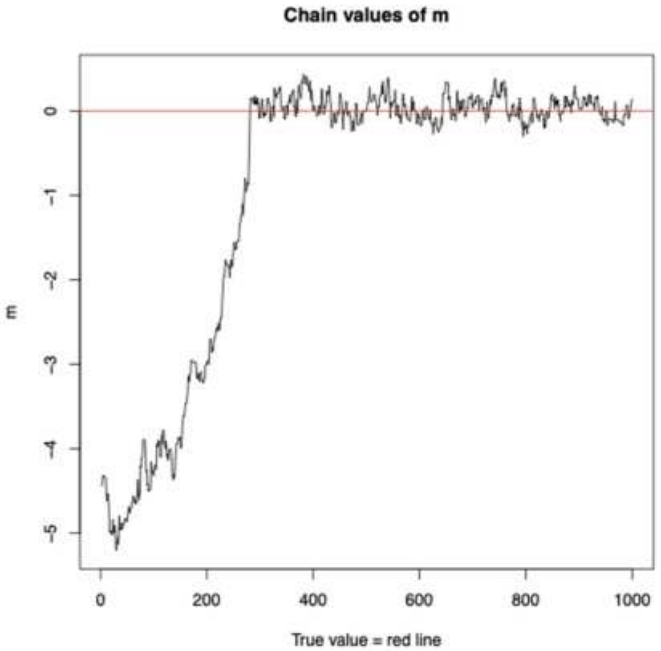
$$\alpha = \min\left(1, \frac{p(\vec{\phi}|\vec{x})q(\vec{\theta}^{t-1}|\vec{\phi})}{p(\vec{\theta}^{t-1}|\vec{x})q(\vec{\phi}|\vec{\theta}^{t-1})}\right)$$

   c. Draw $u \sim U[0,1]$, if $u < \alpha$ set $\vec{\theta}^t = \vec{\phi}$, otherwise set $\vec{\theta}^t = \vec{\theta}^{t-1}$
3. Repeat until the desired number of iterations have been completed

5.3 Markov chain Monte Carlo Methods

MCMC Diagnostics

- Often the starting point is not optimal - and it takes the sampler time to “feel the area out”.
- To account for this it’s common to discard a number of the first samples in the *burn in phase*.
- To diagnose how many burn in samples to discard we can use a *trace plot*



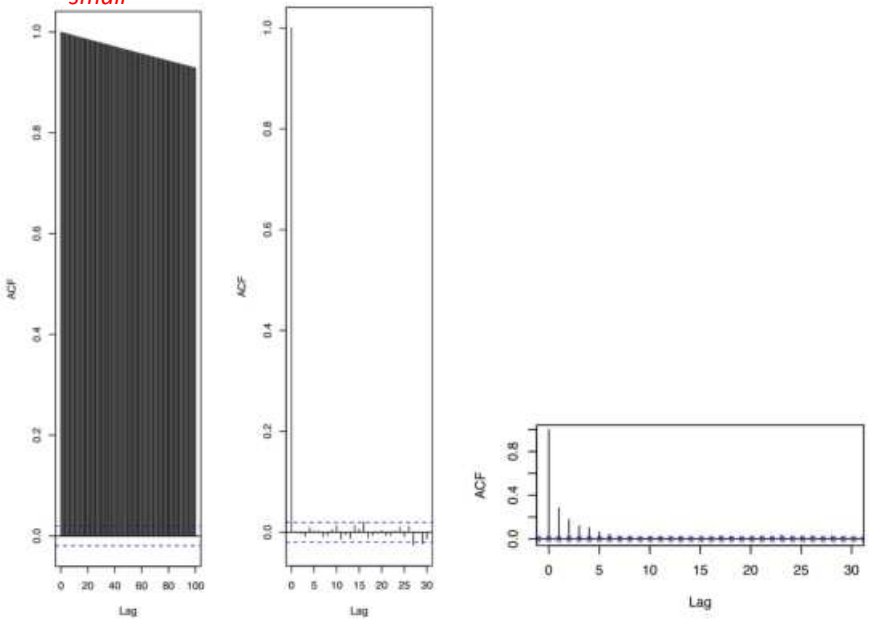
5.3 Markov chain Monte Carlo Methods

MCMC Diagnostics

- MCMC samples are not independent samples
- This modifies the variance to include the correlation
- The correlation can be estimated with the *autocorrelation function* (ACF).
- The autocorrelation at *lag-k* is $\mathrm{cov}(\theta^i, \theta^{i+k})$

$$\rho_k = \frac{\sum_{i=1}^{N-k}(\theta^i - \bar{\theta})(\theta^{i+k} - \bar{\theta})}{\sum_{i=1}^{M}(\theta^i - \bar{\theta})^2}$$

Jumps too small



bad *ideal* *normal*

5.3 Markov chain Monte Carlo Methods

MCMC Diagnostics

- If proposed jumps are too small, both the acceptance rates and autocorrelation function will be high
- If proposed jumps are too large, both the acceptance rates and autocorrelation function will be low
- We want to maximise the rate at which we obtain new independent samples
- We can track this with the *effective sample size* (ESS)

$$\mathrm{ESS} = \frac{M}{1 + 2\sum_{k=1}^{\infty} \rho_k}$$

M - number of samples in chain

5.3 Markov chain Monte Carlo Methods

MCMC Diagnostics

- For complex probability distributions, Markov chains can get "stuck" sampling from a local maximum and not the global maximum
- Therefore it's good practice to have multiple chains sample, and combine the results
- Consistency between chains is calculated with the *Gelman-Rubin statistic*

$$W = \frac{1}{m}\sum_{j=1}^{m}\frac{1}{N-1}\sum_{i=1}^{N}(\theta_{ij}-\bar{\theta}_j)^2$$

within-chain variance

$$B = \frac{N}{m-1}\sum_{j=1}^{m}(\bar{\theta}_j-\bar{\bar{\theta}})^2, \quad \text{where } \bar{\bar{\theta}} = \frac{1}{m}\sum_{j=1}^{m}\bar{\theta}_j.$$

between-chain variance

$$\text{var}(\theta) = \left(1-\frac{1}{N}\right)W + \frac{1}{N}B$$

variance of parameter

Potential scale reduction factor

$$\hat{R} = \sqrt{\frac{\text{var}(\theta)}{W}}$$

R > ~1.1 suggests chains are not converged

5.3 Markov chain Monte Carlo Methods

Speeding up MCMC

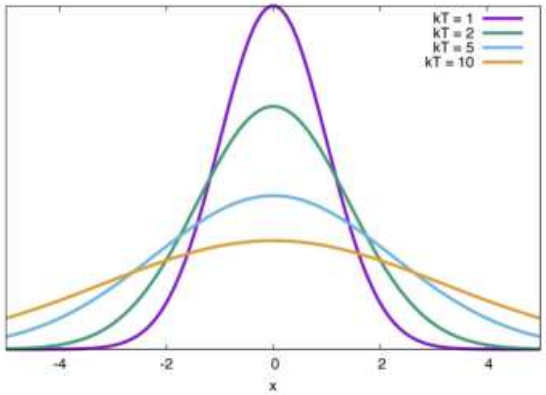
Aside from making a good choice for the proposal distribution, one can apply *annealing:*

- Transform the posterior with

$$p(\boldsymbol{\theta}|\mathbf{x}) \rightarrow [p(\boldsymbol{\theta}|\mathbf{x})]^{\beta}, \quad \text{where } \beta = \frac{1}{kT}$$

kT - temperature

- This smooths out the posterior and makes it easier to sample
- Allows identification of interesting parts of the parameter space



5.3 Markov chain Monte Carlo Methods

Speeding up MCMC

1. *Simulated annealing*:

The temperature is gradually decreased over the course of samping, allowing wide and rapid sampling in the early phase

2. *Parallel tempering*:

Chains are evolved at different temperatures, with a probability at each sampling step that the two chains swap parameters whilst maintaining detail balance

5.3 Markov chain Monte Carlo Methods

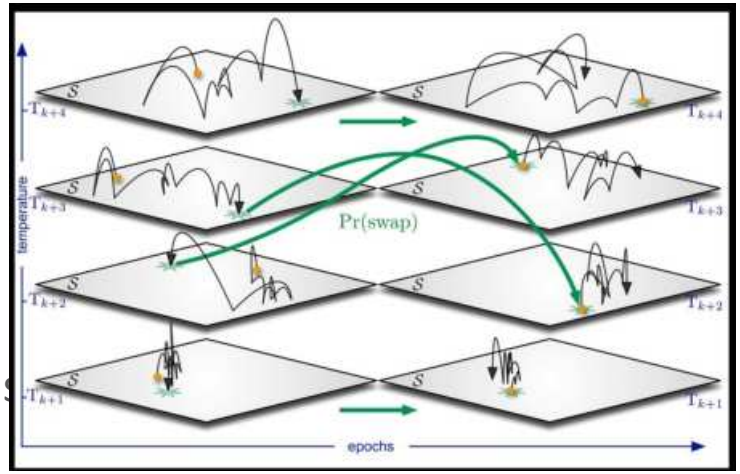
Speeding up MCMC

2. *Parallel tempering*:

Acceptance probability for the swap is

$$\alpha = \min\left(1, \left[\frac{p(\boldsymbol{\theta}^j|\mathbf{x})}{p(\boldsymbol{\theta}^i|\mathbf{x})}\right]^{\frac{1}{T_i}} \left[\frac{p(\boldsymbol{\theta}^i|\mathbf{x})}{p(\boldsymbol{\theta}^j|\mathbf{x})}\right]^{\frac{1}{T_j}}\right)$$

Higher posterior regions that high temperature chains find can be propagated through to low temperature chains and explored



5.4 Nested Sampling

- It is difficult to compute the Bayesian evidence with MCMC

$$Z = \int d\theta \mathcal{L}(\theta)\pi(\theta)$$

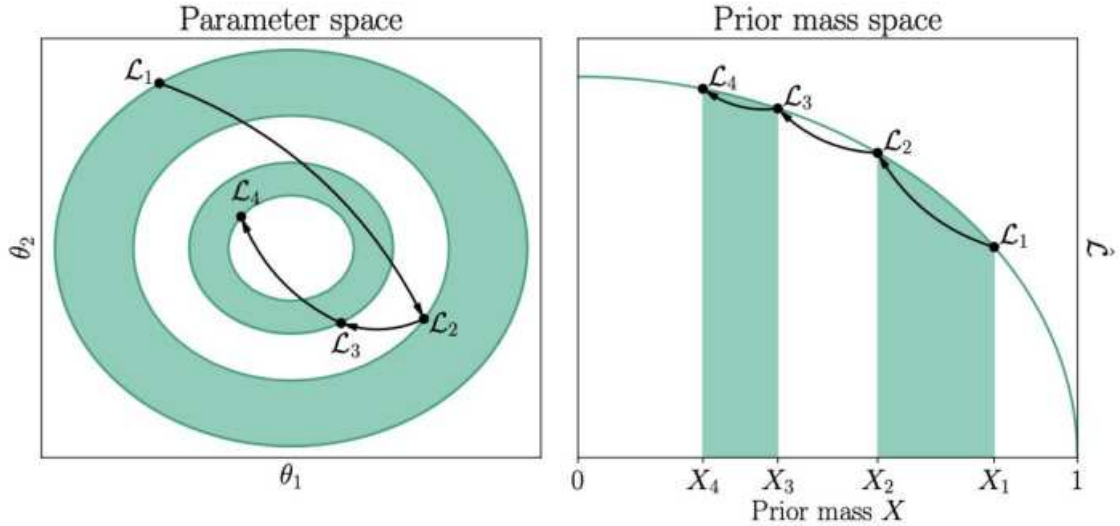
- Nested sampling was developed as an alternative which samples and also computes the evidence
- This is done by transforming the multi-dimensional evidence integral into a one dimensional integral
- The prior volume is written as

$$X(\lambda) = \int_{\mathcal{L}(\boldsymbol{\Theta})>\lambda} \pi(\boldsymbol{\Theta}) d^N\boldsymbol{\Theta} \quad \textit{where} \quad dX = \pi(\boldsymbol{\Theta}) d^D\boldsymbol{\Theta}$$

- The evidence can then be written as $\mathcal{Z} = \int_0^1 \mathcal{L}(X)dX \rightarrow \mathcal{Z} = \sum_{i=1}^{M} \mathcal{L}_i w_i$

5.4 Nested Sampling

$$\mathcal{Z} = \int_0^1 \mathcal{L}(X)dX \rightarrow \mathcal{Z} = \sum_{i=1}^{M} \mathcal{L}_i w_i \qquad 0 < X_M < \cdots < X_2 < X_1 < X_0 = 1$$



Visualize sampling [here]

5.4 Nested Sampling

1. Draw N samples from the full prior
2. Sort the samples in order of likelihood, denote the lowest likelihood as $\mathcal{L}_0$ and remove it from the dataset
3. Draw a new point from the prior given $\mathcal{L} > \mathcal{L}_0$. The prior volume within this iso-likelihood contour is $X_1 = t_1 X_0$ where $\mathbb{P}(t) = Nt^{N-1}$
4. Repeat all steps above, reducing the prior volume $X_i = t_i X_{i-1}$
5. Continue until some stopping criterion is achieved

Algorithm 1: Nested sampling

```
// Live points initialization
Draw K "live" points {θ⃗_1, θ⃗_2, ..., θ⃗_K} from the prior π(θ⃗).
i = 0
// Main sampling loop
while stopping criterion not met do
    Compute the likelihoods for the current set of live points
    Find the live point θ⃗_min associated with the minimum likelihood L_min
    // Likelihood constrained prior sampling loop
    while L(θ⃗') < L_min do
        Sample a new point θ⃗' from the prior π(θ⃗)
    end
    Add θ⃗_min to the list of "dead" points
    Replace θ⃗_min with θ⃗'
    L_i = L_min
    if i = 0 then
        Draw X_i from Beta(K, 1)
    end
    else
        Draw t from Beta(K, 1)
        X_i = t X_{i-1}
    end
    // Check whether to stop.
    Evaluate stopping criterion
    // Increment i
    i = i + 1
end
```

5.4 Nested Sampling

- The evidence is the calculated by

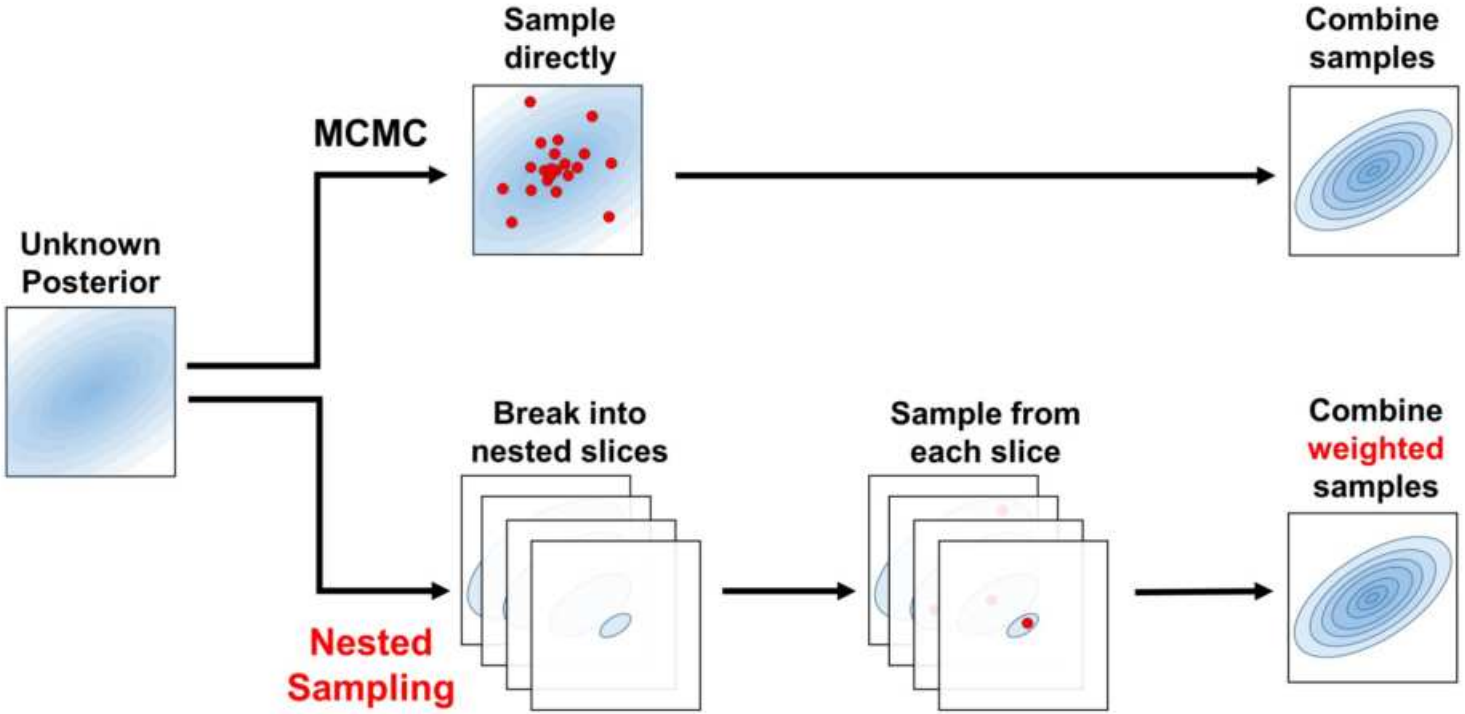
$$\begin{aligned} Z &= \int_0^1 dX\ \hat{\mathcal{L}}(X) \\ &\approx \frac{1}{2}\sum_{i=1}(\mathcal{L}_{i-1}+\mathcal{L}_i)(X_{i-1}-X_i) \\ &\equiv \sum_{i=1} w_i. \end{aligned}$$

- The posterior samples are obtained by resampling with the probability

$$p_i = \frac{w_i}{\sum_{i=1} w_i}$$

- Or the posterior is simply computed on the previously computed points

$$p_i = \frac{\mathcal{L}_i w_i}{\mathcal{Z}}$$


Sample directly
Combine samples
MCMC
Unknown Posterior
Break into nested slices
Sample from each slice
Combine weighted samples
Nested Sampling

6. Machine Learning and Applications

6.1 Machine Learning

- Computers are designed to complete repetitive tasks following a predescribed algorithm that takes an input and gives back an output
- Machine learning is the development of approaches that allow computers to learn/adjust how to perform a task, providing typically a large set of examples for the input and output expected (the *training set*)
- Machine learning algorithms typically consist of function approximators that have a large number of free parameters. These are designed in a way that allow the choice of parameters to be automatically optimised to minimise a specified objective function (the *loss function*).

6.1 Machine Learning

Examples of machine learning implementation:

- ***Regression***: Learn a function predicting real-valued quantities
- ***Sampling***: Generate new samples similar to training examples
- ***Denoising***: Given noisy data, predict clean data
- ***Density estimation***: Given training examples learn a probability density function
- ***Game playing***: What is the best move to make?

6.1 Machine Learning

We require some quantitative measure of performance:

- ***Regression***: The mean square error
- ***Classification:*** Accuracy (the fraction of examples that produce the output)
- ***Density estimation***: The log probability assigned to examples

We must assess this performance on data that was not used in the training dataset (*validation set*)

6.1 Machine Learning

There are different types of machine learning algorithms depending on your task

1. **Unsupervised**
   - ➢ Learn $p(x)$
   - ➢ Examples: density estimation, sampling
2. **Supervised**
   - ➢ Learn $p(y|x)$
   - ➢ Examples: regression, classification

Reinforcement learning allows the algorithm to interact with the environment and produce new samples

6.1 Maximum Likelihood Estimation

Unsupervised learning: Given data $x^i \sim p_{data}(x)$, let $p_{model}(x;\theta)$ be the ML model where we train such that θ is chosen so that $p_{model}(x;\theta)$ is a good approximation for $p_{data}(x)$

$$\theta_{\mathrm{ML}} = \arg\max p_{\mathrm{model}}(\mathbf{X};\theta)$$

$$\theta_{\mathrm{ML}} = \arg\max \prod_{i=1}^{N} p_{\mathrm{model}}(x^i;\theta)$$

$$\theta_{\mathrm{ML}} = \arg\max \sum_{i=1}^{N} \log p_{\mathrm{model}}(x^i;\theta)$$

$$\theta_{\mathrm{ML}} = \arg\max \mathbf{E}_{p_{\mathrm{data(x)}}} \log p_{\mathrm{model}}(x^i;\theta)$$

Negative of this would be the loss function

6.1 Conditional Estimation

Supervised learning: Estimate a conditional probability

Generalise the MLE

$$\theta_{\mathrm{ML}} = \arg\max \sum_{i=1}^{N} \log p_{\mathrm{model}}(y^i | x^i; \theta)$$

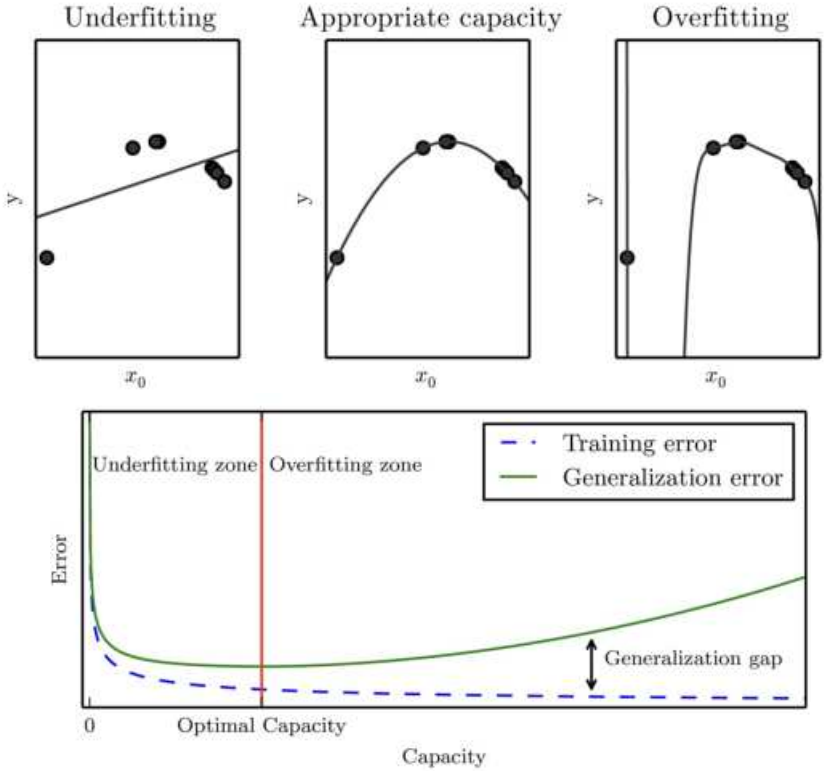
$$\theta_{\mathrm{ML}} = \arg\max \mathbf{E}_{p_{\mathrm{data(x,y)}}} \log p_{\mathrm{model}}(y | x^i; \theta)$$

e.g., source parameters we are interested in

6.1 Overfitting and Underfitting

Higher capacity models run the risk of overfitting. The algorithm must perform well not just on data used for training, but also on new, previously unseen inputs (test data). This is called **generalization**.

Training and test examples should be independent and identically distributed.



6.1 Regularisation

Regularization can be used to avoid overfitting

Regularization helps by simplifying the model, encouraging it to learn the underlying patterns in the data rather than memorizing it.

Principles:

Reduces Model Complexity: adding a penalty for larger weights

Enhances Generalization: A simpler model with smaller weights is less likely to overfit and thus generalizes better to new data.

Controls Overfitting: Regularization helps control the balance between fitting the training data well and keeping the model complexity low.

6.1 Machine Learning for Bayesian Statistics

Instead of just using the maximum likelihood, one can also tread the parameters in a Bayesian way

1. Specify a prior
2. Obtain posterior following Bayes Theorem

$$p(\theta, X) = \frac{p(X|\theta)p(\theta)}{p(X)}$$

This incorporates uncertainty associated to the choice of the parameters

The prior can here act directly as a regulator

6.1 Stochastic Gradient Descent

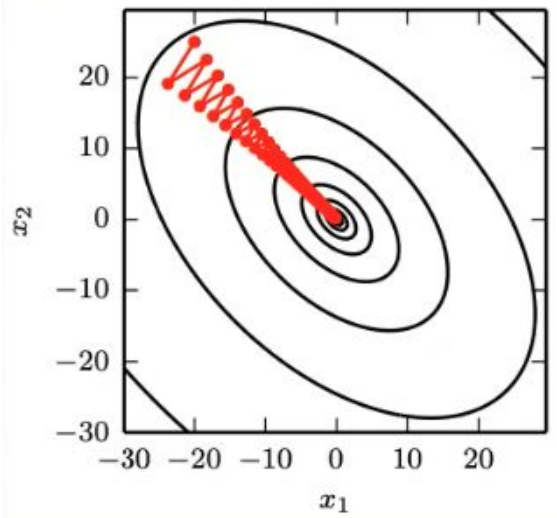
Iterative optimization algorithm used to find the minimum of a function

Minimizing the loss function with respect to the model parameters

$$\theta_1 = \theta_0 - \epsilon \nabla_\theta J|_{\theta_0}$$

learning rate

Instead of the entire dataset, in a stochastic approach you use a small batch of the data - helps with efficiency and not getting stuck in minima



6.1 Machine Learning Summary

1. Dataset (supervised or unsupervised)
2. Model (e.g. regression)
3. Loss function
4. Optimization algorithm (Stochastic gradient descent)

6.2 Neural Networks: Feedforward neural network

Feedforward (no feedback from later layers) neural networks are the classic deep learning model.

Mapping $y=f(x;\theta)$ as a composition of simpler mappings:

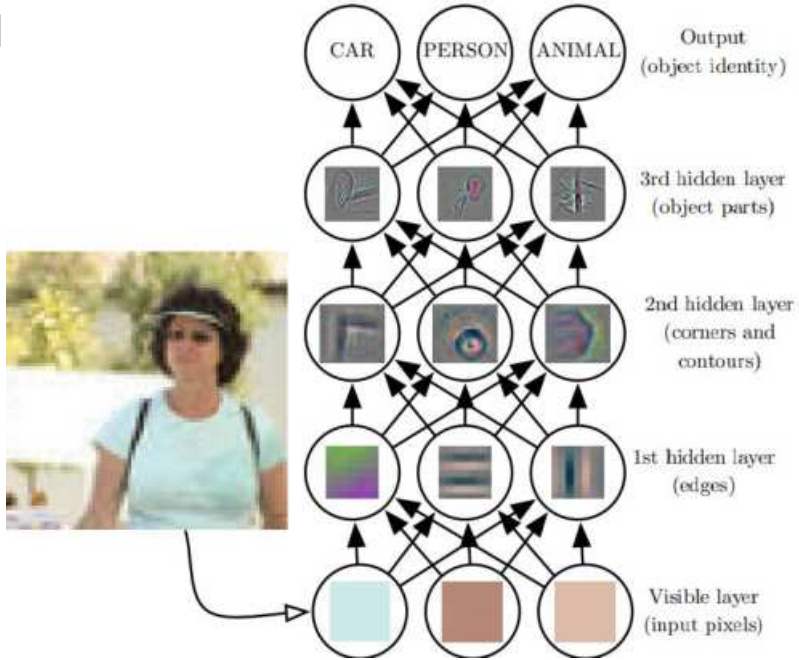
$$f = f^{(d)} \circ f^{(d-1)} \circ \ldots \circ f^{(2)} \circ f^{(1)}$$

output layer *hidden layers*

Form of individual layer: *linear mapping*

$$f^{(i)} = \sigma_j(W^T h + b_i)$$

activation function *weight* *bias*



6.2 Neural Networks: Feedforward neural network

The NN is therefore defined by:

1. Depth
2. Width
3. Choice of activation function

Training uses stochastic gradient descent with gradients calculated using back propagation (the chain rule).

6.2 Neural Networks: Output Layer

The output layer is dependent on the exact problem that you are investigating:

- For regression, you would output the mean value
- For binary classification, you would output a probability [0,1], you can use a Bernoulli distribution

Loss functions are generally rather simple, e.g., linear, but not exponential or anything similar. This is to ensure that during the minimization/optimization nothing is blowing up.

6.2 Neural Networks: Hidden Layers

Activation functions can introduce non-linearities

Most common activation function:

Rectified Linear Unit (ReLU)

$$f(x) = x^+ = \max(0, x) = \frac{x + |x|}{2} = \begin{cases} x & \text{if } x > 0, \\ 0 & \text{otherwise} \end{cases}$$

In addition, it is necessary to choose hyperparameters for depth and width of network.

Deeper and wider will give more representational capacity, although it may be harder to train and longer to run

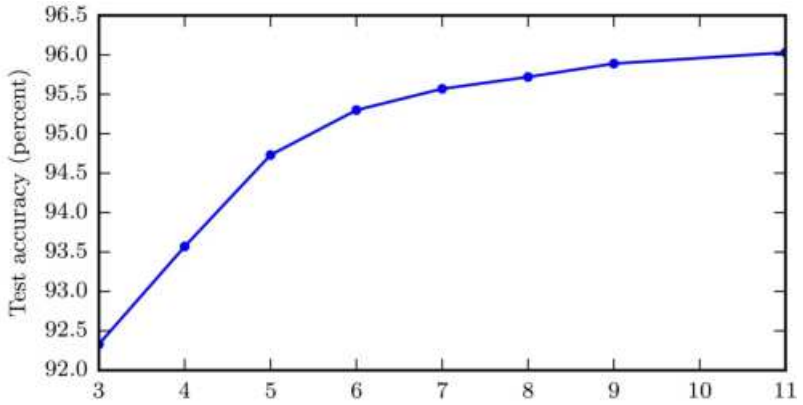
Deeper networks will usually require fewer total parameters than one very wide network

6.2 Neural Networks: Hidden Layers

Universal approximation theorem:

"A feedforward network with a linear output layer and at least one hidden layer can—given a wide enough hidden layer—approximate any (reasonable) function to arbitrary accuracy."

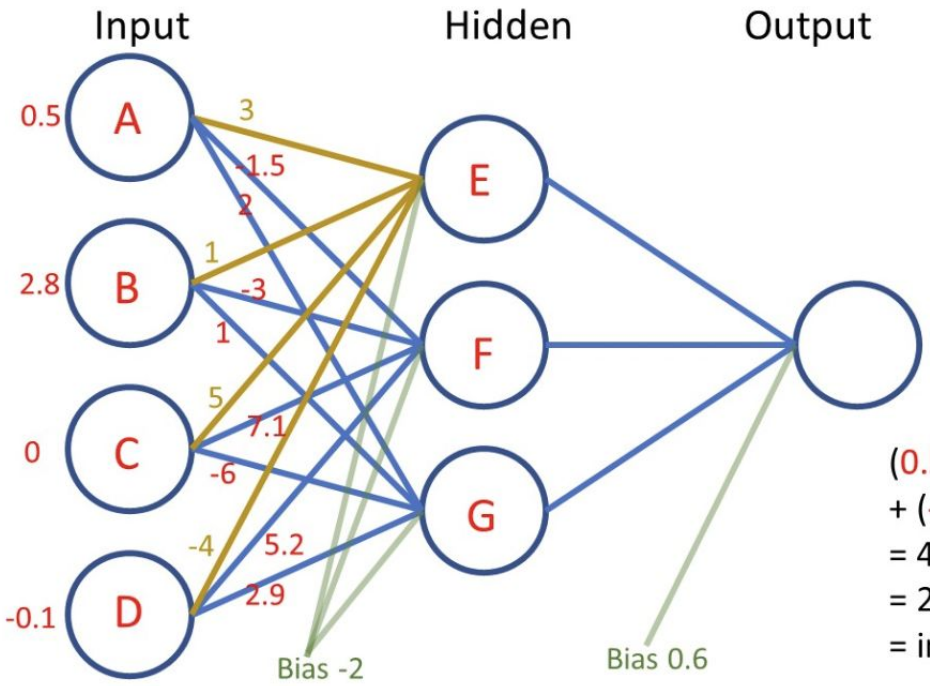
Might maybe lead to a very large hidden layer → deeper networks will usually require fewer total parameters than one very wide network.



6.2 Neural Networks: Example

$$f^{(i)} = \sigma_j(W^T h + b_i)$$

activation function — weight — bias

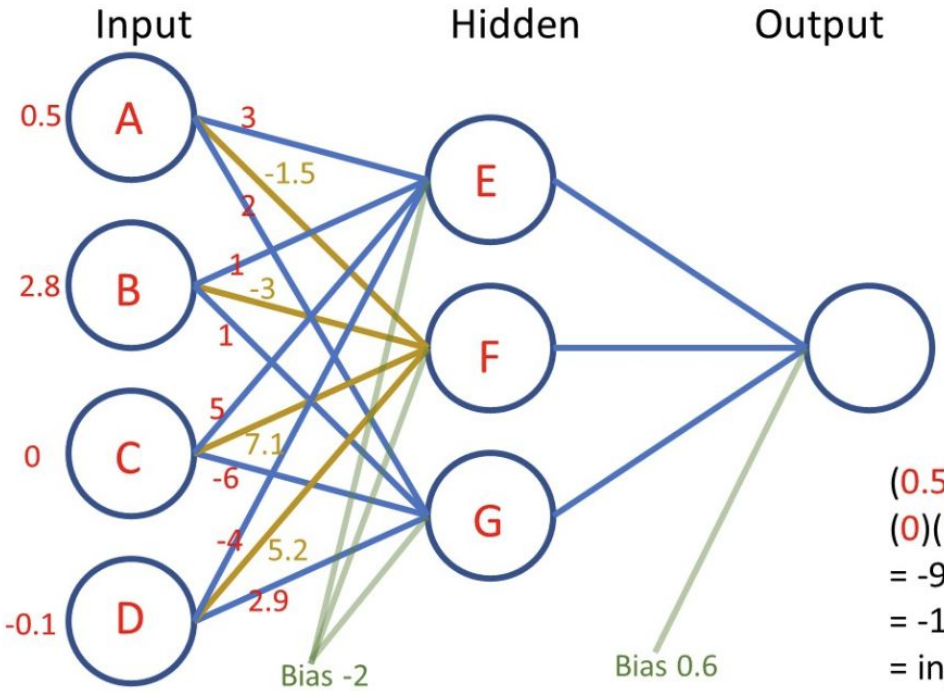


(0.5)(3) + (2.8)(1) + (0)(5)
+ (-0.1)(-4) + (-2)
= 4.7 -2
= 2.7
= input to hidden node E

6.2 Neural Networks: Example

$$f^{(i)} = \sigma_j(W^T h + b_i)$$

activation function — weight — bias

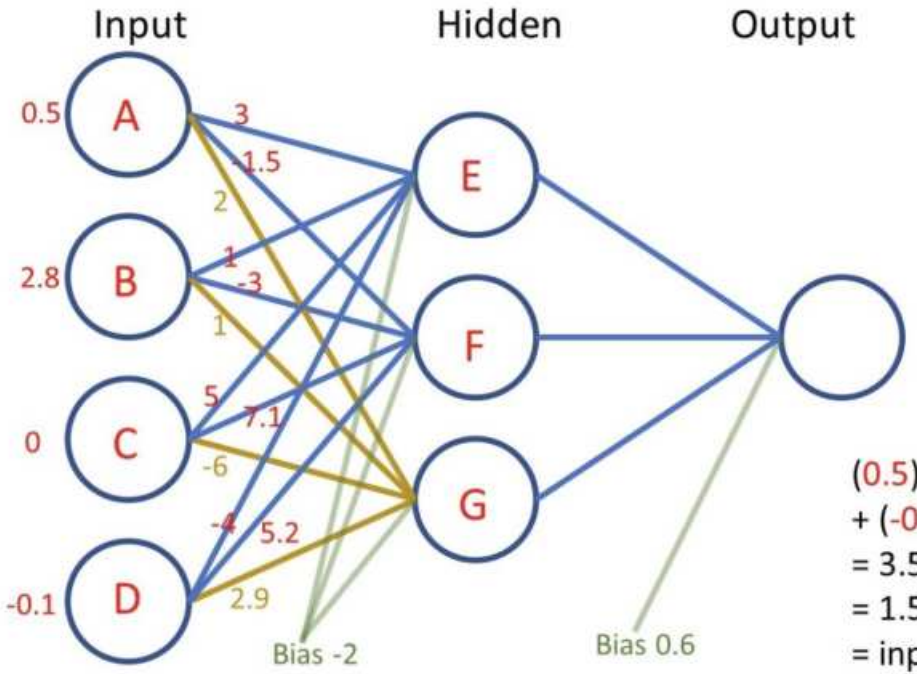


(0.5)(-1.5) + (2.8)(-3) +
(0)(7.1) + (-0.1)(5.2) + (-2)
= -9.67 -2
= -11.67
= input to hidden node F

6.2 Neural Networks: Example

$$f^{(i)} = \sigma_j(W^T h + b_i)$$

activation function — weight — bias



Wx+b

W (weights) x (input) b (bias)

$$\begin{bmatrix} 3 & 1 & 5 & -4 \\ -1.5 & -3 & 7.1 & 5.2 \\ 2 & 1 & -6 & 2.9 \end{bmatrix} \begin{bmatrix} 0.5 \\ 2.8 \\ 0 \\ -0.1 \end{bmatrix} + \begin{bmatrix} -2 \\ -2 \\ -2 \end{bmatrix}$$

$$= \begin{bmatrix} 4.7 \\ -9.67 \\ 3.51 \end{bmatrix} + \begin{bmatrix} -2 \\ -2 \\ -2 \end{bmatrix} = \begin{bmatrix} 2.7 \text{ in to E} \\ -11.67 \text{ in to F} \\ 1.51 \text{ in to G} \end{bmatrix}$$

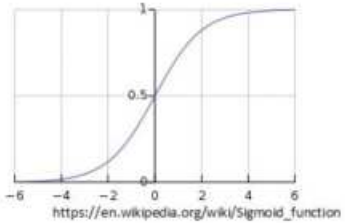
(0.5)(2) + (2.8)(1) + (0)(-6)
+ (-0.1)(2.9) + (-2)
= 3.51 -2
= 1.51
= input to hidden node G

6.2 Neural Networks: Example

$$f^{(i)} = \sigma_j(W^T h + b_i)$$

activation function — weight — bias

sigmoid function $\sigma(z) = \frac{e^z}{1+e^z}$



Node E $\sigma(2.7) = \frac{e^{2.7}}{1+e^{2.7}} = 0.937$

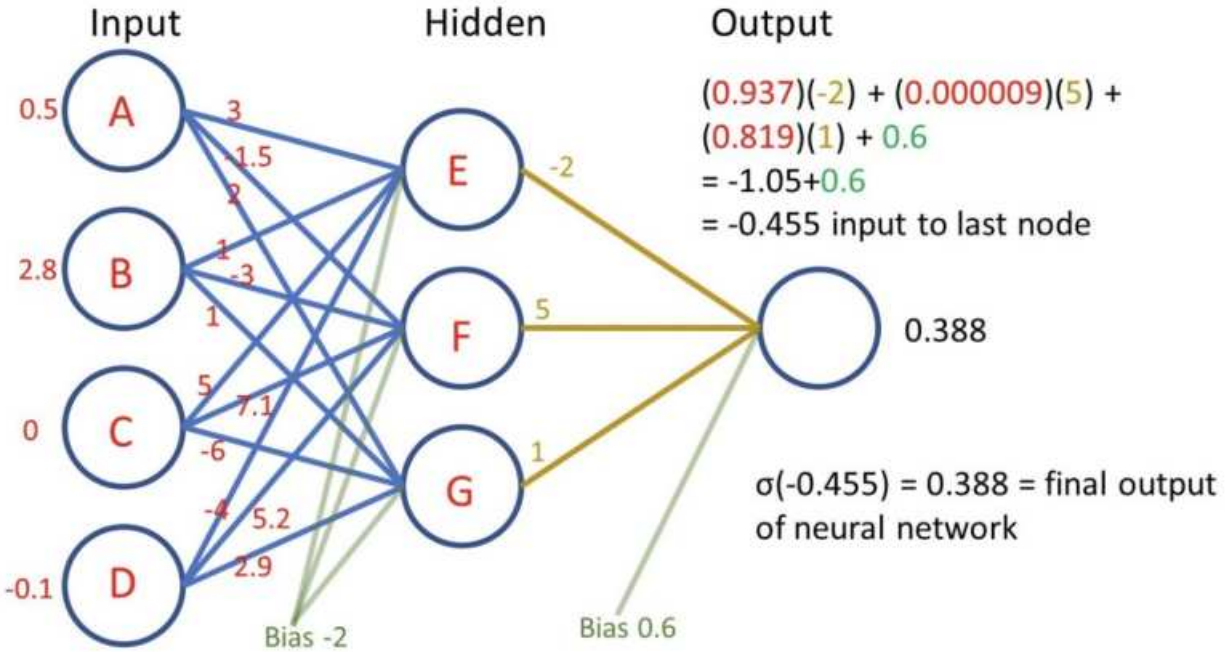
Node F $\sigma(-11.67) = \frac{e^{-11.67}}{1+e^{-11.67}} = 0.000009$

Node G $\sigma(1.51) = \frac{e^{1.51}}{1+e^{1.51}} = 0.819$

6.2 Neural Networks: Example

$$f^{(i)} = \sigma_j(W^T h + b_i)$$

activation function — weight — bias

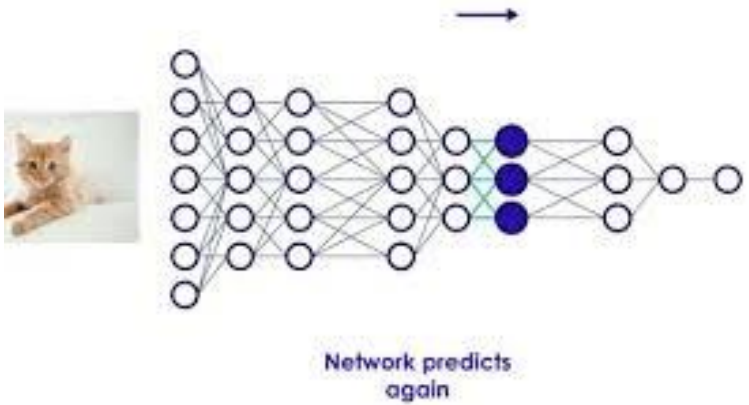


6.2 Neural Networks: Back-propagation

To train the network using some form of gradient descent, it is necessary to be able to efficiently compute gradients with respect to all of the network parameters (weights and biases).

This is accomplished using a form of automatic differentiation called backpropagation.

Relies on compositional nature of neural networks plus the chain rule of calculus and differentiability of all operations.



$$f = f^{(d)} \circ f^{(d-1)} \circ \ldots \circ f^{(2)} \circ f^{(1)}$$

6.2 Neural Networks: Training

1. Initialization

- Weights and Biases Initialization: The neural network starts with random weights and biases.

2. Forward Pass

- Input to Output: The input data is passed through the network layer by layer, with each neuron applying a weighted sum of inputs and an activation function. This process continues until the output layer generates the final predictions.

3. Compute Loss

- Loss Calculation: The difference between the predicted output and the actual target output is calculated using a loss function (e.g., Mean Squared Error, Cross-Entropy Loss). This loss quantifies how well the network is performing.

6.2 Neural Networks: Training

4. Backward Pass (Back-Propagation)

- Gradient Calculation: Starting from the output layer, the algorithm computes the gradient of the loss with respect to each weight and bias:

- Chain Rule of Calculus: The gradients are computed using the chain rule to propagate the error backward through the network layers.

- Partial Derivatives: For each neuron, partial derivatives of the loss with respect to its weights and biases are calculated.

5. Weight Update

- Adjusting Weights and Biases: The computed gradients are used to update the weights and biases to minimize the loss. This is typically done using a method like Gradient Descent or one of its variants (e.g., Stochastic Gradient Descent, Adam).

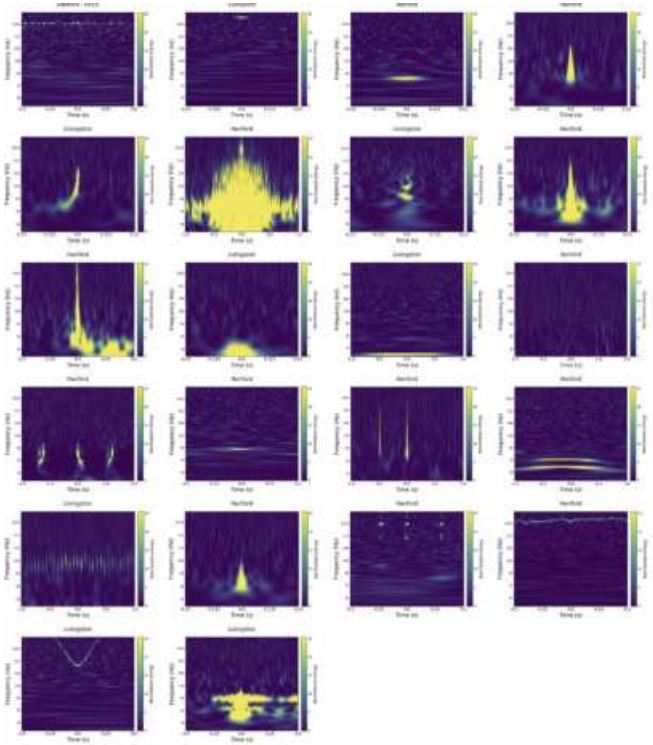
6. Iteration

6.2 Neural Networks: Applications in GW Astronomy

Machine learning for Gravity Spy: Glitch classification and dataset

S. Bahaadini [a], V. Noroozi [b], N. Rohani [a], S. Coughlin [c d], M. Zevin [c d], J.R. Smith [e], V. Kalogera [c d], A. Katsaggelos [a]

Class	Total	# train set	# valid set	# test set	Duration	Frequency	Evolving
1080Lines	328	230	49	49	Long	High	No
1400Ripples	232	162	35	35	Short	High	No
Air Compressor	58	41	8	9	Short	Low	No
Blip	1869	1308	281	280	Short	Mid	Yes
Chirp	66	46	10	10	Short	Mid, Low	Yes
Extremely Loud	454	318	68	68	Long	High, Mid, Low	Yes
Helix	279	195	42	42	Short	Mid	Yes
Koi Fish	830	581	125	124	Short	Mid, Low	Yes
Light Modulation	573	401	86	86	Long	Mid, Low	Yes
Low Frequency Burst	657	460	99	98	Short	Low	Yes
Low Frequency Lines	453	317	68	68	Long	Low	No
No Glitch	181	127	27	27	Long	–	No
None of the Above	88	62	13	13	Short	High, Mid, Low	Yes
Paired Doves	27	19	4	4	Short	Mid, Low	Yes
Power Line	453	317	68	68	Short	Low	No
Repeating Blips	285	200	69	42	Short	Mid	No
Scattered Light	459	321	69	69	Long	Low	Yes
Scratchy	354	248	53	53	Long	High, Mid	Yes
Tomte	116	81	17	18	Short	Low	Yes
Violin Mode	472	330	71	71	Short	High	No
Wandering Line	44	31	6	7	Long	High	Yes
Whistle	305	213	46	46	Short	High	Yes



6.2 Neural Networks: Applications in GW Astronomy

Reduced-order modeling with artificial neurons for gravitational-wave inference

Alvin J. K. Chua,* Chad R. Galley,† and Michele Vallisneri‡
Jet Propulsion Laboratory, California Institute of Technology,
4800 Oak Grove Drive, Pasadena, CA 91109, U.S.A.
(Dated: May 31, 2019)

Gravitational-wave data analysis is rapidly absorbing techniques from deep learning, with a focus on convolutional networks and related methods that treat noisy time series as images. We pursue an alternative approach, in which waveforms are first represented as weighted sums over reduced bases (reduced-order modeling); we then train artificial neural networks to map gravitational-wave source parameters into basis coefficients. Statistical inference proceeds directly in coefficient space, where it is theoretically straightforward and computationally efficient. The neural networks also provide analytic waveform derivatives, which are useful for gradient-based sampling schemes. We demonstrate fast and accurate coefficient interpolation for the case of a four-dimensional binary-inspiral waveform family, and discuss promising applications of our framework in parameter estimation.

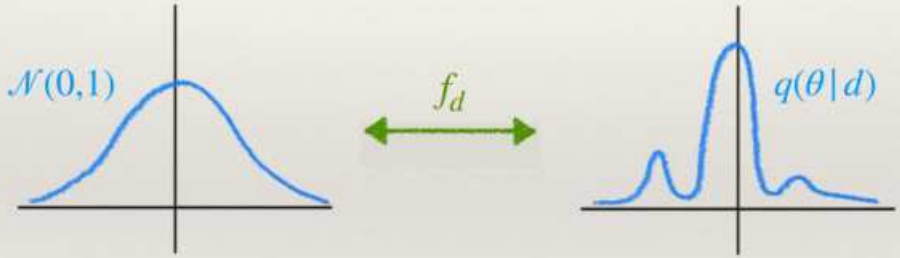
$$h(\theta) = \sum_i \langle h(\theta)|e_i\rangle e_i \coloneqq \sum_i \alpha_i(\theta) e_i \equiv \alpha(\theta)$$

6.2 Neural Networks: Applications in GW Astronomy

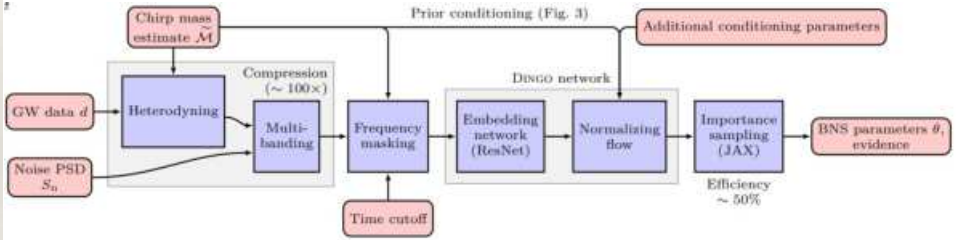
Real-time gravitational-wave inference for binary neutron stars using machine learning

Maximilian Dax,[1,*] Stephen R. Green,[2,†] Jonathan Gair,[3] Nihar Gupte,[3,4] Michael Pürrer,[5,6] Vivien Raymond,[7] Jonas Wildberger,[8] Jakob H. Macke,[1,9] Alessandra Buonanno,[3,4] and Bernhard Schölkopf[1,8]

A **normalizing flow** $f_d : u \mapsto \theta$ defines a complex distribution in terms of a simple one



$$q(\theta|d) = \mathcal{N}(0,1)^D\left(f_d^{-1}(\theta)\right)\left|\det J_{f_d}^{-1}\right|$$



End of Course!

$$p(\vec{\theta}|\mathbf{x}) = \frac{p(\mathbf{x}|\vec{\theta})p(\vec{\theta})}{p(\mathbf{x})}$$

Please do the course evaluation!