Gravitational-Waves Class:
Theory of Probability

Matteo Breschi
matteo.breschi@uni-jena.de

Theoretisch-Physikalisches Institut,
Friedrich-Schiller-UniversitÃďt Jena

So.Se. 2020

Contents

Chapter 1

Introduction

The term *probability* is largely used in the scientific community since, with this concept, we are able to predict the result of many events belonging to different fields of knowledge. The concept of probability is related to statistics, i.e. the interpretation of randomness. A clear example is the game theory or

However, what is a probability under a mathematical and more rigorous point of view? This theory can be studied using two different approaches:

- **Frequentist approach**: the probabilities describe frequencies of outcome in a random experiments, where with frequencies we mean the average fraction of positive (or negative) results; this approach implicitly assumes that the experiment can be repeated as many times we want.

- **Bayesian approach**: a more general use of probabilities is referred to the *degree of belief* in propositions that do not involve random variable, for example: the probability that Mr. A was the murder of Mr. B, given the evidence.

We are interested in the latter case, the Bayesian approach, because we are able to use probabilities to describe *inferences*, i.e. to associate quantitative values which predict the outcome of an event or describe the adaptability of a model given a set of data. For these notes, we use as reference textbooks the ones listed in the Bibliography, [1–6].

The modern logical formulation of statistics was developed by Thomas Bayes in his published works [1] of 1763 and was successfully used by Pierre-Simon Laplace (1812) in various scientific fields, from astronomy to medicine. These works are elegantly shown in his book [2], in which he asserted: *la thÃĺorie des probabilitÃĺs n'est que le bon sens reduit au calcul.* However, these works was largely forgotten and discredited until they were rediscovered by Harold Jeffreys (1939) and, in more recent times, they have been expounded.

The problem was not really matter of substance but concept. To the pioneers such as Bayes and Laplace, a probability represented a plausibility: how much they thought that something was true, based on the evidence at hand. To the 19th century scholars, however, this seemed too vague and subjective an idea to be the basis of a rigorous mathematical theory. So they redefined probability as the long-run relative frequency with which an event occurred, given many repeated (experimental) trials. Since frequencies can be measured, probability was now seen as an objective tool for dealing with random phenomena.

Although the frequency definition appears to be more objective, its range of validity is also far more limited. For example, Laplace used his probability theory to estimate the mass of Saturn, given orbital data that were available to him from various astronomical observatories.

He computed the posterior probability (the meaning of this term will be clear later) for the mass of the planet, given the data and all the relevant background information, such as a knowledge of the laws of classical mechanics. Laplace stated that: *it is a bet of 11,000 to 1 that the error of this result is not 1/100th of its value.* He would have won the bet, as another 150 years' accumulation of data has changed the estimate by only 0.63%! According to the frequency definition, however, we are not permitted to use probability theory to tackle this problem. This is because the mass of Saturn is a constant and not a random variable; therefore, it has no frequency distribution and so probability theory cannot be used.

The frequency definition of probability merely gives the impression of a more objective theory. In reality it just makes life more complicated by hiding the difficulties under the rug, only for them to resurface in a less obvious guise. Indeed, it is not even clear that the concept of *randomness* central to orthodox statistics is any better-defined than the idea of *uncertainty* inherent in Bayesian theory of probability. For example, we might think that the numbers generated by a call to a function like `rand` on a computer constitutes a random process: the frequency of the numbers will be distributed uniformly between 0 and 1, and their sequential order will appear haphazard. The illusory nature of this randomness would become obvious, however, if we knew the algorithm and the seed for the function `rand` (for then we could predict the sequence of numbers output by the computer). At this juncture, some might argue that, in contrast to our simple illustration above, chaotic and quantum systems provide examples of physical situations which are intrinsically random. Either way, randomness is what we call our inability to predict things which, in turn, reflects our lack of knowledge about the system of interest. This is again consistent with the Bayes and Laplace view of probability, rather than the asserted physical objectivity of the frequentist approach.

The concerns about the subjectivity of the Bayesian view of probability are understandable, and the aim of creating an objective theory is quite laudable. Unfortunately, the frequentist approach does not achieve this goal: neither does its concept of randomness appear very rigorous, or fundamental, under scrutiny and nor does the arbitrariness of the choice of the statistic make it seem objective. In fact, the presumed shortcomings of the Bayesian approach merely reflect a confusion between subjectivity and the difficult technical question of how probabilities should be assigned. The popular argument goes that if a probability represents a *degree of belief*, then it must be subjective because my belief could be different from yours. The Bayesian view is that a probability does indeed represent how much we believe that something is true, but that this belief should be based on all the relevant information available. While this makes the assignment of probabilities an open-ended question, because the information at my disposal may not be the same as that accessible to you, it is not the same as subjectivity. It simply means that probabilities are always conditional, and this conditioning must be stated explicitly. As Jaynes has pointed out, objectivity demands only that two people having the same information should assign the same probability; this principle has played a key role in the modern development of the Bayesian approach.

1.1 Cox's Axioms

In 1946, Richard Cox tried to get away from the controversy of the Bayesian versus frequentist view of probability. He decided to look at the question of plausible reasoning afresh, from the perspective of logical consistency. He found that the only rules which met his requirements were those of probability theory. Using the approach of Bayes and Laplace, he observed that the degrees of belief can be mapped onto probabilities if satisfy some simple consistency rules, called the **Cox's axioms**. Calling x an event and $P(x)$ its degree of belief, the Cox's axioms postulate:

1. Degrees of belief can be ordered, so if $P(x) > P(y)$ and $P(y) > P(z)$ then
$$P(x) > P(z)\,.$$
This condition implies that the degrees of belief can be mapped onto real numbers.

2. Exists a function f which relate a proposition x to its negation $\bar{x}$, such that:
$$P(x) = f\big[P(\bar{x})\big]\,.$$

3. Exists a function g which relate a conjunction of propositions $x \cap y$ to the conditional proposition $x|y$ and the proposition y, such that:
$$P(x \cap y) = g\big[P(x|y), P(y)\big]\,.$$

Let us observe that $P(x \cap y)$ is the probability of the event x *and* the event y, i.e. the probability that both events are realized. Usually the intersection symbol is replaced by a comma in order to achieve a more elegant form, i.e. $P(x \cap y) = P(x, y)$, however when it is needed we will explicitly write this symbol (see Sec. 1.2). Moreover, the conditional probability $P(x|y)$ is the probability of x given y, i.e. the probability of x under the assumption that y is occurred.

Notation: From this moment, if x is an event, we use $P(x)$ to denote the probability of the event x and $\bar{x}$ is the logical negation of the event x, with probability $P(\bar{x}) = 1 - P(x)$. Then, if y is another event, we can construct the following terms:

- $P(x \cup y)$, the probability that x or y occur;
- $P(x \cap y) = P(x, y)$, the probability that x and y occur;
- $P(x|y)$, the probability that x occurs given that y occurred.

Note that we use the symbols $\cup$, $\cap$ to denote the logical operator, respectively *or*, *and*. This is motivated by the fact that in the theory of sets, these logical operators are represented by the operations of union and intersection.

Thus probabilities can be used to describe assumption and describe inferences given those assumptions. The rules of probability ensure that if two people make the same assumptions and receive the same data then they will draw identical conclusions. This more general use of probability to quantify beliefs is known as the Bayesian viewpoint. It is also known as the subjective interpretation of probability, since the probabilities depend on assumptions. Advocates of a Bayesian approach to data modeling and pattern recognition do not view this subjectivity as a defect, since in their view, you cannot do inference without making assumptions.

From the Cox's axioms follow the sum and the product rule of probabilities, that is

$$P(x) + P(\bar{x}) = 1\,, \tag{1.1}$$

$$P(x \cap y) = P(x|y) \cdot P(y)\,, \tag{1.2}$$

The first equation imposes that the sum (or the integral) of the probabilities of all possible results is normalized to unity. In the second equation we write the conditional probability of x given y, that means the probability of x assuming y as true.

1.2 Basic calculations

In this section we recall the basic information needed to carry out simple probability computations, such as additions and multiplications, for different cases. First of all we remark that, if x and y are two events, $x \cap y \equiv (x, y)$ is the intersection of two events and it can be red as "*x and y*", while $x \cup y$ is the union of two events and it can be red as "*x or y*".

1.2.1 Addition

If x and y are two *disjointed* events (i.e. these two events cannot happen at the same time in a single experiment), the probability that x or y $(x \cup y)$ will occur can be written as

$$P(x \cup y) = P(x) + P(y)\,. \tag{1.3}$$

Exercise: Compute the probability of rolling 2 or 3 on a 6-sided die.

If x and y are two *compatible* events, in general they could happen simultaneously and then favorable cases can arise fo both events that can be counted both for $P(x)$ and $P(y)$. In this condition, if we apply Eq. (1.3), these favorable cases are double counted. From this argumentation, it follows that the generalization of Eq. (1.3) can be written as

$$P(x \cup y) = P(x) + P(y) - P(x \cap y)\,. \tag{1.4}$$

From Eq. (1.3) and Eq. (1.4), it follows that if two events are disjointed, the intersection of their representative subsets is emply, $P(x \cap y) = 0$.

Exercise: Suppose we draw a card from a 52 card deck. Which is the probability that the drawn card is a king or is of aces?

1.2.2 Multiplication

If x and y are two *independent* events (i.e. not causally connected, in other words the realization of x does not affect the realization of y and viceversa), the probability that x and y $(x \cap y)$ will occur can be written as

$$P(x \cap y) = P(x) \cdot P(y)\,. \tag{1.5}$$

Exercise: Suppose we toss a coin and we roll a die. Which is the probability of rolling a 6 on the die and a cross on the coin?

Exercise: Suppose we roll two die. Which is the probability of rolling a 5 and 6? (independently from the order)

The multiplication rule Eq. (1.5) can be generalized for the case of non independent events. In this case we have to involve the conditional probability $P(x|y)$ which is the probability that x occurs once y has occurred. The generalization of Eq. (1.5) coincides with Eq. (1.2),

$$P(x \cap y) = P(x|y) \cdot P(y) = P(y|x) \cdot P(x)\,. \tag{1.6}$$

Then it is obvious to observe that, if x and y are independent then

$$P(x|y) = P(x|\bar{y}) = P(x)\,.$$

1.3 Bayes' Theorem

The relations in Eq. (1.1)-(1.2) represent the basic algebra of probability. Many other results can be derived from them, amongst the most useful are the **Bayes' theorem**,

$$P(x|y) \cdot P(y) = P(y|x) \cdot P(x)\,, \tag{1.7}$$

and the **marginalization rule**,

$$P(x) = \int P(x,y)\,dy = \int P(x|y)\,P(y)\,dy\,, \tag{1.8}$$

where we used Eq. (1.2) for the second equality and the integrals are extended to the entire domain of the variable of interest y.

The importance of these properties to data analysis becomes apparent if we replace x and y by the statistical quantities of importance. Consider, under certain hypothesis $\mathscr{H}$, a physical event expressed by the set of parameters $\boldsymbol{\theta}$ and supposing to have carried out an experiment obtaining the data $\mathbf{d}$. So, thanks to the Bayes' theorem we can write,

$$\underbrace{P(\mathbf{d}|\boldsymbol{\theta},\mathscr{H})}_{\text{likelihood function } (\mathcal{L})} \cdot \underbrace{P(\boldsymbol{\theta}|\mathscr{H})}_{\text{prior prob. } (\Pi)} = \underbrace{P(\boldsymbol{\theta},\mathbf{d}|\mathscr{H})}_{\text{joint prob.}} = \underbrace{P(\boldsymbol{\theta}|\mathbf{d},\mathscr{H})}_{\text{posterior prob. } (\mathcal{P})} \cdot \underbrace{P(\mathbf{d}|\mathscr{H})}_{\text{evidence } (\mathcal{Z})}\,. \tag{1.9}$$

Usually in the literature, Eq. (1.9) is expressed using a more compact form,

$$\mathcal{L}(\boldsymbol{\theta})\,\Pi(\boldsymbol{\theta}) = \mathcal{P}(\boldsymbol{\theta})\,\mathcal{Z}\,,$$

however, in this notes we will use the full forms since this reduced expression does not highlight the meaning of the different quantities and it could lead to misunderstandings. Every term in Eq. (1.9) is conditioned from the initial hypothesis $\mathscr{H}$. In general during our analysis, we always start making some assumption, because there is no such thing as an absolute probability, and so the probability of an event, is the conditional probability on that assumptions. Although the conditioning on some hypothesis $\mathscr{H}$ is often omitted in calculations, to reduce algebraic cluttering, we must never forget its existence.

The power of Bayes' theorem lies in the fact that it relates the quantity of interest, the probability that the model is true given the data (that is the posterior probability), to the term we have a better chance of being able to assign, the probability that we would have observed the measured data if the hypothesis was true. The **prior probability** $P(\boldsymbol{\theta}|\mathscr{H})$ represents our state of knowledge, or ignorance, about the truth of the hypothesis before we have analyzed the current data. This is modified by the experimental measurements through the **likelihood function** $P(\mathbf{d}|\boldsymbol{\theta},\mathscr{H})$, which is not properly a probability since it describes the plausibility of a model, given specific observed data. The result yields the **posterior probability** $P(\boldsymbol{\theta}|\mathbf{d},\mathscr{H})$, representing our state of knowledge about the truth of the model in the light of the data. In a sense, Bayes' theorem encapsulates the process of learning since if we are interested in measuring the same quantity with different experiments we are allowed to use the posterior probability as prior probability for the new measurement.

We should note, however, that the equality of Eq. (1.7) is often replaced with a proportionality, because the term $P(\mathbf{d}|\mathscr{H})$ is omitted. This is fine for many data analysis problems, such as those involving parameter estimation, since the missing denominator is simply a normalization constant, which do not depend explicitly on the hypothesis. In some situations, like model selection, this term plays a crucial role. For that reason, it is given the name of **evidence**

$P(\mathbf{d}|\mathscr{H})$. This crisp single word captures the significance of the entity, however this quantity is also labelled as marginal likelihood, since using Eq. (1.7) and Eq. (1.8), it can be computed as

$$P(\mathbf{d}|\mathscr{H}) = \int_{\boldsymbol{\Theta}} P(\mathbf{d}|\boldsymbol{\theta},\mathscr{H})\, P(\boldsymbol{\theta}|\mathscr{H})\, d\boldsymbol{\theta}\,, \tag{1.10}$$

where $\boldsymbol{\Theta}$ is the entire domain of $\boldsymbol{\theta}$. Combining Eq. (1.10) with Eq. (1.9), we get

$$P(\boldsymbol{\theta}|\mathbf{d},\mathscr{H}) = \frac{P(\mathbf{d}|\boldsymbol{\theta},\mathscr{H})\, P(\boldsymbol{\theta}|\mathscr{H})}{\int_{\boldsymbol{\Theta}} P(\mathbf{d}|\boldsymbol{\theta},\mathscr{H})\, P(\boldsymbol{\theta}|\mathscr{H})\, d\boldsymbol{\theta}}\,. \tag{1.11}$$

This latter form highlights the fact that the evidence works like a renormalization term for the determination of the posterior distribution $P(\boldsymbol{\theta}|\mathbf{d},\mathscr{H})$.

Exercise: You are planning a barbecue with your friends, but the morning sky is cloudy. You have the following information:

- 50% of all rainy days start with a cloudy morning;
- 40% of all days start with a cloudy morning (cloudy mornings are common);
- Only 3 of 30 days (10%) tend to be rainy in this period of the year.

What is the chance of rain during the day?

Exercise: Let us assume that we are in a institute for medical research and we are conducting a study on a particular genetic defect. We know that:

- 1% of people have the genetic defect;
- 90% of tests for the gene detect the defect (*true positives*);
- 9.6% of the tests are *false positives*, i.e. you get positive result for the test when you should have received a negative results.

If a person gets a positive test result, what are the odds they actually have the genetic defect? [To solve this problem you should use the discrete version of Eq. (1.11)].

Chapter 2

Probability Density Functions

If x is the result of an experiment, in general x depends on the particular realization of the experiment itself. Then, x is a **random variable**. There are two types of random variables: *discrete* variables, which can be enumerable with a finite list of numbers, and *continuous* variables. In these notes, we will focus on the second case, the continuous variables, since they are able to fit our aims. We note that, for utilitarian purposes, we can refer to a random variable also with the term *parameter* and the parameter's space is its domain.

In the analysis of continuous variables, it is useful to introduce the *probability density function*. If we denote the random variable with x and we consider a range in the random variable's domain $A \equiv [x_1, x_2] \in X$, then we can define the probability density function (pdf) $p(x)$ of the probability $P(x)$ as the function such that

$$P\big(x_1 \leq x \leq x_2\big) = \int_{x_1}^{x_2} p(x)\, dx\,. \tag{2.1}$$

If x is a continuous outcome of an experiment, its pdf (also called probability distribution) $p(x)$ is a function of x such that $p(x) \geq 0$ for $x \in X$, where X is the domain of x and

$$\int_X p(x)\, dx = 1\,. \tag{2.2}$$

Eq. (2.2) is the generalization to continuous variable of Eq. (1.1). This equation tells us an obvious fact, i.e. the probability of $x \in X$ is 100%, which means that we can be sure the x will fall in its domain. Sometimes, it could happen that the integral over the entire domain of your pdf differs from the unity, i.e.

$$\int_X p(x)\, dx = N\,.$$

This means that your pdf is not *normalized*, and then the probability described by your function lies outside the boundaries $[0, 1]$. Obviously, we can solve this issue defining a new pdf $p'(x)$ such that $p'(x) = p(x)/N$.

Notation: From this moment, we refer to a probability density function of the variable x, or pdf, with lower-case letter, such as $p(x)$. The integral of $p(x)$ over the range A, which represent the probability that $x \in A$, is labelled with the capital letter, such as $P(x \in A)$ and it is called *cumulative probability function* (cpf).

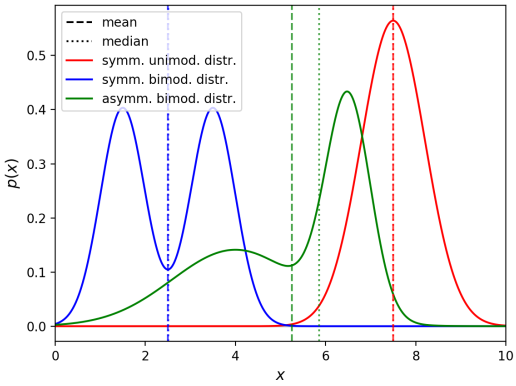

Figure 2.1: Example of unimodal symmetric pdf (red), bimodal symmetric pdf (blue) and bimodal asymmetric pdf (green). The vertical dashed lines represent the respective mean, while the dotted lines are the median values. The red pdf has only one mode and it coincides with the mean. The blue pdf has two equiprobable modes and then the mean value is located in their center, in correspondence of the local minimum. The green pdf has two different modes, then the mean value is going to be shifted toward the most probable peak.

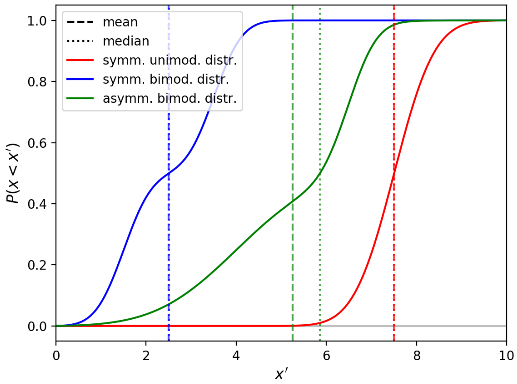

Figure 2.2: The figure shows the cumulative density probability functions $P(x < x')$ for the three pdf in Fig. 2.1. It is possible to see that for symmetric distributions, the mean and the median are coincident. For cumulative probabilities, it is simple to identify the median point, since it is the value $\mu_{1/2}$ such that $P(\mu_{1/2} < x') = 1/2$. In fact, applying Eq. (1.1), we get $P(\mu_{1/2} < x') = 1 - P(\mu_{1/2} > x') = 1/2$, and then also $P(\mu_{1/2} > x') = 1/2$. This is not valid for asymmetric pdf, where mean and median do not coincide.

2.0.1 Mean and Variance

If $x \in X$ is a random variable and $p(x)$ its pdf, we define the **mean** of the pdf the quantity

$$\mu = E[x] = \int_X x\, p(x)\, dx\,. \tag{2.3}$$

This term is also called *expectation value* of the variable x or we can generalize the definition to any function $f(x)$ of a random variable. We note also that the expectation value E is a linear operator, and so:

$$E[x + y] = E[x] + E[y], \quad E[\alpha x] = \alpha E[x]\,,$$

where x and y are two random variables and α is a constant real number. Often we rely on the mean value in order to decide if an outcome is reasonable or not, however this value coincides with the peak of the pdf only for *symmetric* and *unimodal* distributions. The term "symmetric" means that the pdf is symmetric with respect to the mean value. The term "unimodal" means that the pdf has only one peak, otherwise we refer to *bimodal* pdf or in general *multimodal* pdf. An example is given in Fig. 2.1.

Notation: Eq. (2.7) is the definition of the mean but it defines also the expectation value operator E. This linear operator can be generalized to any function f of the random variable x as

$$E[f(x)] = \int_X f(x)\, p(x)\, dx\,. \tag{2.4}$$

Let us note that for the case of equiprobable outcomes $p(x_i) = 1/N$, $\forall\, i = 1, \ldots, N$, the definition in Eq. (2.7) returns the usual arithmetic mean,

$$\mu = E[x] = \sum_i x_i \cdot p(x_i) = \frac{1}{N} \sum_i x_i. \tag{2.5}$$

Note that the condition $p(x_i) = 1/N$ is imposed by normalization,

$$\sum_i p(x_i) = 1\,,$$

which is the natural discrete form of Eq. (2.2).

However, in general the mean does not coincides with a relevant value of the distribution, so we define other quantity: the **median** $\mu_{\frac{1}{2}}$ is the value of the random variable such that:

$$\int_{-\infty}^{\mu_{\frac{1}{2}}} p(x)\,dx = \int_{\mu_{\frac{1}{2}}}^{+\infty} p(x)\,dx\,, \tag{2.6}$$

and the **mode** μ^* which is the most probable value of the random variable and it correspond to the peak of the distribution.

The definition in Eq. (2.7) can be generalized. In general, when we have a random variable, we refer to the momentum of n-th order of the variable to denote the integral

$$E[x^n] = \int_X x^n\, p(x)\,dx\,. \tag{2.7}$$

This quantities are very import to characterize a pdf. As we saw the momentum of first order is the mean of the distribution. Then, we observe that the momentum of zeroth order is the unity, or in general the normalization constant of the pdf. The momentum of second order is useful to quantify the spread of the distribution around the mean value: indeed the momentum of second order centered around the mean value is known as **variance** of a distribution,

$$\mathrm{Var}(x) = E\Big[\big(x - E[x]\big)^2\Big] = E[x^2] - \big(E[x]\big)^2\,, \tag{2.8}$$

and its square root $\sqrt{\mathrm{Var}(x)}$ is called standard deviation. Furthermore, the momentum of third order centered around the mean value is called **skewness** and it quantify with the symmetry of the distribution around the mean value.

Exercise: Prove the relation in Eq. (2.8).

2.0.2 Confidence Intervals

If we want to report the result of a measurement in a scientific report or article, it is not always convenient to use mean and standard deviation. Indeed this properties work very well for symmetric pdf, but in general the pdf of your measurement is not symmetric and then, using mean and standard deviation, you are going to loose the information regarding the symmetric (or asymmetry) of your measurement. For some other cases, these quantities cannot be computed since the integrals do not converge.

Then, if x is a random variable and $p(x)$ its pdf, we define the **confidence interval** of the $N\%$ isoprobability contour (or simply the $N\%$ confidence interval) a region $\mathrm{CI}_N \equiv [a,b]$ of the variable's domain such that

$$P(x \in \mathrm{CI}_N) = \int_{\mathrm{CI}_N} p(x)\,dx = N \times 10^{-2}\,. \tag{2.9}$$

In order to avoid misunderstandings, we note that the boundary of CI_N are usually evaluated through the computation of the integrals of the tails, i.e.

$$\begin{aligned} P(x < a) &= \int_{-\infty}^{a} p(x)\,dx = \frac{100-N}{2} \times 10^{-2}\,,\\ P(x > b) &= \int_{b}^{+\infty} p(x)\,dx = \frac{100-N}{2} \times 10^{-2}\,. \end{aligned} \tag{2.10}$$

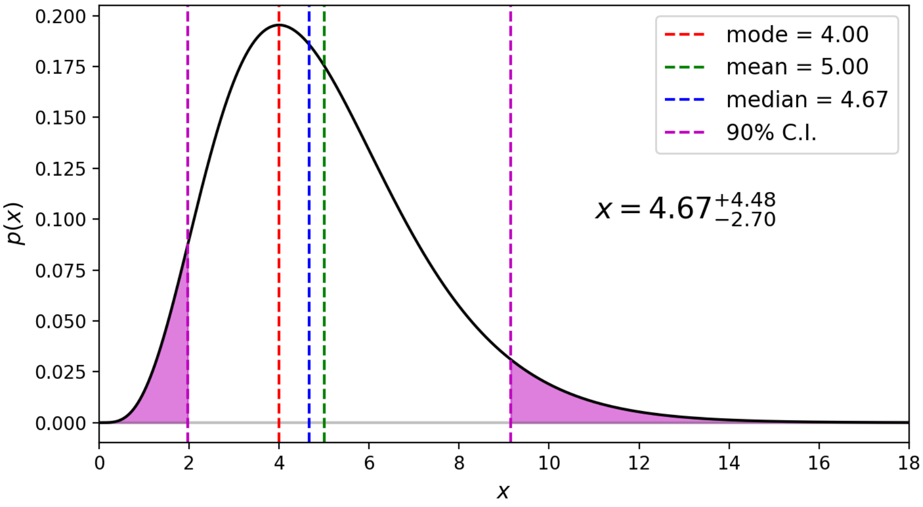

Figure 2.3: Example of unimodal asymmetric distribution. The red, green and blue vertical lines are respectively mode, mean and median values. It is possible to observe that in general this values do not coincides. Moreover the purple vertical lines represents the 90% confidence interval and the purple region is the area corresponding to the 10% of the entire integral. The text $x = 4.67^{+4.48}_{-2.70}$ is an example of how to report this particular measurement in a report using the median and the its differences with respected to the boundaries of the 90% credible interval.

Let us suppose we want to compute the 90% credible interval of a random variable, Eq. (2.10) is telling you the following: you can evaluate the boundaries a, b computing the integral of the tails and stopping when every single integral is 5%, since when the two tail contributions are 5% each (10% in total), the central portion is the remaining 90%. In other words, a is the 5th percentile and b is the 95th percentile of the distribution.

We can understand that if we report a central value (i.e. mean, median, . . .) and the confidence interval of a specified contour (i.e. 66%, 90% or 95%) we are able to discriminate if a pdf is symmetric or not. Usually it is common to use median value and 90% confidence interval to report the result of a measurement, since the median is a quantity more coherent with the definition Eq. (2.10) and mostly because it suffers less the effects of statistical fluctuations with respect to the other central quantities.

2.1 Error Propagation

Another useful tool in the statistic analysis is the *error propagation*. With this term we mean the behavior of a set of parameters under a given transformation; in other words, given the set of parameters $\mathbf{x}$ and a transformation $\mathbf{x} \to \mathbf{y} = f(\mathbf{x})$, how can we relate the probability $p(\mathbf{x})$ to the probability $p(\mathbf{y})$? In order to give an answer to this question, we recall the property of a probability for which the integration of $p(\mathbf{x})$ over the entire domain must be equal to the unity (Eq. (2.2)). This property must be true for the probability p in every analytical form, such that

$$\int p(\mathbf{x})\, d\mathbf{x} = 1 = \int p(\mathbf{y})\, d\mathbf{y}\,.$$

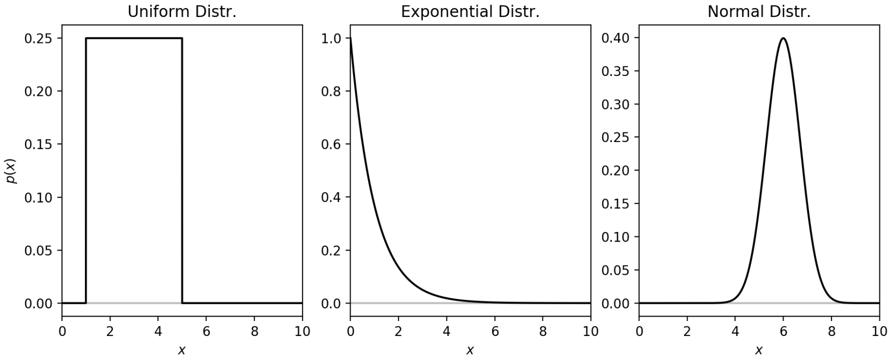

Figure 2.4: Example of uniform, exponential and normal distributions.

From this relation we find easily that the transformation function that allows us to write p through the new parameters is the determinant of the Jacobian of this transformation,

$$p(\mathbf{y}) = p(\mathbf{x}) \left| \frac{\partial \mathbf{x}}{\partial \mathbf{y}} \right| . \tag{2.11}$$

We can understand that using Eq. (2.11) we are able to compute how the uncertainties and correlations propagate from an initial set of variable $\mathbf{x}$ to a second set $\mathbf{y}$.

Exercise: Compute the Jacobian for the transofrmation from cartesian coordinates (x, y, z) to polar coordinates (r, θ, ϕ). If we assume a uniform distribution for the variables (x, y, z), which is the distribution for (r, θ, ϕ)?

2.2 Standard Distributions

In this section we introduce some useful distributions for continuous random variables. However, we note that we are omitting several very important distributions related with discrete variables, such as the *binomial* pdf (useful for true-false cases), the *Poissonian* (used when the random variable is an integer count) pdf and many others.

Notation: If x is a random variable and $p(x)$ its pdf, then we denote with the form $x \sim p(x)$ the sentence "*x is drawn from the distribution p*" or on the other way around "*p is the probability distribution for the variable x*".

2.2.1 Uniform Distribution

The uniform distribution is used when we are completely ignorant regarding a particular outcome of a variable. This pdf is specified only by two values a and b which correspond to the lower and upper bounds of the distribution. If x is a random continuous variable, then $\mathrm{U}(x|a,b)$ is defined as

$$\mathrm{U}(x|a,b) = \begin{cases} \frac{1}{b-a} & \text{if } x \in [a,b] \,, \\ 0 & \text{otherwise} \,. \end{cases} \tag{2.12}$$

An example of uniform pdf is shown in Fig. 2.4. We can see that a and b are the bounds of our pdf and there are no chance our variable x will be outside this boundaries. But inside this boundary every value has the same probability, i.e. we are completely ignorant on the possible result of x.

Exercise: Compute mean and variance for the variable $x \sim \mathrm{U}(x|a,b)$.

2.2.2 Exponential Distribution

The exponential distribution is characterized by a single parameter $\lambda \geq 0$ and it is defined as

$$\mathrm{e}(x|\lambda) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0\,, \\ 0 & \text{if } x < 0\,. \end{cases} \tag{2.13}$$

It is simply to compute the cumulative probability function (cpf) in the domain $x \leq x_0$, where x_0 is a possible value of x,

$$P(x \leq x_0) = 1 - e^{-\lambda x_0}\,, \tag{2.14}$$

and this form is very important is many fields of physics. Indeed Eq. (2.14) is the distribution function used to indicate the average time interval between two successive counts, i.e. decay time of radioactive nuclei (where $1/\lambda$ is called mean-life) or the length of the path of a particle in a homogeneous medium (where $1/\lambda$ is called mean-free-path).

The reason why this pdf is able to describe such events underlies in its most important property: using Eq. (1.1) and Eq. (2.14), we can write

$$\begin{aligned} P(x \geq x_1 + x_2) &= 1 - P(x < x_1 + x_2) \\ &= e^{-\lambda(x_1+x_2)} \\ &= e^{-\lambda x_1} \cdot e^{-\lambda x_2} \\ &= P(x \geq x_1) \cdot P(x \geq x_2)\,. \end{aligned}$$

A distribution with such a property is called *memory-less*, since after every iteration, the probability for x to be above a certain value x_i does not preserve the information on the previous outcome x_{i-1}. This condition can also be written as

$$P(x \geq x_1 + x_2 | x \geq x_1) = P(x \geq x_2)\,. \tag{2.15}$$

Exercise: Compute mean and variance for the variable $x \sim \mathrm{e}(x|\lambda)$.

Exercise: Let us assume to observe, starting from a moment t_0, a radioactive nucleus with mean-life $1/\lambda = 1\,\mathrm{s}$. We ask ourselves:

- What is the probability that the nucleus has not decayed after $1\,\mathrm{s}$?
- What is the probability that the nucleus has not decayed after $10\,\mathrm{s}$?
- What is the probability that the nucleus has not decayed between $10\,\mathrm{s}$ and $11\,\mathrm{s}$?

2.2.3 Normal Distribution

The normal distribution, also called the *Gaussian* distribution, is the most important and used pdf. The normal distribution is characterize by two values μ, σ and it is defined as

$$\mathrm{N}(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}\,. \tag{2.16}$$

n	$P(x \in I_n)$
0.318639	25%
0.674490	50%
1	68.2689492%
1.281552	80%
1.644854	90%
2	95.4499736%
3	99.7300204%
4	99.993666%
5	99.9999426697%
6	99.9999998027%

Table 2.1: Table reporting the confidence interval levels for different values of n, which define the region $I_n = [\mu - n\sigma, \mu + n\sigma]$. This notation is used very often within the scientific community, however remember that this notation underlies the fact that you are assuming that the pdf of your measurement is Gaussian.

We can observe that $\mathrm{N}(x|\mu, \sigma)$ is defined on the entire domain of the real numbers, it has only one mode and it is symmetric with respect to the mean.

However, even if the normal pdf fits perfectly the description of a canonical distribution, the integral

$$\int_a^b \mathrm{N}(x|\mu, \sigma)\, dx \,, \tag{2.17}$$

has no analytical expression. Obviously, this does not mean that it cannot be evaluated (numerically, in example) with the desired accuracy level. Usually, in order to avoid this problem, it is common to define the *error function* $\mathrm{erf}(x)$ such as

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\, dt \,. \tag{2.18}$$

With this definition, if $x \sim \mathrm{N}(x|\mu = 0, \sigma = 1)$, we can compute the probability that x will fall in an interval $[a, b]$ as

$$\frac{\sqrt{2}}{4}\Big[\mathrm{erf}(b) - \mathrm{erf}(a)\Big] = P(x \in [a, b]) \,.$$

This can be very useful to evaluate the confidence level of a pdf in a given region of the parameter's space. The generalization for this evaluation can be found applying the transformation $t \to (z - \mu)/(\sqrt{2}\sigma)$ in Eq. (2.18) and rescaling the result for σ.

At this point, it is interesting to note that if we compute (numerically) the integral

$$\int_{\mu-\sigma}^{\mu+\sigma} \mathrm{N}(x|\mu, \sigma) \approx 0.68 \,.$$

This means that the credible interval $\mu \pm \sigma$ of a normal distribution correspond to the 68% isoprobability contour. Then we can extend the integral to the region $\mu \pm 2\sigma$ and we get that the results roughly correspond to the 95% c.i., while $\mu \pm 3\sigma$ is the 99.7% credible region. Then we can generalize this property and create a table which relates a value n, such that $I_n = [\mu - n\sigma, \mu + n\sigma]$, to the probability that the observed event will fall in I_n, as it is shown in Tab. 2.1.

Another very important property of the normal distribution is that it is a *limit distribution*. Let us explain this concept with an example: the binomial distribution is used to characterized discrete variables with two possible outcomes, i.e. true-false, head-tail, 0-1 and so on (see Sec. 2.3). The binomial distribution is defined for integer values k as

$$\mathrm{Bin}(k|n, p) = \binom{n}{k} p^k (1 - p)^{n-k} \,, \tag{2.19}$$

where p is the probability of one of the two possible outcomes (then $1-p$ is the probability of the other one), k represents the number of positive results (i.e. the one corresponding to the probability p) and n is the total number of draws. This distribution as mean $E[k] = np$ and variance $\mathrm{Var}(x) = np(1-p)$. If we take the limit $n \to +\infty$ of a binomial distribution, then we can prove that the normal distribution is the result of this limit. The same is valid for the Poissonian distribution, which is another pdf for discrete variables, defined as

$$\mathrm{P}(k|\mu) = \frac{\mu^k}{k!} e^{-\mu} . \tag{2.20}$$

This pdf reach the form of a normal distribution in the limit $\mu \to +\infty$. We note that, if $k \sim \mathrm{P}(k|\mu)$, then $E[k] = \mu$ and $\mathrm{Var}(k) = \mu$. However, the fact that the normal distribution is the limit expression for other discrete pdf underlies the assumption that we are extending the domain of the discrete variable k to the entire set of the real number.

In general, the normal distribution is an excellent approximation for the pdf of a measurement since this pdf is able to describe statistical fluctuations due to summation of several independent contributions. In physics, this happens very often when there are no systematic errors.

Exercise: Compute mean and variance for the variable $x \sim \mathrm{N}(x|\mu, \sigma)$.

Exercise: Suppose we have two independent events x, y such that $x \sim \mathrm{N}(x|\mu_1, \sigma_1)$ and $y \sim \mathrm{N}(x|\mu_2, \sigma_2)$. How can we estimate the pdf of the quantity $x + y$?

2.2.4 Other Distributions

Here we comment other important distributions. The first pdf we mention is the *Lorentzian* distribution, defined as

$$\mathrm{L}(x|\mu, \gamma) = \frac{\gamma}{\pi} \frac{1}{(x-\mu)^2 + \gamma^2} . \tag{2.21}$$

This function is often used in atomic and particle physics and it has some interesting properties. First of all we note that Eq. (2.21) is normalized. Then, the median exist and it coincides with μ and the distribution is symmetric with respect to this value. However the mean and the higher momenta are not defined since the integral

$$\frac{\gamma}{\pi} \int \frac{x}{x^2 + \gamma^2}\, dx = \frac{1}{2} \log(\gamma^2 + x^2) ,$$

leads to an indeterminate form $\infty - \infty$ when it is evaluated at $\pm\infty$. Then, the mean is not defined for this distribution and the same for higher momenta. This is also one of the reason why usually the median is preferred to the mean.

Another important distribution is the χ^2 distribution. This pdf is defined only for real positive numbers $x \geq 0$ as

$$\chi^2(x|n) = \frac{x^{\frac{n-2}{2}}\, e^{-\frac{x}{2}}}{2^{\frac{n}{2}}\, \Gamma\left(\frac{n}{2}\right)} , \tag{2.22}$$

where n is labelled as degree-of-freedom and Γ is the Euler's function. This distribution characterize the sum of the residuals of Gaussian variable. This means the following: suppose we perform n measurements of the same quantity and we collect a set of values $\{z_1, z_2, z_3, \dots, z_n\}$, all these values are different realizations of the same quantity and this means that they are drawn from the same pdf, $z_i \sim \mathrm{N}(x|\mu, \sigma)$ $\forall i = 1, \dots, n$. Then the sum of the residuals x, defined as

$$x = \sum_{i=1}^{n} \left(\frac{z_i - \mu}{\sigma} \right)^2 ,$$

is distributed as a χ^2 distribution, i.e. $x \sim \chi^2(x|n)$, with degree-of-freedom n. This pdf has existing mean and variance, and they are respectively n and $2n$.

We want to mention one last useful pdf, but let us introduce it though an example. Suppose that x is a random variable such that $x \sim \mathrm{U}(x|a,b)$. Then is we perform the transformation $x \to y = e^x$, how can we write $p(y)$? Using Eq. (2.11) and observing that $x = \log(y)$, we can write

$$p(y) = p(x)\,\frac{dx}{dy}\,, \quad \text{where } \frac{dx}{dy} = \frac{1}{y}\,,$$

$$p(y) \propto \frac{1}{y}\,.$$

Then we can understand that $p(y) \propto 1/y$ is the distribution which reproduce an uniform pdf under the transformation $x \to y = e^x$. This fact can be expressed also on the other way around as follows: if we assume that $\log(x) \sim \mathrm{U}(x|a,b)$, then $x \sim 1/x$.

2.3 Example: Toss a Coin

In this section we want to examine the behaves of the posterior during an set of observations and let us begin with a simple coin-tossing experiment. A very important check is knowing how to recognize when a coin is fair. By fair, we mean that we would be prepared to lay an even 50 : 50 bet on the outcome of a flip being a head or a tail. In ascribing the property of fairness to the coin we are, of course, assuming that the coin-tosser was not skilled enough to be able to control the initial conditions of the flip (such as the angular and linear velocities).

A sensible way of formulating this problem is to consider a large number of contiguous propositions about the range in which the bias-weighting of the coin might lie. If we denote the bias-weighting (or in general the parameter) by θ, and we consider a range in the parameter's space $[\theta_1, \theta_2]$, then we can define the probability density function (pdf) $p(\theta)$ of the probability $P(\theta)$ as the function such that

$$P\big(\theta \in [\theta_1, \theta_2]\big) = \int_{\theta_1}^{\theta_2} p(\theta)\, d\theta\,. \tag{2.23}$$

Furthermore, $\theta = 0$ and $\theta = 1$ can represent a coin which produces a tail or a head on every flip, respectively. There is a continuum of possibilities for the value of θ between these limits, with $\theta = 1/2$ indicating a fair coin. Our state of knowledge about the fairness, or the degree of unfairness, of the coin is then completely summarized by specifying how much we believe these various propositions to be true. If we assign a high probability to one (or a closely-grouped few) of these propositions, compared to the others, then this would indicate that we were confident in our estimate of the bias-weighting. If there was no such strong distinction, then it would reflect a high level of ignorance about the nature of the coin.

In the light of the data, and the above discussion, our inference about the fairness of this coin is summarized by the conditional pdf: $P(\theta|\mathbf{d}, \mathscr{H})$, that is the posterior. To estimate this posterior pdf, we need to use Bayes' theorem Eq. (1.9); it relates the pdf of interest to two others, which are easier to assign. The prior distribution represent what we know about the coin given only the initial informations; since we know nothing about the coin, we should keep a very open mind about its nature and a simple probability assignment which reflects this is a uniform prior,

$$p(\theta|\mathscr{H}) = \begin{cases} 1 & \text{if } 0 \le \theta \le 1\,, \\ 0 & \text{otherwise}\,. \end{cases}$$

This prior state of knowledge, or ignorance, is modified by the data through the likelihood function: $P(\mathbf{d}|\boldsymbol{\theta}, \mathscr{H})$. It is a measure of the chance that we would have obtained the data that we actually observed, if the value of the bias-weighting was given (as known). If, in the conditioning information $\mathcal{H}$, we assume that the flips of the coin were independent events, so that the outcome of one did not influence that of another, then the probability of obtaining the data "*k heads in N tosses*" is given by the binomial distribution:

$$p(d|\theta, \mathscr{H}) = \binom{N}{k} \theta^k \left(1-\theta\right)^{N-k} .$$

This solution is reasonable because θ is the chance of obtaining a head on any flip, and there were k of them, and $1-\theta$ is the corresponding probability for a tail, of which there were $N-k$. According to Eq. (1.9), we can compute the posterior which represents our state of knowledge about the nature of the coin in the light of the data. To get a feel for this result, Fig. 2.5 shows how this pdf evolves as we obtain more and more data pertaining to the coin. This is done with the aid of data generated in a computer simulation. The panel in the top left-hand corner shows the posterior pdf for θ given no data (it coincides with the prior pdf); it indicates that we have no more reason to believe that the coin is fair than we have to think that it is double-headed, double-tailed, or of any other intermediate bias-weighting.

As far as we go with the extractions, the information we get modify our posterior distribution: we can see that at $N = 1$ we have a head, then the posterior is biased from the single result we obtained. For $N = 2$, we get a tail and a head, which is a pretty fair result and the posterior is centered around $\theta = 1/2$. With increasing N, the posterior get peaked around the expectation value of θ, which coincides with $1/2$ if the coin is fair.

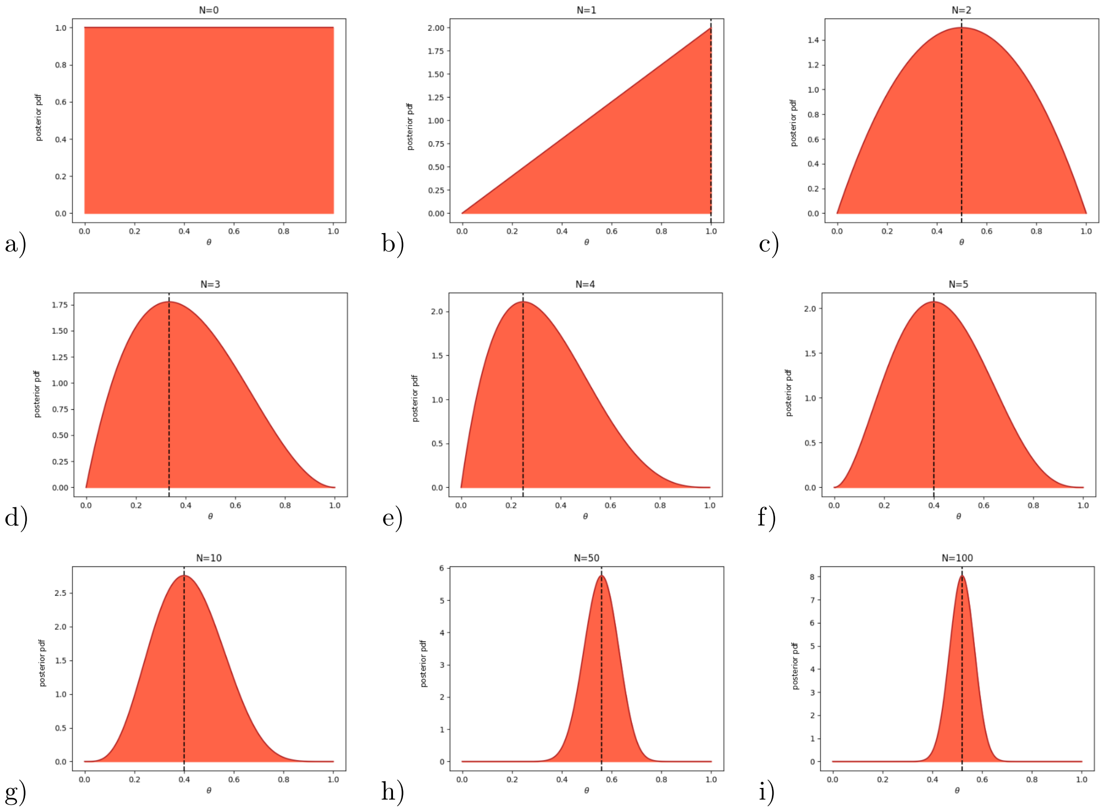

Figure 2.5: We show the evolution of the posterior probability density as new extractions are made and the vertical dashed lines represent the peak k/N. Increasing the number of tossed coin, we reach the value $\theta = 1/2$ which means that our coin is fair. Note that the posterior distributions with $N < 10$ are strongly affected by the statistical fluctuations.

Chapter 3

Applications

In the following sections, we will present applications and elaboration about the concepts learnt in the previous paragraphs. In the first section we discuss the quadratic approximation of a generic pdf, which correspond to the Gaussian case. In the second section we explain the principles of Bayesian model selection, based on the computation of the evidence. In the last section, we introduce the hierarchical models and the concept of hyper-parameters.

3.1 Quadratic Approximation

The number of inference problems that can be tackled by Bayesian inference methods is enormous. The aim of inference is to find the most probable explanation for some data. While this most probable hypothesis may be of interest, and some inference methods do locate it, this hypothesis is just the peak of a probability distribution, and it is the whole distribution that is of interest.

In Sec. 2.3, we have seen how the posterior pdf encodes our inference about the value of a parameter, given the data and the relevant background information. Often, however, we wish to summarize this with just two numbers: the best estimate and a measure of its reliability. Since the probability (density) associated with any particular value of the parameter is a measure of how much we believe that it lies in the neighborhood of that point, our best estimate is given by the maximum of the posterior pdf. If we denote the quantity of interest by θ and its posterior pdf $p(\theta|\mathbf{d},\mathscr{H})$, then the best estimate of its value θ^* is given by the **maximum posterior** condition

$$\left.\frac{d}{d\theta}\,p(\theta|\mathbf{d},\mathscr{H})\right|_{\theta=\theta^*} = 0\,. \tag{3.1}$$

Then θ^* is the mode of the posterior distribution $p(\theta|\mathbf{d},\mathscr{H})$. In the case of uniform prior distribution using Eq. (1.7), we found the usual *maximum likelihood principle.* However, we should also check the sign of the second derivate to ensure that θ^* represents a maximum rather than a minimum. Moreover, the second derivative gives us information about the width of the distribution around the peak. In order to consider the behavior of the function in the neighborhood of a point, we perform a Taylor expansion, or to be more precise we expand the logarithmic of the posterior pdf about the point $\theta = \theta^*$,

$$\log p(\theta|\mathbf{d},\mathscr{H}) = \log p(\theta^*|\mathbf{d},\mathscr{H}) + \frac{1}{2}\left[\frac{d^2}{d\theta^2}\,\log p(\theta^*|\mathbf{d},\mathscr{H})\right]_{\theta=\theta^*}\left(\theta-\theta^*\right)^2 + \dots \tag{3.2}$$

where the first derivate vanish since we are around the maximum and Eq. (3.1) holds. The first

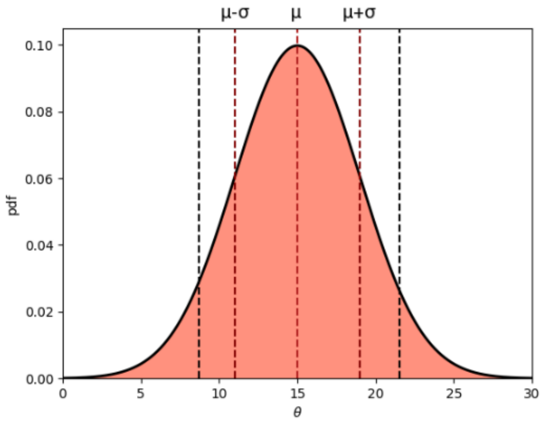

Figure 3.1: Example of normal distribution with $\mu = 15$ and $\sigma = 4$. The red dashed lines represents the mean μ and the values $\mu \pm \sigma$ (which corresponds roughly to the 68% confidence interval), while the black lines represent the 90% confidence interval. The quadratic approximation is very useful since it is able to describe the result of a measurement affected by summation of independent and uncorrelated sources of noise.

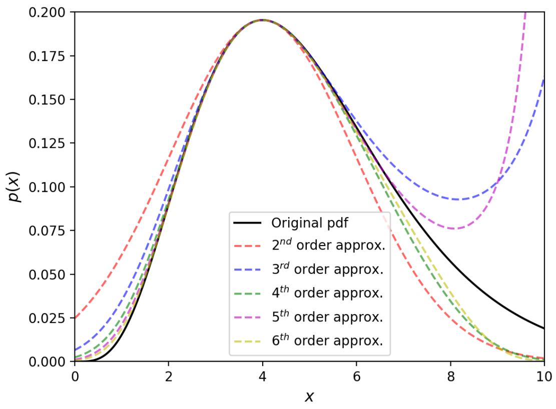

Figure 3.2: Example of approximations with a Taylor series, Eq. (3.2), of an asymmetric pdf. It is possible to see that the second order approximation (red) is able to reproduce the original pdf around the peak $x^* = 4$, but it cannot reproduce the asymmetries. Then we can see that higher order approximations are able to improve the description of the original pdf in the region around x^*.

term is a constant and tells us nothing about the shape of the distribution. The quartic term is, therefore, the dominant factor determining the width of the posterior.

Consider now a distribution that can be expressed only through the quartic term in the previous equation and all the other terms in the expansion vanishes ,

$$p(\theta|\mathbf{d}, \mathscr{H}) = a \cdot \exp\left[\frac{1}{2}\, b\big(\theta - \theta^*\big)^2\right],$$

where a is the normalization constant,

$$a = p(\theta^*|\mathbf{d}, \mathscr{H})\,,$$

and b is defined as

$$b = \left[\frac{d^2}{d\theta^2} \log p(\theta|\mathbf{d}, \mathscr{H})\right]_{\theta=\theta^*}.$$

This kind of distribution coincides with the normal distribution and it is usually written as,

$$\mathrm{N}(\theta|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(\theta-\mu)^2}{2\sigma^2}\right], \tag{3.3}$$

and plotted in Fig. 3.1. The value $\mu \equiv \theta^*$ represent the mode of the pdf and the distribution is symmetric with respect to it, under this assumptions. Its width is characterized by the parameter σ, also called the standard deviation, which in this case corresponds to

$$\sigma = \sqrt{-b}\,. \tag{3.4}$$

So, our inference about the quantity of interest θ is conveyed concisely by the statement $\theta = \mu \pm \sigma$. However, this notation is valid only for quadratic approximation.

Increasing the order of approximation of the pdf, we are able to describe the asymmetries of the original pdf, as it is shown in Fig. 3.2. The higher-order terms capture the morphology of the original pdf, in an interval centered around the mode θ^*. Increasing the order of approximation, the accuracy level increase as well in an increasingly larger range. However, it is possible to see that odd orders of expansion (third, fifth and so on) could generate divergencies in region far from the peak of the pdf, as it is for $x \to +\infty$ in Fig. 3.2. For this reason, we suggest to use always even orders of expansion in the case of high-order approximations, since their behavior for $x \to \pm\infty$ can be always take under control. Furthermore, we observe that, since the expansion is performed around the mode θ^*, this value remains always a maximum point (at least local) for every term in the expansion.

3.1.1 Multidimensional Distributions

We consider an estimation problem involving more than one parameter. Although the posterior pdf is of a higher dimensionality, being a function of several variables, it still encodes our inference about the values of the parameters, given the data and the relevant background information. As before, we often wish to summarize it with just a few numbers: the best estimates and a measure of their reliabilities. Since the probability density associated with any particular set of values for the parameters is a measure of how much we believe that they lie in the neighborhood of those values, our optimal estimate is given by the maximum of the posterior pdf.

If we denote with $\{\theta_i\}$ the set of parameters, the best estimate of their values is given by the generalization of Eq. (3.1)

$$\left.\frac{\partial p}{\partial \theta_i}\right|_{\theta_i^*} = 0\,, \tag{3.5}$$

where $p(\theta_1, \theta_2) \equiv p(\theta_1, \theta_2|\mathbf{d}, \mathscr{H})$ and θ_i^* are the values that maximize the posterior pdf. We also need a further test, analogous to Eq. (3.1), to ensure that we are dealing with a maximum and not a minimum or a saddle-point; we must evaluate the second derivates. As we done above, we expand using a Taylor series and we suppose that $\{\theta_i\} \equiv \{\theta_1, \theta_2\}$. Then, we can write

$$\begin{aligned}\log p(\theta_1, \theta_2) = \log p(\theta_1^*, \theta_2^*) + \frac{1}{2}\Bigg[& \left.\frac{\partial^2 \log p}{\partial \theta_1^2}\right|_{\theta_1^*, \theta_2^*} \big(\theta_1 - \theta_1^*\big)^2 \\ & + \left.\frac{\partial^2 \log p}{\partial \theta_2^2}\right|_{\theta_1^*, \theta_2^*} \big(\theta_2 - \theta_2^*\big)^2 \\ & + 2\left.\frac{\partial^2 \log p}{\partial \theta_1 \partial \theta_2}\right|_{\theta_1^*, \theta_2^*} \big(\theta_1 - \theta_1^*\big)\big(\theta_2 - \theta_2^*\big)\Bigg] \\ & + \ldots\end{aligned}$$

where we have assumed that $\partial^2 p/\partial\theta_1\partial\theta_2 = \partial^2 p/\partial\theta_2\partial\theta_1$. We can use a matrix notation, and limiting our analysis to normal distributions, i.e. the quadratic approximation, we can write,

$$\mathrm{N}_2\big(\theta_i|\theta_i^*, C_{ij}\big) = \frac{1}{\sqrt{(2\pi)^2|C|}} \exp\left[-\frac{1}{2}\,(\theta_i - \theta_i^*)\, C_{ij}^{-1}\, (\theta_j - \theta_j^*)\right]. \tag{3.6}$$

The matrix C_{ij} is called the matrix containing the second derivates of the pdf, $|C|$ is its determinant and in the quadratic approximation $\theta_i^* \equiv \mu_i$. In the θ_1, θ_2 plane the quantity between brackets is an ellipse centered in (θ_1^*, θ_2^*) and the orientation and the eccentricity are determined by C_{ij}. The directions of the principal axes formally correspond to the eigenvectors of the second-derivative matrix. It is also easy to see, in the quadratic approximation, that the marginalized pdf of θ_1 and θ_2 maintain a Gaussian profiles.

We are able to introduce the variance of the multidimensional pdf, which gives a measure of its spread. It is formally defined to be the expectation value of the square of the deviations from the mean; this is given by Eq. (2.8) for every parameter. For the one-dimensional normal distribution of Eq. (3.3), this integral yields the result

$$\mathrm{Var}(\theta) = \sigma^2 \,,$$

where σ is the standard deviation of Eq. (3.4). This definition of the error-bar can be extended to pdfs of more than one variable. Explicitly, for the two-dimensional case we have been considering the marginalized posterior,

$$\begin{aligned}\sigma_i^2 = E\Big[\big(\theta_i - E[\theta_i]\big)^2\Big] &= \int \big(\theta_i - E[\theta_i]\big)^2\, p(\theta_1, \theta_2)\, d\theta_1 d\theta_2 \\ &= \int \big(\theta_i - E[\theta_i]\big)^2\, p(\theta_i)\, d\theta_i \,.\end{aligned}$$

The idea of variance can be broadened to consider the simultaneous deviations of both θ_1 and θ_2; this **covariance**, which we will denote as $\mathrm{Cov}(\theta_1, \theta_2)$, is given by

$$\mathrm{Cov}(\theta_1, \theta_2) = \sigma_{12}^2 = E\Big[\big(\theta_1 - E[\theta_1]\big)\big(\theta_2 - E[\theta_2]\big)\Big] \,, \tag{3.7}$$

and is a measure of the correlation between the inferred parameters. If an over-estimate of one usually leads to a larger than average value for the other, then the difference $\theta_2 - E[\theta_2]$ will tend to be positive when $\theta_1 - E[\theta_1]$ is positive; if the same is true for under- estimates, so that $\theta_2 - E[\theta_2]$ is usually negative when $\theta_1 - E[\theta_1]$ is as well, the expectation value of the product of the deviations will be positive: the covariance will then be greater than zero. If there is an anti-correlation, so that an over-estimate of one is accompanied by an under-estimation of the other, then the covariance will be negative. When our estimate of one parameter has little, or no, influence on the inferred value of the other, then the magnitude of the covariance will be negligible in comparison to the variance terms; in other words, $\sigma_{12}^2 \ll \sqrt{\sigma_1^2 \sigma_2^2}$.

Furthermore, using the definition Eq. (3.7), we are able to construct the relation between σ_{ij}^2 (variances and covariances) and the matrix C_{ij} in Eq. (3.6): indeed, we obtain $\sigma_{ij}^2 = C_{ij}^{-1}$. For this reason σ_{ij}^2 is often called **covariance matrix**. So, according with Eq. (3.3), we will indicate a n-dimension normal distribution using the matrix formalism as

$$\mathrm{N}_n(\boldsymbol{\theta}|\boldsymbol{\mu}, \boldsymbol{\sigma}) = \frac{1}{\sqrt{(2\pi)^n |\boldsymbol{\sigma}|}}\, e^{-\frac{1}{2}\left[(\theta_i - \mu_i)\, \sigma_{ij}^{-1}\, (\theta_j - \mu_j)\right]} \,, \tag{3.8}$$

where $\boldsymbol{\mu}$ is the vector of means and $\boldsymbol{\sigma} = \sigma_{ij}^2$ is the covariance matrix. We are also able to extend the previous methods to n-dimensional normal distribution. The maximum of the multivariate Gaussian is defined by the vector $\boldsymbol{\theta}^* = \{\theta_1^*, \theta_2^*, \theta_3^*, \dots\}$ and calling $\nabla_i = \partial/\partial\theta_i$; the condition for finding its components, in Eq. (3.5), can be written compactly as

$$\boldsymbol{\nabla} \log p(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} = 0 \,,$$

where the gradient is evaluated in $\boldsymbol{\theta}^*$. By comparison with the standard one-dimensional Gaussian of Eq. (3.3), we found that that the spread of the posterior should be related to the inverse of the second- derivative matrix. Indeed, the covariance matrix is give by

$$\mathrm{Cov}(\boldsymbol{\theta}) = \sigma_{ij}^2 = -\Big(\nabla_i \nabla_j \log p\Big)^{-1} \,. \tag{3.9}$$

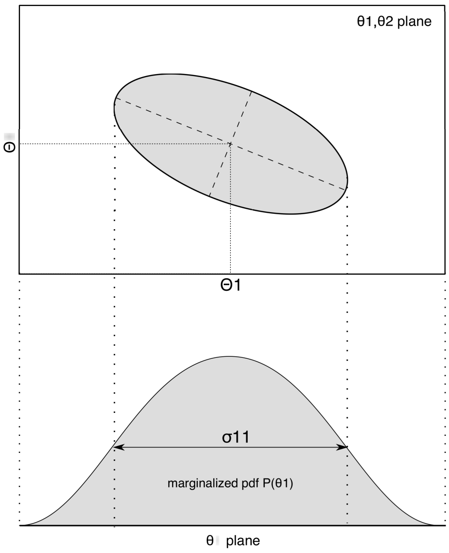

Figure 3.3: Example of 2-dimensional normal distribution and its projection on the θ_1 plane (marginalized over θ_2). Its contour in the (θ_1, θ_2) plane is an ellipse centered in (θ_1^*, θ_2^*) and with semiaxes defined by the correlation matrix $\sigma_{ij} = -(\nabla_i \nabla_j \log p)^{-1/2}$. The pdf marginalized over the θ_2 axis, $p(\theta_1) = \int p(\theta_1, \theta_2)\, d\theta_2$, is shown with the corresponding standard deviation $\sigma_{11} = -(\nabla_1^2 \log p)^{-1/2}$. It is possible to observe that, when. we perform the marginalization, we have lost all the information about θ_2 and how it is correlated with θ_1.

This equation is the generalization of Eq. (2.8) and Eq. (3.7). The square root of the diagonal elements $i = j$ corresponds to the marginal error-bars for the associated parameters; the off-diagonal components $i \neq j$ tell us about the correlations between the inferred values of θ_i and θ_j.

3.2 Model Selection

If we want to infer about on a set of data, we would have to choose a functional form based on the relevant background information available, including theoretical considerations, the results of calibration measurements or merely an approximation to simplify the algebra; in all cases, the underlying assumptions need to be stated clearly (in the conditioning on $\mathscr{H}$). Suppose we have a set of measures of an observable and we must decide if this measure is distributed like a Gaussian or like a Lorentzian. How can we decide which is better?

The type of question posed above is often called *model selection*, or model comparison. Naïvely, we might think that a choice between proposed alternatives can be made on the basis of how well they fit the data. A little reflection soon reveals a potential difficulty in that more complicated models, defined by many parameters, will always be able to give better agreement with the experimental measurements. Let us begin with an elementary formulation due to Jeffreys (1939); we call it the story of Mr A and Mr B [4].

Suppose that we have the vector of data $\mathbf{d}$, Mr A has a theory and thinks that the noisy measurements of y against x are described by the law $y = 0$; Mr B has also a theory and he believes that $y = \lambda$, but is not sure about the value of the constant λ; there could also be a Mr C who is willing to allow the possibility of a non-zero slope, so that $y = \lambda_1 x + \lambda_0$ and therefore has two adjustable coefficients; and so on. It is clear that we need to evaluate the posterior probabilities for A and B being correct to ascertain the relative merit of the two theories. Intuitively, a good value that can describe the validity of a model over another is the ratio of the posteriors,

$$\mathcal{B} = \frac{p(\mathrm{A}|\mathbf{d}, \mathscr{H}_\mathrm{A})}{p(\mathrm{B}|\mathbf{d}, \mathscr{H}_\mathrm{B})} \, . \tag{3.10}$$

Let us start by applying Bayes' theorem to both the numerator and the denominator, this gives

$$\frac{p(\mathrm{A}|\mathbf{d},\mathscr{H}_{\mathrm{A}})}{p(\mathrm{B}|\mathbf{d},\mathscr{H}_{\mathrm{B}})} = \frac{p(\mathbf{d}|\mathrm{A},\mathscr{H}_{\mathrm{A}})}{p(\mathbf{d}|\mathrm{B},\mathscr{H}_{\mathrm{B}})} \cdot \frac{p(\mathrm{A}|\mathscr{H}_{\mathrm{A}})}{p(\mathrm{B}|\mathscr{H}_{\mathrm{B}})} \,,$$

because the term $p(\mathbf{d}|\mathscr{H}_{\mathrm{A,B}})$ cancels out, top and bottom, since the data do not depend on the hypothesis. The second term on the right-hand side reflects our relative prior preference for the alternative theories; to be fair, we can take it to be unity. To assign the probabilities involving the experimental measurements, we need to be able to compare the data with the predictions of A and B: the larger the mismatch, the lower the corresponding probability. This calculation is straightforward for Mr A, but not for Mr B; the latter cannot make predictions without a value for λ. To circumvent this difficulty, we can use the sum and product rule to relate the probability we require to other pdfs which might be easier to assign. In particular, we can express

$$p\big(\mathbf{d}|\mathrm{B},\mathscr{H}_{\mathrm{B}}\big) = \int p\big(\mathbf{d},\lambda|\mathrm{B},\mathscr{H}_{\mathrm{B}}\big)\, d\lambda = \int p\big(\mathbf{d}|\lambda,\mathrm{B},\mathscr{H}_{\mathrm{B}}\big)\, p\big(\lambda|\mathrm{B},\mathscr{H}_{\mathrm{B}}\big)\, d\lambda\,. \tag{3.11}$$

The first term in the integrand is the ordinary likelihood function and the second term is the prior pdf, and using Eq. (1.7) we can replace with the posterior pdf product the evidence. Since the normalization constraints on the posterior pdf are valid, we ca rewrite the ratio in Eq. (3.10), also called the **Bayes' factor**, using the evidences as

$$\mathcal{B}^{\mathrm{A}}_{\mathrm{B}} = \frac{p(\mathbf{d}|\mathscr{H}_{A})}{p(\mathbf{d}|\mathscr{H}_{B})} = \frac{\int p(\mathbf{d}|\lambda_{A},\mathscr{H}_{A})\, p(\lambda_{A}|\mathscr{H}_{A})\, d\lambda_{\mathrm{A}}}{\int p(\mathbf{d}|\lambda_{B},\mathscr{H}_{B})\, p(\lambda_{B}|\mathscr{H}_{B})\, d\lambda_{\mathrm{B}}}\,. \tag{3.12}$$

Then, if $\mathcal{B}$ is very much greater than one, then we will prefer A's theory; if it is very much less than one, then we prefer that of B; and if it is of order unity, then the current data are insufficient to make an informed judgement. As usual, probability theory warns us immediately that the answer to our question depends partly on what we thought about the two theories before the analysis of the data.

To proceed further analytically, we have to make some considerations and approximations: for first, the assumption of Mr A, $\lambda_{\mathrm{A}} = 0$, is reflected in the choice of the prior $p(\lambda_{\mathrm{A}}|\mathscr{H}_{A}) = \delta(\lambda_{\mathrm{A}})$, where we use the Dirac δ function. Then, we assume that, a priori, Mr B is only prepared to say that λ_{B} must lie between the limits λ_{min} and λ_{max}, assigning an uniform prior within this range:

$$p(\lambda_{\mathrm{B}}|\mathscr{H}_{\mathrm{B}}) = \begin{cases} 1/\big(\lambda_{\mathrm{max}} - \lambda_{\mathrm{min}}\big) & \text{for } \lambda_{\mathrm{max}} \geq \lambda_{\mathrm{B}} \geq \lambda_{\mathrm{min}}\,, \\ 0 & \text{otherwise}\,. \end{cases}$$

Let us also take it that there is a value $\lambda^* \neq 0$ which yields the closest agreement with the measurements; the corresponding probability $p(\mathbf{d}|\lambda_{\mathrm{B}} = \lambda^*, \mathscr{H}_{\mathrm{B}})$ will be the maximum of B's likelihood function. As long as this adjustable parameter lies in the neighborhood of the optimal value, $\lambda^* \pm \sigma_\lambda$, we would expect a reasonable fit to the data; this can be represented by the Gaussian pdf,

$$p(\mathbf{d}|\lambda) = \frac{1}{\sqrt{2\pi}\sigma_\lambda} \exp\left[- \frac{\big(\lambda - \lambda^*\big)}{2\sigma_\lambda^2} \right].$$

This is the likelihood function used to estimate the agreement of a model with the data. Then, we get

$$p(\mathbf{d}|\mathscr{H}_{\mathrm{A}}) = \int \delta(\lambda_{\mathrm{A}})\, p(\mathbf{d}|\lambda_{\mathrm{A}},\mathscr{H}_{\mathrm{A}})\, d\lambda_{\mathrm{A}}\,,$$

and

$$p(\mathbf{d}|\mathscr{H}_{\mathrm{B}}) = \frac{1}{\lambda_{\mathrm{max}} - \lambda_{\mathrm{min}}} \int_{\lambda_{\mathrm{min}}}^{\lambda_{\mathrm{max}}} p(\mathbf{d}|\lambda_{\mathrm{B}},\mathscr{H}_{\mathrm{B}})\, d\lambda_{\mathrm{B}}\,.$$

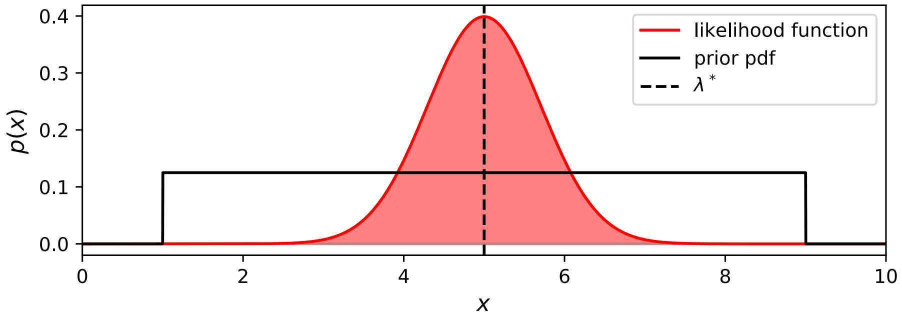

Figure 3.4: Graphical representation of Mr B's theory: the black line is the prior pdf, which reflects our information before the measurement. The red line is the likelihood function for the parameter λ_B once he get the measurement λ^*.

Assuming that the sharp cut-offs at λ_max and λ_min do not cause a significant truncation of the Gaussian pdf, its integral will be equal to 1, and then the Bayes factor becomes

$$\mathcal{B}^\text{A}_\text{B} = p(\mathbf{d}|\lambda = 0) \cdot \left(\lambda_\text{max} - \lambda_\text{min}\right) . \tag{3.13}$$

We note that often the likelihood function is defined without normalization constant. Then the quantities computed during these previous steps could differ, but if the method is applied consistently between the two analyzed, the final result will be the same.

The first term $p(\mathbf{d}|\lambda = 0) = \text{N}(\lambda = 0|\lambda^*, \sigma_\lambda)$ measures how well the prediction of Mr. A's model agrees with the data. This term is divided by 1, which is the integral of Mr. B's likelihood. Then, under a more general point of view, we can see this term as the evaluation of the agreement of each model with the data. In our case $p(\mathbf{d}|\lambda = 0) \ll 1$, then this term strongly prefers Mr. B's theory. However, the goodness-of-fit cannot be the only thing that matters; if it was, we would always prefer more complicated explanations. indeed, probability theory tells us that there is another term to be considered. As assumed earlier in the evaluation of the marginal integral of Eq. (3.11), the prior range $\lambda_\text{max} - \lambda_\text{min}$ will generally be much larger than the uncertainty σ_λ permitted by the data. This means that the final term in Eq. (3.13) acts to penalize B for the additional parameter (and it will be larger increasing the number of parameters). For this reason, it is used to say that the Bayes' factor include the effect of the *Ockham's razor*. This factor can play a crucial role when both theories give comparably good agreement with the measurements and it becomes increasingly important if B's theory fails to give a significantly better fit as the quality of the data improves.

In our example, Mr. B's theory is strongly preferred since the measurement leads to the value 5.0 ± 0.7, which lies very far from the prediction of Mr. A. This result is also expressed from the Bayes' factor in Eq. (3.13) which, in our case, corresponds to $\mathcal{B}^\text{A}_\text{B} = 9.5 \times 10^{-10}$.

3.2.1 Information

A good experiment was one that yielded a likelihood function which was much more sharply peaked than the prior, otherwise, very little is learnt from the measurements. This basic idea can be quantified through the notion of **entropy**,

$$\mathcal{H} = \int p(\boldsymbol{\theta}|\mathbf{d}) \log_2 \left[\frac{p(\boldsymbol{\theta}|\mathbf{d})}{p(\boldsymbol{\theta})}\right] d\boldsymbol{\theta} , \tag{3.14}$$

with logarithm to the base 2, since the information is measured in *bits*. For example: consider a parameter θ that has just two equivalent states. Then the prior would be $p(\theta) = (1/2, 1/2)$ and the posterior can be $(1, 0)$ or $(0, 1)$. This situation yields to the information,

$$\mathcal{H} = 1 \cdot \log_2(2) + 0 \cdot \log_2(0) = 1\,,$$

which means that we need just one bit to store a single information, 0 or 1. In this calculation, we used the limit $x \log_2 x \to 0$ for $x \to 0$. Similarly, for four equivalent states with a prior $p(\theta) = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$, we gain the information $\mathcal{H} = 2$, which corresponds to the number of digits necessary to express the numbers from 0 to 3 in binary code. So, the quantity $\mathcal{H}$ give us the minimum value of bits needed to store an information and Eq. (3.14) is the corner-stone of information theory, founded by Shannon (1948).

Now suppose that we design an experiment to measure $\boldsymbol{\theta}$ and the observations produce the data $\mathbf{d}$, and we know the likelihood function $p(\mathbf{d}|\boldsymbol{\theta})$. Though these formulae are general, we are only designing the equipment and have not yet acquired specific data. Nevertheless, we can combine Eq. (1.9) and Eq. (1.10) with Eq. (3.14) to discover the amount of information,

$$\mathcal{H}(\mathbf{d}) = \int \frac{p(\mathbf{d}|\boldsymbol{\theta})\, p(\boldsymbol{\theta})}{p(\mathbf{d})} \log_2 \left[\frac{p(\mathbf{d}|\boldsymbol{\theta})}{p(\mathbf{d})} \right] d\boldsymbol{\theta}\,, \tag{3.15}$$

that we have about $\boldsymbol{\theta}$ if we did the experiment. Even, before acquiring the data, we can obtain the expected information as

$$E[\mathcal{H}] = \int \mathcal{H}(\mathbf{d})\, p(\mathbf{d})\, d\mathbf{d}\,. \tag{3.16}$$

This quantity represent the benefit of the experiment, quantifying the amount of information about $\boldsymbol{\theta}$ which the experiment is expected to provide.

3.3 Hierarchical Models

Hierarchical regression models are useful as soon as there are predictors at different levels of variation. Using a hierarchical approach, we write the parameters of the prior as functions of other unknown parameters. The sub-models combine to form the hierarchical model, and Bayes' theorem is used to integrate them with the observed data and account for all the uncertainty that is present. The hierarchical approach is very important also for the clusters analysis and for the extrapolation of informations from populations, which compose many social and natural phenomena: the students in a school, patients in a medical study, the property of an astronomical cluster and so on.

It is important to notice that non-hierarchical model are inappropriate for hierarchical data; in fact with few parameters, generally, the model cannot fit the data, while using many parameters, we incur in the problem of overfitting.

3.3.1 Exchangeability

Consider a set of observation $j = 1, \ldots, n$ with the respective vector of data $\mathbf{d}$ and parameters $\boldsymbol{\theta}$. If no information, except $\mathbf{d}$, is available to distinguish any of the θ_j from any other, and no grouping or ordering can be made on it, we must assume symmetry among the parameters and the prior. Strictly, we have to assume that, if we *exchange* the parameters $\theta_1, \ldots, \theta_n$, the prior pdf $p(\theta_1, \ldots, \theta_n)$ is invariant under permutation of these parameters.

The simplest exchangeable distribution has each θ_j coming from an independent distribution governed by an unknown parameter ϕ, i.e. the different θ_j are independent and identically distributed variables,

$$p(\boldsymbol{\theta}|\phi) = \prod_{j=1}^{n} p(\theta_j|\phi)\,. \tag{3.17}$$

Marginalizing, we found the usual prior

$$p(\boldsymbol{\theta}) = \int p(\boldsymbol{\theta}|\phi)\, p(\phi)\, d\phi\,, \tag{3.18}$$

where $p(\phi)$ is the prior pdf that we assume for the new parameter ϕ. This probability density is sometimes called the **hyper-prior**, since it represents the condition of a more fundamental set of parameters, i.e. ϕ which is labelled as *hyper-parameter*. It is also easy to generalize the formalism to the case in which the hyper-prior has a multidimensional support.

If we are interested in the analytical form of the posterior distribution of ϕ, we can proceed in two different ways:

1. If we know the analytic for of the joint posterior $p(\phi, \boldsymbol{\theta}|\mathbf{d})$ and the marginal posterior of $\boldsymbol{\theta}$, i.e. $p(\boldsymbol{\theta}|\phi, \mathbf{d})$, we can directly apply the Bayes' theorem and get

   $$p(\phi|\mathbf{d}) = \frac{p(\phi, \boldsymbol{\theta}|\mathbf{d})}{p(\boldsymbol{\theta}|\phi, \mathbf{d})}\,. \tag{3.19}$$

   However, this relation could have some problems in the normalization constants, since they could depend on ϕ as well as $\mathbf{d}$.

2. If we combine Eq. (1.7) with Eq. (3.18), we can write the joint posterior pdf in terms of the priors and the usual likelihood,

   $$p(\boldsymbol{\theta}, \phi|\mathbf{d}) = p(\mathbf{d}|\boldsymbol{\theta})\, p(\boldsymbol{\theta}|\phi)\, p(\phi)\,, \tag{3.20}$$

   and, marginalizing we get

   $$p(\phi|\mathbf{d}) = \int p(\boldsymbol{\theta}, \phi|\mathbf{d})\, d\boldsymbol{\theta}\,. \tag{3.21}$$

Thanks to Eq. (3.19) or Eq. (3.21), we can write the marginal posterior pdf of the hyper-parameter ϕ and then we are able to infer on it.

Let us make some example: if we trow a die, every output have probability $\theta_j = 1/6$ for $j = 1, \ldots, 6$. If we include, in the physics model, the imperfection of the die, the θ_j become exchangeable, but they cannot be modeled as independent, since $\sum_j \theta_j = 1$. This yields that the prior cannot be an independent and identically distributed mixture, but the analytic form is still invariant under permutations of j.

3.3.2 Gaussian Linear Models

Let us consider a two-level model, so we have a vector of data $\mathbf{d}$, which is related to some vector of parameters $\boldsymbol{\theta}$ by a linear relation, say

$$\mathbf{d} = \mathbf{A}\,\boldsymbol{\theta} + \mathbf{e}\,,$$

where $\mathbf{A}$ is the matrix that define the linear relation and $\mathbf{e}$ is a gaussian error, with mean $\langle\mathbf{e}\rangle = 0$ and covariance matrix $\mathrm{Cov}(\mathbf{e}) = \mathbf{V}$. Then, we assume that the parameters $\boldsymbol{\theta}$ is defined by another linear relation with the hyper-parameters $\boldsymbol{\phi}$, say

$$\boldsymbol{\theta} = \mathbf{X}\,\boldsymbol{\phi} + \boldsymbol{\epsilon}\,,$$

where $\mathbf{X}$ is another matrix and $\boldsymbol{\epsilon}$ is a gaussian error with zero mean and covariance $\mathrm{Cov}(\boldsymbol{\epsilon}) = \mathbf{W}$. So, $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ are respectively the first-level and the second-level parameters and $\mathbf{A}$, $\mathbf{X}$ are the first-level and second-level design matrix. We observe that in general the two parameters vectors have different dimension, say n and m. The aim of Bayesian inference is therefore to be able to infer on the parameters $\boldsymbol{\theta}$ and on the $\boldsymbol{\phi}$, even not knowing anything a priori about them, basing our assertions on the posterior pdf $p(\boldsymbol{\theta}|\mathbf{d})$ and $p(\boldsymbol{\phi}|\mathbf{d})$.

In order to extract informations from the data, we can write the likelihood function $p(\mathbf{d}|\boldsymbol{\theta})$ and the prior $p(\boldsymbol{\theta})$, recalling the Gaussian structures of the uncertainties $\mathbf{e}$, $\boldsymbol{\epsilon}$, as

$$\begin{aligned}\log p(\mathbf{d}|\boldsymbol{\theta}) &\propto -\frac{1}{2}\big[(\mathbf{d}-\mathbf{A}\boldsymbol{\theta})^{\mathrm{t}}\,\mathbf{V}^{-1}\,(\mathbf{d}-\mathbf{A}\boldsymbol{\theta})\big]\,,\\ \log p(\boldsymbol{\theta}) &\propto -\frac{1}{2}\big[(\boldsymbol{\theta}-\mathbf{X}\boldsymbol{\phi})^{\mathrm{t}}\,\mathbf{W}^{-1}\,(\boldsymbol{\theta}-\mathbf{X}\boldsymbol{\phi})\big]\,,\end{aligned}$$

then, the posterior, using Eq. (1.7), can be expressed as $p(\boldsymbol{\theta}|\mathbf{d}) \propto p(\mathbf{d}|\boldsymbol{\theta})\,p(\boldsymbol{\theta})$. Keeping only the terms in $\boldsymbol{\theta}$, we get

$$p(\boldsymbol{\theta}|\mathbf{d}) = \mathrm{N}_n(\boldsymbol{\theta}|\boldsymbol{\mu}_\theta, \boldsymbol{\Sigma}_\theta)\,,$$

where the mean and the covariance are

$$\boldsymbol{\mu}_\theta = \boldsymbol{\Sigma}_\theta\Big(\mathbf{A}^{\mathrm{t}}\mathbf{V}^{-1}\mathbf{d} + \mathbf{W}^{-1}\mathbf{X}\,\boldsymbol{\phi}\Big) \quad , \quad \boldsymbol{\Sigma}_\theta^{-1} = \mathbf{A}^{\mathrm{t}}\mathbf{V}^{-1}\mathbf{A} + \mathbf{W}^{-1}\,. \tag{3.22}$$

Now we can apply the Bayes' theorem to obtain the posterior pdf of $\boldsymbol{\phi}$,

$$p(\boldsymbol{\phi}|\mathbf{d}) = \frac{p(\mathbf{d}|\boldsymbol{\phi})\,p(\boldsymbol{\phi})}{p(\mathbf{d})}\,,$$

where $p(\boldsymbol{\phi})$ represent the hyper-prior of $\boldsymbol{\phi}$, that it is assumed uniform since we know anything about the hyper-parameters. So the posterior becomes equal to the likelihood function $p(\mathbf{d}|\boldsymbol{\phi})$, and defining the hierarchical relation,

$$\mathbf{d} = \mathbf{A}\mathbf{X}\boldsymbol{\phi} + \mathbf{A}\boldsymbol{\epsilon} + \mathbf{e} = \tilde{\mathbf{A}}\,\boldsymbol{\phi} + \tilde{\mathbf{e}}\,,$$

where

$$\tilde{\mathbf{A}} = \mathbf{A}\mathbf{X} \quad , \quad \tilde{\mathbf{e}} = \mathbf{A}\boldsymbol{\epsilon} + \mathbf{e} \quad , \quad \tilde{\mathbf{V}} = \mathrm{Cov}(\tilde{\mathbf{e}}) = \mathbf{V} + \mathbf{A}\mathbf{W}\mathbf{A}^{\mathrm{t}}\,.$$

Then, we can write the posterior pdf of the hyper-parameters as a Gaussian,

$$p(\boldsymbol{\phi}|\mathbf{d}) = \mathrm{N}_m(\boldsymbol{\phi}|\boldsymbol{\mu}_\phi, \boldsymbol{\Sigma}_\phi)\,,$$

with mean and variance

$$\boldsymbol{\mu}_\phi = \boldsymbol{\Sigma}_\phi\,\tilde{\mathbf{A}}^{\mathrm{t}}\tilde{\mathbf{V}}^{-1}\mathbf{d} \quad , \quad \boldsymbol{\Sigma}_\phi = \Big(\tilde{\mathbf{A}}^{\mathrm{t}}\tilde{\mathbf{V}}^{-1}\tilde{\mathbf{A}}\Big)^{-1}\,, \tag{3.23}$$

We have achieved the posterior distribution for the first-level and the second-level parameters, expressed in terms of the data and the covariance matrix. In the simplest case of an univariate model, where

$$d = a\cdot\theta + e \quad , \quad \theta = x\cdot\phi + \epsilon\,,$$

calling the variances $\mathrm{Var}(e) = v^2$ and $\mathrm{Var}(\epsilon) = w^2$, we get

$$\mu_\theta = \frac{1}{\sigma_\theta^2}\left(\frac{d}{v^2} + \frac{\phi}{w^2}\right) \quad , \quad \sigma_\theta^2 = \left(\frac{1}{v^2} + \frac{1}{w^2}\right)^{-1} ,$$

$$\mu_\phi = \frac{d}{a\,x} \quad , \quad \sigma_\phi^2 = \frac{1}{x^2}\left(w^2 + \frac{v^2}{a^2}\right) .$$

Chapter 4

Monte Carlo Methods

The Monte Carlo methods are computational statistical techniques to computing integrals using random positions, called **samples**, whose distribution is carefully chosen. The MC methods are extremely general, and the basic recipes allow us, in principle, to solve any problem in statistical physics.

The target probability distribution might be a distribution from statistical physics or a conditional distribution arising in data modeling. Then, this distribution can be useful to evaluate statistical quantities of interest, such as mean and variance of the different parameters. The Monte Carlo (MC) methods use random paths through the parameters' space in order to identify the bulge of the distribution. This method effectively coincides with a thermalization process, where the potential is described by the target probability function.

Let us make an example: presently, we know that the value of π is $3.1415926535\ldots$, but for a moment we conjecture to forget the value of π. So, how can we get a sufficiently good estimate? A good idea is the following: we draw a circle of radius r and a circumscribed square, with side $\ell = 2\,r$. From geometry, we know that the ratio of these areas is

$$\frac{A_{\text{circle}}}{A_{\text{square}}} = \frac{\pi r^2}{\ell^2} = \frac{\pi}{4}\,.$$

So, if we throw a sufficiently large number N of points, uniform distributed in plane inside the square, and we count the number N_{in} of points inside the circle, then the ratio

$$\frac{N_{\text{in}}}{N} = \frac{\pi}{4}\,,$$

and we are able to estimate the value of π multiplying the previous ratio by 4. The result of the algorithm is shown in Fig. **??**. The procedure can be expressed using the *pseudocode* as it is shown in Alg. 1.

The power of the MC methods manifests itself especially in multidimensional problems. Suppose to generalize the previous example to a n-dimensional problem. Then, the probability that the i-th outcome falls into or out of the circumference is described by a binomial distribution

$$N_{\text{in}} \sim \text{Bin}(p, N)\,,$$

with $p = \pi/4$. Then we have $\langle N_{\text{in}} \rangle = pN$ and $\text{Var}(N_{\text{in}}) = Np(1-p)$. So, we can write the convergence speed of the algorithm using the relation

$$\left|\frac{\pi}{4} - \frac{N_{\text{in}}}{N}\right| < \varepsilon(N) \quad \Rightarrow \quad \varepsilon(N) \propto \frac{1}{\sqrt{N}}\,. \tag{4.1}$$

Algorithm 1

```
$N_{\mathtt{in}} \leftarrow 0$                                   ▷ Initialize the counter
for $i \leftarrow [1, N]$ do
    $x_i \leftarrow$ U(x|-1,1)                                ▷ Get a random value on the $x$-axes
    $y_i \leftarrow$ U(y|-1,1)                                ▷ Get a random value on the $y$-axes
    if $x_i^2 + y_i^2 < 1$ then        ▷ If the sampled point $(x_i, y_i)$ is inside the circle...
        $N_{\text{in}} \leftarrow N_{\text{in}} + 1$                                   ▷ ...count it
    end if
end for
$\pi \leftarrow 4 \cdot N_{\text{in}} / N$                                    ▷ Estimate $\pi$
```

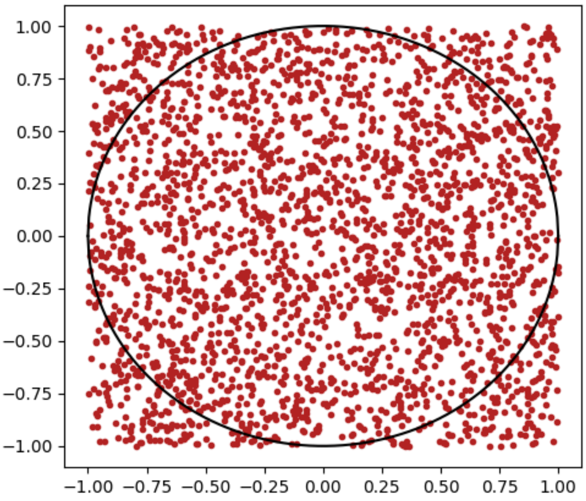

Figure 4.1: Example of estimation of π using the Monte Carlo method implemented in Alg. 1. The run was performed using a total number of points $N = 2500$ and $N_{\text{in}} = 1958$ of them felt inside the circumference. Then we estimated the value of $\pi \approx 3.1328$. However, this measurement has an error of $\varepsilon(N = 2500) \approx 0.0088$. If we want to increase the accuracy we have to increase the number of extracted points. Let us say we need $\varepsilon(M) = 10^{-4}$, then using Eq. (4.1), we get $\frac{\varepsilon(N)}{\varepsilon(M)} = \frac{\sqrt{M}}{\sqrt{N}}$, which means $M \approx 2 \times 10^7$.

This result tells us an obvious fact: increasing the total number of thrown points N we increase the accuracy of the estimation; however Eq. (4.1) is valid for every MC process in parameters' space with arbitrary dimension. Otherwise, if we choose to compute the integral analytically we incur in the rest of Lagrange, which brings an uncertainty proportional to $N^{-2/n}$ (where N is the total number of points and n is the dimension of the problem). So, for $n > 4$ we can easily see that the statistical evaluation of an integral becomes favored than the analytical approach.

In general, the Monte Carlo methods are able to generate a set of samples from a posterior pdf and to estimate the integrals, that, in the case of interest, are the expectation values of the physical observables. Usually this techniques use a chain, in which every step is an iteration of a partial algorithm. If the outcome of a step is determined (or conditioned) by the previous one, the algorithm is called **Markov Chain Monte Carlo** (MCMC).

4.1 Rejection Sampling

The simplest application of a MC method is the rejection sampling; our purpose is to generate samples from a probability density $p(x)$ starting from a prior distribution $\Pi(x)$ (usually taken as flat in a prior range), which is simpler and easier to control. We assume a one-dimensional problem, then we can write the pseudocode for the rejection sampling algorithm as it is shown in Alg. 2.

Algorithm 2

```
iter← 0                                   ▷ Initialize the counter
while iter< $N$ do
    $x^* \sim \Pi(x)$                     ▷ Generate an element from $\Pi(x)$
    $p^* \sim \mathrm{U}[\,0,1\,]$        ▷ Generate a random probability
    if $p^* < p(x^*)$ then                ▷ If $x^*$'s probability is large enough ...
        Collect $x^*$
        iter←iter+1
    end if
end while
```

The algorithm stops when it produces N samples. We can see that this method is very roughly: we generate a point x^* in the prior, then if the probability $p(x^*)$ is greater than a random value p^* between 0 and 1, we accept the new point, otherwise we discard it and we iterate the algorithm until `iter` (the iterator) reaches N. We note that the random extraction of p^* is fundamental in order to describe the target pdf, since it introduce some indetermination in the algorithm.

This method is able to generate a good set of independent samples, but usually it is not the best way to obtain a sufficiently description of the posterior pdf $p(x)$, since the rejection sampling is not able to return smooth functions. Moreover, a rejection sampling is usually very computational expensive, i.e. it requires a lot of time to complete the loop, since the acceptance ratio could be very low, especially if the prior distribution is very different form the posterior.

4.2 Metropolis-Hastings Sampling

A powerful implementation of a MCMC method comes from Metropolis and Hastings. In contrast to rejection sampling, it is not necessary that $\Pi(x)$ look at all similar to $p(x)$ in order for the algorithm to be practically useful, because using this approach the prior density depends on the current state x^* and it can be any fixed density from which we can draw samples. More exactly, calling $\Pi(x)$ the prior distribution of the variable x of interest, $q(s)$ is a **proposal distribution** that determines the step s in the parameter's space and $p(x)$ is the posterior pdf, also called in this context, **target distribution**. We describe a simplified version of the pseudocode for a Metropolis-Hastings sampling as it is shown in Alg. 3.

The parameter α is arbitrary parameter (usually sampled in a flat prior between 0 and 1, as it is shown in our example), its role is to introduce more randomness in the algorithm. We can see that the i-th outcome depends on the previous result x_{i-1}, or, more precisely, on the posterior probability of the previous outcome. However, it is important to analyze the convergence of Metropolis sampling; since the algorithm start from a random point in the prior, it need some steps to reach the range of the posterior maximum.

This approach become more interesting increasing the number of parameters, in fact MCMC methods play a crucial role in the evaluation of multidimensional integrals. Especially when the dimensionality of the problem increases, it became harder to compute analytically the values of the interested integrals, so a simulations became a powerful tool.

The Metropolis Sampling evaluates with a good approximation the target distribution and the

Algorithm 3

```
$x_0 \sim \Pi(x)$                              ▷ Sample an initial value from the prior
for $i \leftarrow [1, N]$ do
    $s_i \sim q(s)$                            ▷ Sample a random step
    $x_i \leftarrow x_{i-1} + s_i$             ▷ Compute the new point
    $\alpha_i \leftarrow \mathrm{U}(\alpha|0,1)$   ▷ Assign random value
    if $p(x_i) > \alpha_i \cdot p(x_{i-1})$ then
        Collect $x_i$                          ▷ Collect the new point for the new interation
    else
        $x_i \leftarrow x_{i-1}$               ▷ Use the old point for the new iteration
    end if
end for
```

statistical quantities, such as the median value and the variance, and give us back a satisfactory set of posterior samples. However, this algorithm (and MCMC methods in general) is not able to help us in the model selection, since these codes are not optimize to estimate the value of the evidences of different models. An algorithm able to estimate the evidence with a very good accuracy is the *nested sampling*, but this topic will not be treated in these notes.

We observe that Alg. 3 can be generalized for the case of multi-dimensional parameter space. In this case the value x becomes a point in a n-dimensional parameters' space $\mathbf{x} = (x^{(1)}, x^{(2)}, \ldots, x^{(n)})$. Now every sample is an array of n elements $\mathbf{x}_i = (x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(n)})$ and then also the proposal distribution $q(\mathbf{x})$ must admit a n-dimensional domain.

4.2.1 Exploring the Parameters' Space

In an MCMC chain, we start with some arbitrary initial position in the parameters' space and we move our vector in a random direction according to the proposal distribution. However, the proposal distribution can be informed from the set of collected samples, i.e. we can compute the correlation of the samples and move the point consistently in order to make more efficient steps.

A standard practice is to start with a multidimensional normal distribution with mean zero and covariance matrix equal to the identity (or another diagonal matrix), after a certain number m of steps, we can evaluate the covariance matrix of our collected samples and we replace it instead of the previous one, and so on reaching the optimal covariance matrix for our steps in order to run across a quicker path for maximize the likelihood and so, to make the iteration faster and more efficient. This technique is called **adaptive** MC, since the algorithm adapt his jobs over the previous data. We can easily implement this option inserting the follows algorithm inside the *for* cycle, as it is shown in Alg. 4.

Algorithm 4

```
for $i \leftarrow [1, N]$ do
    ...                                        ▷ The previous code
    if $i \bmod m = 0$ then
        $C \leftarrow \mathrm{cov}(\mathbf{x})$    ▷ Adapt the covariance of the step
    end if
end for
```

Moreover, since a chain start from a random point, it is possible that a single sequence of extracted points is not enough to explore properly the parameters' space. This is very common with multimodal target distribution: if a chain falls in a local maximum, then with a Metropolis sampling, this chain is not able to escape from this maximum and find the other one(s). This is due to the architecture of this algorithm. In order to avoid this problem, usually the sampling is performed with **multiple parallel chains**, where each of those is a Metropolis sampling which started from its random initial point (different from the others). In this way, different chains are able to explore different regions of the parameters' space and moreover, in the case of multimodal target distribution, the importance of every peak is given by the ratio between the number of chains that converged into the peak and the total number of chains.

4.2.2 Extraction of Independent Samples

Fig. 4.2 shows an example of the evolution of a chain in a Metropolis-Hastings sampling. We performed a uni-dimensional sampling with a gaussian target distribution and proposal distribution, such that

$$p(x) \sim \mathrm{N}(x|\mu = 9.38, \sigma = 0.02)\,, \quad q(s) \sim \mathrm{N}(s|\mu = 0, \sigma = 0.01)\,.$$

In Fig. 4.2, it is possible to observe that the initial value is randomly chosen and then the algorithm explores the parameter's space, moving in order to find the region that maximize the likelihood. However, zooming on the chain, we can see that not all the points are independent from the others. The reason is the following: the algorithm saves a point in the chain at every iteration, but it actually moves toward a new point only when its likelihood increase with respect to the likelihood of the previous one, otherwise the same point is kept. So we are collecting series of identical numbers that are not independent at all. Then the sampled chain could appear independent on large scale but if we go to analyze in detail it's easy to see that the algorithm repeats the same value until it accepts a new one.

First of all, the initial part of the chain, called the *shrinking* or *burn-in*, is not extracted from the correct distribution. In fact the first value is randomly chosen from the algorithm and the subsequent ones range in the parameters' space until they reach the bulge of the likelihood. So we cannot consider these first points as good ones, they do not came from our researched distribution, we have to wait until the thermalization process of the algorithm converges. Usually, the first half of the chain is discarded, so if we start with a chain of N elements, we obtain a chain of $N/2$ elements.

Then we have the problem of repeated points. In order to avoid this problem and to extract an independent set of samples, we have to evaluate the **autocorrelation function** (ACF) of the chains and determine its autocorrelation length (ACL). Then we will take one samples every ACL from the remaining chain. For a stochastic stationary process, the ACF function depends on the lag ℓ, i.e. the difference in the positions of the elements. This function is even with respect to ℓ and then we can take into account just the positive semi-axis of ℓ. We can evaluate the ACF of a set of samples $\{\theta_i\}$ as follows,

$$\mathrm{ACF}_\ell\,(\theta) = \sum_{k=1}^{N} \frac{\big(\theta_k - E[\theta]\big)\big(\theta_{\ell+k} - E[\theta]\big)}{\sigma_k\,\sigma_{\ell+k}}\,, \tag{4.2}$$

where $E[\theta]$ is the mean of the samples and σ_k is the standard deviation of the k-th element.

We clearly expect that for values of the lag around zero the autocorrelation is very close to the unity (since every point is correlated with itself) and then it decreases fluctuating around zero, if the samples are actually independent. An example of ACF is shown in Fig. 4.3. The ACL is defined from the ACF as the first value of ℓ for which the ACF becomes zero, i.e. the lower intersection between the ACF and the lags axis. If we have a set of multidimensional parameter, i.e. $\boldsymbol{\theta} = (\theta^{(1)}, \theta^{(2)}, \dots)$, the ACL of the entire set is taken to be the larger ACL for every single parameter. We note also that the ACL for a set of sample is an integer number.

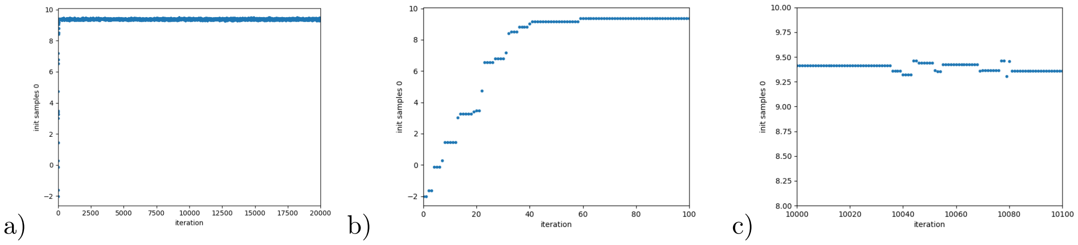

Figure 4.2: The figure shows a chain (of 20000 elements) collected with a Metropolis-Hastings algorithm. In the first panel (a) it is shows the overall chain. The second panel (b) represents a zoom on the first iteration in order to visualize better the shrinking. The third panel (c) is a zoom around the center of the chain: one can observe the autocorrelation between the elements and the high recurrence of the same elements.

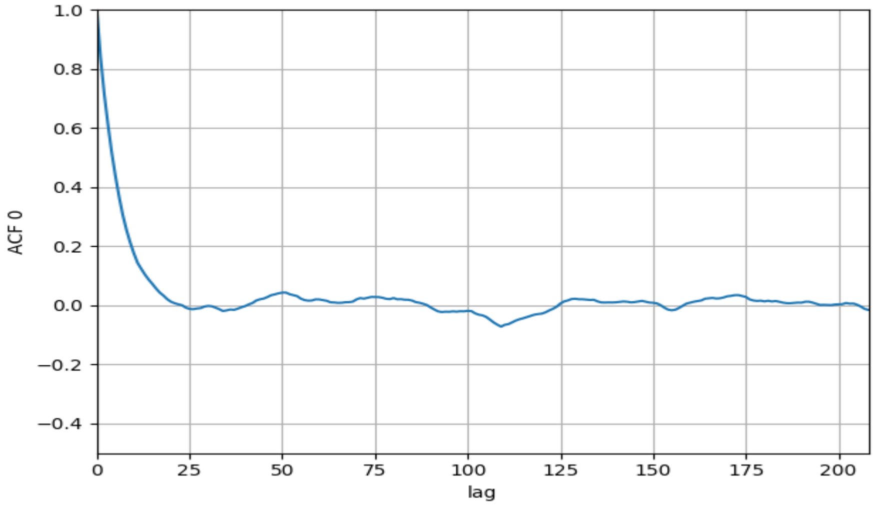

Figure 4.3: ACF of the chain shown in Fig. 4.2. As we expect in zero the ACF is equal to 1, and then decrease reaching zero for the first time at $\ell = 24$. So we set ACL= 24. We are allowed to use this value because after $\ell = 24$, the ACF fluctuates around zero, proving that above a lag of 24 elements there are no correlations (in average) between the elements of the chain.

Let us summarized: suppose that a Metropolis-Hastings algorithm returns a list of N elements, labelled as $\{\boldsymbol{\theta}_i\}$. This points are not independent and this list contains the shrinking portion, which is not supported by the target distribution a priori. So we throw away the initial segment of the chain (usually the first half of the chain) and we remain with a $N/2$-dimensional array labelled as $\{\boldsymbol{\theta}'_i\}$. Then we evaluate the ACL, in order to construct an independent set of samples, labelled as $\{\hat{\boldsymbol{\theta}}_i\}$: we append to this list the first element of $\{\boldsymbol{\theta}'_i\}$, the ACL-th one, the (2×ACL)-th one and so on until $(M+1) \times \text{ACL} > N/2$. At this point we have obtained an array with $M = N/(2 \cdot \text{ACL})$ elements which corresponds to an independent set of samples extracted from the target distribution. We can compute these processes as it is shown in Alg. 5, where n is the number of parameters and N is the initial number of points in the chain.

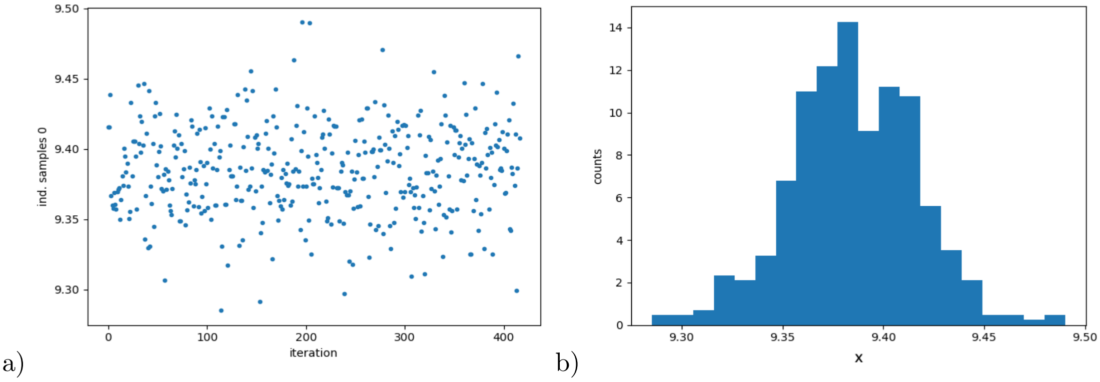

Figure 4.4: The figures show the final set of independent samples. Left panel is the scatter plot of the 417 elements of $\hat{x}$. The right panel shows the histogram of the samples extracted from the target distribution. From this information, we can extrapolate the most important stochastic quantities as the mean, the variance and the evidence.

Algorithm 5

```
Estimate $\mathrm{ACF}(x)$
$\mathrm{ACL} \leftarrow$ first root of $\mathrm{ACF}(x)$                ▷ Estimate ACL
$\{x'\} \leftarrow x_{[N/2,N]}$                              ▷ Throw away the shrinking
for $i \leftarrow [1, N/(2 \cdot \mathrm{ACL})]$ do
    $\hat{x}_i \leftarrow {x'}_{i \cdot \mathrm{ACL}}$                    ▷ Collect an element each ACL iteration
end for
```

In our example, we start with 20000 samples and in the end we get ACL= 24, so the effective number of independent elements is 417 that describe satisfactory the posterior distribution of our parameters. For higher-dimension problems, the number of samples required to describe the target distribution increases, since the algorithm has to investigate the different correlations and explore properly a large parameters' space. It's important to observe that we have thrashed about the 98% of the initial set and this is happened for a unidimensional problem. We will see that increasing the dimensionality of the space (in other words increasing the number of free parameters), the efficiency falls down very rapidly. So when we are going to perform MCMC techniques in multi-dimensional spaces, we must increase the numbers of initial points to be able to generate a sufficient statistic.

Now from the samples $\hat{x}$ we can reproduce the posterior distribution through an histogram (Fig. 4.4) and analyze its statistical property, such as mean and standard deviation,

$$x = 9.37 \pm 0.03\,,$$

or median and 90% credible interval,

$$x = 9.38 \pm 0.05\,.$$

Although the value of the median is preferred.

4.2.3 Convergency and Efficiency

It can be shown that for any proposal distribution $q(s)$, the probability distribution described by x_i tends to the target distribution $p(x_i)$ for $i \to +\infty$. The Metropolis method is an example

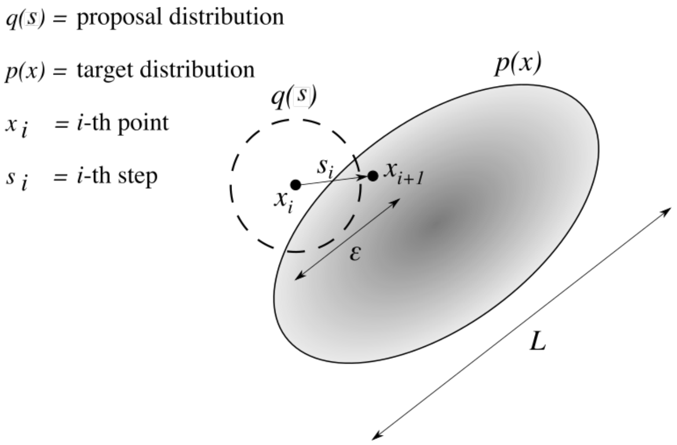

Figure 4.5: Graphical representation, in a 2-dimensional parameters' space, of a random step in a Metropolis-Hastings sampling algorithm. The terms $p(x)$ and $q(s)$ denote respectively the target and the proposal distributions. At every iteration, a step s_i is extracted from $q(s)$ and it defines the random-walk through the parameters' space. Then, the i-th point of the chain x_i is moved according with the random step $x_{i+1} = x_i + s_i$. If $p(x_{i+1}) > p(x_i)$, the new point is accepted, otherwise it is discarded. The figure shows also the characteristic length-scales L and ε.

of a MCMC. In contrast to rejection sampling, where the accepted points $\{x_i\}$ are independent samples from the desired distribution, Markov chain Monte Carlo methods involve a Markov process in which a sequence of states $\{x_i\}$ is generated, each sample x_i having a probability distribution that depends on the previous value, x_{i-1}. Since successive samples are dependent, the Markov chain may have to be run for a considerable time in order to generate samples that are effectively independent samples from the target $p(x)$.

The Metropolis method is widely used for high-dimensional problems. Many implementations of the Metropolis method employ a proposal distribution with a length scale ε that is short relative to the longest length scale L of the probable region (Fig. 4.5). A reason for choosing a small length scale is that for most high-dimensional problems, a large random step from a typical point (that is, a sample from $p(x)$) is very likely to end in a state that has very low probability; such steps are unlikely to be accepted. If ε is large, movement around the state space will only occur when such a transition to a low-probability state is actually accepted, or when a large random step chances to land in another probable state. So the rate of progress will be slow if large steps are used.

The disadvantage of small steps, on the other hand, is that the Metropolis method will explore the probability distribution by a **random walk**, and a random walk takes a long time to get anywhere, especially if the walk is made of small steps. Recall that the first aim of Monte Carlo sampling is to generate a number of *independent* samples from the given distribution (a dozen, say). If the largest length scale of the state space is L, then we have to simulate a random-walk Metropolis method for a time

$$T \approx \left(\frac{L}{\varepsilon}\right)^2 \tag{4.3}$$

before we can expect to get a sample that is roughly independent of the initial condition âĂŞ and thatâĂŹs assuming that every step is accepted: if only a fraction f of the steps are accepted on average, then this time is increased by a factor $1/f$.

This rule of thumb gives only a lower bound; the situation may be much worse, if, for example, the probability distribution consists of several islands of high probability separated by regions of low probability.

Bibliography

[1] Thomas Bayes, *An essay towards solving a problem in the doctrine of chances.* A. M. F. R. S. 53 *Phil. Trans. R. Soc.* (1763).

[2] Pierre Simon, Marquis de Laplace, *ThÃľorie analytique des probabilitÃĺs.* Courcier Imprimeur, Paris (1812).

[3] D. J. C. MacKey, *Information Theory, Inference, and Learning Algorithms.* Cambridge University Press (2003).

[4] D. S. Sivia with J. Skilling, *Data Analysis, a Bayesian Tutorial.* Oxford University Press (2006).

[5] W. Krauth, *Statistical Mechanics: Algorithms and Computations.* Oxford University Press (2006).

[6] A. Gelman, G. B. Carlin, H. S. Stern and D. B. Rubin, *Bayesian Data Analysis.* Chapman and Hall, CRC (2013).