

Random Notes on Probabilities

Alamino, R.C.

May 21, 2009

Preface

These notes are a collection of techniques and facts related to inference, statistics and probabilities. They are not intended to be rigorous in any sense. Most of the formulas included here are true for well behaved distributions, but some may fail in pathological cases, specially anything involving analytical continuations.

The main idea is to put together in one single document useful tricks that are commonly used by all professionals in the area but are not usually written in textbooks. Some because they are considered obvious (or at least almost) as they are very easy to derive, some because they are not used too often and some because they seem useless, which may very well be the truth. I will leave it up to the reader to judge those matters.

The reader may feel that the topics in the text are not too much correlated and that it resembles more a collage than a whole picture. The reader is absolutely right in this aspect and that is the reason why I am calling it *random* notes on probability. As connections are everywhere in mathematics, it is possible that this feeling disappears as the text continues to expand, however the title already express my intention of not worrying to much about that.

The level of the notes is very basic and I believe they are easily understood by undergraduate students with a minimal knowledge of probability theory probably in the form of a basic course. Clearly, as it is been a long time since I was an undergraduate student, my judgement may be distorted in this respect. However, I will need to wait for some feedback to know the extent of my mistake.

These notes are an ongoing project, so I would be very grateful to receive comments, critics and/or suggestions about anything from content to layout (specially because Latex always give me a hard time on this matter). Many proofs and details are lacking and I intend to expand the text as much as it

is possible with time.

I would like to add some thanks to Prof. Nestor Caticha, who I would say is close to be a direct contributor as he provided me with one of the proofs of the normalisation of the Dirichlet and as most of what I have learnt about the topics here I learnt in my Ph.D. thesis. I also would like to thank my wife, for so many reasons that I cannot even start to write here.

Roberto C. Alamino

Contents

1	Probabilities	1
1.1	Basic Rules and Definitions	1
1.2	Probability Mixtures	4
2	Dirac Deltas	7
2.1	Basic Definitions	7
2.2	Mathematical Properties	10
2.3	Jacobians	11
3	Entropy	13
3.1	Maximum Entropy	14
3.2	Kullback-Leibler Projection	15
4	Gaussians	17
4.1	Moments	18
4.2	Limit Theorems	20
5	Gammas and Betas	23
5.1	The Beta Distribution	23
5.2	The Gamma Distribution	24
6	Dirichlets	29
6.1	Normalisation	30
6.1.1	Method One: Complex Integration	30
6.1.2	Method Two: Analytical Continuation	33
6.2	Moments	34

Chapter 1

Probabilities

1.1 Basic Rules and Definitions

One of the potentially most confusing things in probability theory is the notation. The distinction between random variables and its values is a notorious example and usually is solved by the convention that random variables are symbolised by capital letters, for instance X , while their values are represented by small letters, as x . Unfortunately, in these notes, we will not adopt this particular convention in the hope that this distinction will be clear from the context. Indeed, we will not use it simply because along this text we will not gain too much in doing so. However, there is another kind of situation where a clarification of the notation will make our equations more understandable. We will define what we will call a **probability operator** by the symbol \mathcal{P} . What I mean by this is that \mathcal{P} can be simply substituted by the words *the probability of* whenever it appears. Therefore, when we write the expression $\mathcal{P}(x)$ we want to symbolise *the probability of x* and not some particular function of x named \mathcal{P} . There are cases where the probability of x will indeed assume a particular functional form dependant on x and possibly some set of parameters. For these cases, *if* we want to emphasise the particular functional form, we will usually use capital Latin letters in normal font, like P , Q etc, although this rule will not be so rigid. In some places, we may also indicate that x is distributed according to some probability distribution $P(x)$ by writing $x \sim P(x)$.

Also, we will be completely sloppy with the words event, proposition and random variable. Although rigorously the difference may be important, for

the desperation of the reader we will not pay attention to that in the main text and hope, once more, that the context will serve as a guide to choose the correct interpretation.

Another convention that will be used is that, although x can assume either a discrete or a continuous number of values, we will use for both cases the same notation \sum_x to indicate a summation over all the values of x . In the case these values are continuous, this should be understood as the integral $\int dx$, where the integration region is the whole set where x is defined. Using this notation, we define the average of some function $f(x)$ over x by

$$\langle f(x) \rangle_x = \sum_x \mathcal{P}(x) f(x). \quad (1.1)$$

When the variable over which the average is being taken is clear from the context, we may drop the subscript x in the angle brackets. Sometimes, when there is more than one possibility and we want to specify that the average is taken with $x \sim P(x)$ for a particular $P(x)$, we use the distribution as a subscript in the form $\langle f(x) \rangle_{P(x)}$.

One important rule that will be largely used in the rest of this text is the one defining conditional probabilities. Let us consider two variables x and y . The **conditional probability** of x given y then reads

$$\mathcal{P}(x|y) = \frac{\mathcal{P}(x, y)}{\mathcal{P}(y)}, \quad (1.2)$$

where $\mathcal{P}(x, y)$ is the joint probability of both variables. By the obvious fact that the order of the variables in the notation for the joint probability is immaterial, this definition implies the equality

$$\mathcal{P}(x, y) = \mathcal{P}(x|y)\mathcal{P}(y) = \mathcal{P}(y|x)\mathcal{P}(x), \quad (1.3)$$

which can be easily generalised to more than two variables and is known as the **chain rule** of probability.

The last equality in the above equation can be rearranged and then rewritten as

$$\mathcal{P}(x|y) = \frac{\mathcal{P}(y|x)\mathcal{P}(x)}{\mathcal{P}(y)}. \quad (1.4)$$

As $\mathcal{P}(x|y)$ is a probability, it must be normalised and therefore we have that

$$\mathcal{P}(y) = \sum_x \mathcal{P}(y|x)\mathcal{P}(x), \quad (1.5)$$

which is sometimes called the **partition function** in applications in physics and symbolised by the letter Z . Due to the fact that the normalisation is an obvious condition, with an obvious expression, it is sometimes common to write equation (1.4) as a proportionality by dropping the normalisation

$$\mathcal{P}(x|y) \propto \mathcal{P}(y|x)\mathcal{P}(x). \quad (1.6)$$

An important property of all the above rules is that they are true both for the case of discrete and continuous variables.

In order to satisfy the reader's impatience, let us finally say that equation (1.4) is also known as **Bayes' Rule** or **Bayes' Theorem**. Sometimes people will give this name to equation (1.3), which is *wrong!* In fact, Bayes' Theorem is a little bit more than simply equation (1.4). It is this equation plus an interpretation in terms of an inference process. There is an intrinsic "time order" in Bayes' Rule. This order is not causal, and this must be always stressed once that it is the source of much confusion in applications, but it is an *inference* order.

Each term in Bayes' Rule receives an interpretation in terms of this inference process. The conditional probability $\mathcal{P}(x|y)$ is called the **posterior distribution** of x given that the occurrence of y was observed. Its value is then equal to the product of $\mathcal{P}(y|x)$, the **likelihood** of y if we consider x as given, times $\mathcal{P}(x)$, the **prior distribution** of x . In most applications, x corresponds to a set of parameters of the system under study, while y is a set of data generated by this same system. The likelihood describes a model of the system and encodes how the system with parameters x would generate the observed data y . The prior encodes all information we might have about the parameters x . For instance, symmetry properties or the range in which its values fall. The construction of the prior distribution from the available information about x is a complex problem which, among other things, is related to the construction of the Lagrangean in physical theories. We will touch slightly on the problem of choosing a prior when we discuss the maximum entropy method in section 3.1.

Used in combination with the marginalisation rule

$$\mathcal{P}(x) = \sum_y \mathcal{P}(x, y), \quad (1.7)$$

conditional probabilities give rise to the *simple-looking* equation

$$\mathcal{P}(x) = \langle \mathcal{P}(x|y) \rangle_y, \quad (1.8)$$

where I have stressed the term *simple-looking* because in fact this equation is very far-reaching. More than people usually think or, if they think, more than they really tell. We are going to use it a lot but, before, we will have to be acquainted to some more techniques.

These are the basic probability rules we will need in the following chapters. There are two very interesting books about Bayesian inference, namely [1] and [2] which discuss in depth things like the role of the prior and the likelihood, but we will not extend ourselves too much into these areas in these notes.

1.2 Probability Mixtures

In inference problems it is sometimes convenient to define what is called a **probability mixture** which may also be called a **mixture model** in some cases. We can define continuous or discrete mixtures. A discrete mixture is a **convex combination** of probability distributions. A convex combination is defined as a linear combination of a set of N probability distributions $P_i(x)$ given by

$$Q(x) = \sum_i p_i P_i(x) \quad (1.9)$$

with

$$\sum_i p_i = 1, \quad 0 \leq p_i \leq 1. \quad (1.10)$$

Many times, we are interested in constructing mixtures of distributions that are in the same **parametric family**. A parametric family of distributions is defined by its functional form. All members of the family have the same functional form and differ from each other only by the values of a set of parameters θ that define the family. Examples of parametric families are the Gaussians, the Gamma and Beta distributions and the Dirichlets, all of them will have a specially dedicated chapter in these notes.

In order to differentiate between the different families, we will usually denote them by calligraphic capital letters, except for the letter \mathcal{P} which we are already using as our probability operator. However, as letters, specially calligraphic, are limited we will allow some freedom of notation which we will make clear when necessary. Let us denote a general parametric family by the letter \mathcal{F} . Then, to indicate that it gives the probability for some variable x and to make explicit that it depends on the parameters θ , we will write

it as $\mathcal{F}(x|\theta)$. This notation is completely consistent with our notation for conditional probabilities for the values of x are defined for some given θ to which, we will see soon, we can even attach a probability distribution for its values.

Now, equipped with the knowledge of parametric families, we can construct what is called a **parametric mixture** which is simply the probability mixture restricted to some parametric family

$$Q(x) = \sum_i p_i \mathcal{F}(x|\theta_i), \quad (1.11)$$

where the index i corresponds to i different values of the parameters θ . Obviously, the above equation can be written also as

$$Q(x) = \sum_{\theta} p(\theta) \mathcal{F}(x|\theta), \quad (1.12)$$

where the sum over the parameters is restricted to some set of interest and we substituted the index i on p by a notational dependence on the value of θ . Using our convention that the sum over a variable can be understood as an integral if the variable assume continuous values, the same expression then can define a **continuous mixture** if we consider $p(\theta)$ a function of θ such that $p(\theta) \geq 0$.

The reader obviously noticed already where we are trying to arrive. The conditions on the coefficients of the mixtures obviously give them the properties of probability distributions and, in fact, we can interpret them exactly as that: the probability distributions of the parameters of the mixture. Note that even in the case of a non-parametric mixture, we can consider i as a kind of generalised parameter and p_i its distribution. Being so, the formula for the mixture acquires a form that we are already used to

$$Q(x) = \langle \mathcal{F}(x|\theta) \rangle_{\theta}, \quad (1.13)$$

and it turns out that in applications, finding the coefficients of the mixture turns out to be equivalent to find the distribution of θ . This is particularly powerful because now we can use Bayes' Rule to infer the posterior probability of θ given some data! But again, this is best explained in other texts like [1, 2].

Chapter 2

Dirac Deltas

2.1 Basic Definitions

The **Dirac delta** was introduced by Dirac to represent point mass or charge distributions in a continuous way. The physicist's definition is

$$\delta(x - x_0) = \begin{cases} \infty, & x = x_0 \\ 0, & \text{otherwise,} \end{cases} \quad (2.1)$$

such that

$$\int_{-\infty}^{+\infty} dx \delta(x - x_0) = 1, \quad (2.2)$$

$$\int_{-\infty}^{+\infty} dx \delta(x - x_0) f(x) = f(x_0), \quad (2.3)$$

for every function $f(x)$, where x here is understood as a one-dimensional real variable. Actually, the integrals can be taken in any neighbourhood of x_0 with the same result. For questions of consistency however, if x_0 is one of the integration limits, then we have

$$\int_{x_0}^{+\infty} dx \delta(x - x_0) = \int_{-\infty}^{x_0} dx \delta(x - x_0) = 1/2. \quad (2.4)$$

It is common to say that $\delta(x - x_0)$ is a Dirac delta centred at x_0 . Of course the Dirac delta cannot be a usual function, although it is pretty common to see references to it as the *Dirac delta function* or simply a *delta function*.

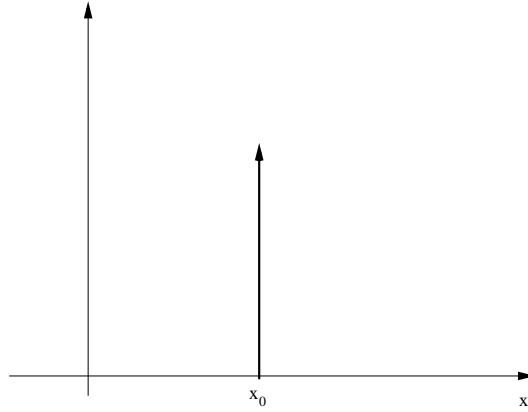


Figure 2.1: Graphical representation of a Dirac delta distribution centred at x_0 .

We can consider it as some kind of colloquial language and I will not worry too much about that. Rigorously, the Dirac delta is a distribution and can be defined as a limit of a sequence of functions (see for instance [3]). It also can be seen as the limit of a continuous probability distribution when its variance goes to zero. As the variance measures the spread around the mean value, if we make the variance of a probability distribution go to zero, it becomes a Dirac delta centred in its mean value. For instance, one of the most interesting and useful limits that gives a delta is of a Gaussian distribution

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right], \quad (2.5)$$

which leads to

$$\delta(x-\mu) = \lim_{\sigma^2 \rightarrow 0} \mathcal{N}(x|\mu, \sigma^2). \quad (2.6)$$

When I said that the Dirac delta is a distribution, you can understand it as a probability distribution, although as we have seen, one where the random variable is completely concentrated in one point. This obviously means that $\delta(x-x_0)$ is the probability distribution that encodes the fact that x is actually no random variable, it is exactly equal to x_0 always. This probability is sometimes graphically represented by a vertical arrow at x_0 like in figure 2.1.

The Dirac delta can be easily extended to multidimensional or many-variables. Consider the variable $\mathbf{x} = (x_1, x_2, \dots, x_n)$, with integer n . If we

want to express that it is exactly equal to $\mathbf{y} = (y_1, y_2, \dots, y_n)$ by means of a delta, this can be done by writing

$$\delta(\mathbf{x} - \mathbf{y}) = \prod_{i=1}^n \delta(x_i - y_i), \quad (2.7)$$

which is also written sometimes as

$$\delta^n(\mathbf{x} - \mathbf{y}), \quad (2.8)$$

to make explicit the fact that it is an n -dimensional delta.

Obviously, we can use the Dirac delta also to express the fact that $x = f(y)$ in the form

$$\mathcal{P}(x|y) = \delta(x - f(y)), \quad (2.9)$$

and here is where things begin to become interesting. In fact, we can use the Dirac delta to find the probability distribution of x simply by using a property that we have seen in the first chapter in equation (1.8), namely

$$\begin{aligned} \mathcal{P}(x) &= \langle \mathcal{P}(x|y) \rangle_y \\ &= \langle \delta(x - f(y)) \rangle_y. \end{aligned} \quad (2.10)$$

For example, suppose that we would like to calculate the probability of the sum of two random variables, say $x + y$. If we define a function $z(x, y) = x + y$, then we actually want the probability distribution of z . The usual way to solve this, that you can find in any probability book, is to go through calculating the cumulative distribution function and so on. However, the beauty of viewing the Dirac delta as a probability distribution is that we can, using the above formula, write the probability of z as

$$\mathcal{P}(z) = \langle \mathcal{P}(z|x, y) \rangle_{x,y}, \quad (2.11)$$

where we are averaging over the joint distribution of x and y . Now, if you are given x and y , z is already defined, there is no choice but it being $x + y$ and then we have

$$\mathcal{P}(z) = \langle \delta(z - (x + y)) \rangle_{x,y} \quad (2.12)$$

$$= \sum_{x,y} \delta(z - x - y) \mathcal{P}(x, y) \quad (2.13)$$

$$= \sum_x \mathcal{P}(x, z - x), \quad (2.14)$$

which, although looking pretty obvious, is very general in the sense that the joint distribution of x and y can be anything, even presenting correlations between both variables. In the particular case where x and y are independent, this is just the familiar **convolution** formula for the probability of the sum, with the convolution of two distributions $P(x)$ and $Q(x)$ given by

$$(P * Q)(x) = \int dy P(x - y)Q(y). \quad (2.15)$$

Let me repeat and add something. The power of the above derivation relies on its generality, but not only that. It relies also in the brevity of the calculation. As another example, let us see how the moments of the variable $y = bx$, where b is a constant, are related to the moments of x . The mean, for instance, becomes

$$\begin{aligned} \langle y \rangle_y &= \sum_y \mathcal{P}(y)y \\ &= \sum_y \left[\sum_x \mathcal{P}(y|x)\mathcal{P}(x) \right] y \\ &= \sum_x \mathcal{P}(x) \sum_y y \delta(y - bx) \\ &= b \langle x \rangle_x. \end{aligned} \quad (2.16)$$

And this can be extended to other moments giving the well-known results $\langle y^n \rangle_y = b^n \langle x^n \rangle_x$. Although in this case the results are pretty obvious, the method can be used in much more general and complicated situations.

2.2 Mathematical Properties

In order to be able to do more interesting things with the Dirac delta, we need first take a look at some of its mathematical properties which are going to be very useful in our future calculations. An important property is

$$\delta(f(x)) = \sum_i \frac{\delta(x - x_i)}{|f'(x_i)|}, \quad (2.17)$$

where the x_i 's are the roots of the equation $f(x) = 0$. From this property follows some of the well known simpler properties like the two below

1.

$$\delta(\alpha x) = \frac{\delta(x)}{|\alpha|}, \quad (2.18)$$

2.

$$\delta(x^2 - \alpha^2) = \frac{1}{2|\alpha|} [\delta(x + \alpha) + \delta(x - \alpha)]. \quad (2.19)$$

It is usually very convenient to use the so-called integral representation of the delta, which is really its Fourier transform given in its n -dimensional form as

$$\delta(\mathbf{x} - \mathbf{y}) = \frac{1}{(2\pi)^n} \int_{-\infty}^{+\infty} d\mathbf{k} e^{i\mathbf{k}\cdot(\mathbf{x}-\mathbf{y})}. \quad (2.20)$$

It is also possible to define the n -th derivatives of the Dirac delta by

$$\int_{-\infty}^{+\infty} dx \left[\frac{d^n}{dx^n} \delta(x - x_0) \right] f(x) = (-1)^n \left[\frac{d^n f(x)}{dx^n} \right]_{x=x_0}, \quad (2.21)$$

which is written in one dimension.

The Dirac delta can also be seen as the derivative of another distribution called the **Heaviside step function** $\theta(x)$, defined by

$$\theta(x) = \begin{cases} 1, & x > 0 \\ 1/2, & x = 0 \\ 0, & x < 0 \end{cases} \quad (2.22)$$

which, put in symbols just to make it look nice, is

$$\delta(x) = \frac{d}{dx} \theta(x). \quad (2.23)$$

2.3 Jacobians

Another interesting property related to equation (2.10) is that we can use Dirac deltas to change variables under integrals. Let us say that we want to calculate the integral

$$\int dx f(g(x)). \quad (2.24)$$

Usually we would change variables from x to $y = g(x)$ to try to simplify the integral. Doing this requires a change in the domain of integration and

the appearance of a **Jacobian**. But let us see how to do that using a delta function

$$\begin{aligned}\int dx f(g(x)) &= \int dx \int dy \delta(y - g(x)) f(y) \\ &= \int dy \left[\int dx \delta(y - g(x)) \right] f(y),\end{aligned}\tag{2.25}$$

and it is easy to see that the term between square brackets is a combination of the Jacobian *plus* a change in the integration domain. Although it looks only a formal expression, it turns out that many times it is possible to really calculate this term by using the integral representation of the delta and doing the integration on x ! I said that it is completely related to the average rule, given by equation (2.10), because you can write the intuitively clear expression

$$\langle f(g(x)) \rangle_x = \langle f(y) \rangle_y,\tag{2.26}$$

and then open it as

$$\begin{aligned}\int dx \mathcal{P}(x) f(g(x)) &= \int dy \mathcal{P}(y) f(y) \\ &= \int dy \langle \delta(y - g(x)) \rangle_x f(y) \\ &= \int dy \left[\int dx \delta(y - g(x)) \mathcal{P}(x) \right] f(y).\end{aligned}\tag{2.27}$$

Now, if we pick $\mathcal{P}(x) = 1/V$, where V is a finite volume where x is defined, we get back the change of variables formula. Of course you can go the other way round and use the change of variables to find the equality of the averages or the formula for the probability of y in terms of the probability of x . It works either way.

Chapter 3

Entropy

For every probability distribution $P(x)$ we can define an **entropy** functional given by

$$S[P] = - \sum_x P(x) \ln P(x). \quad (3.1)$$

Note that I said that entropy is a **functional**, not a function, as it takes as its argument some function $P(x)$ and gives back a number. In statistical physics entropy is interpreted as a measure of disorder. The classic work of Shannon [4, 5] which gave birth to information theory showed that the entropy of a probability distribution can be understood as the average lack of information or the average surprise when we sample from the distribution. The association of entropy with information is a powerful one with many ramifications not only in information theory but also in physics and other sciences. In inference, this association gave origin to the **maximum entropy** principle, which we will analyse in more detail in the next section.

Entropy has a series of nice mathematical properties, like concavity and positivity. It also has a lot of subtleties in which we will not spend our time for a whole books had already been written about it. The information theoretic view on entropy can be studied in the standard information theory reference [6]. For the physics point of view, the most basic text is probably Reif's book [7].

3.1 Maximum Entropy

The method of maximum entropy, among other things, is mainly used to choose a prior distribution for some random variable. There is a very complete account of it in Jaynes' book [2], so I will pass over all the details here and give just an overview. Suppose we have some information about a probability distribution in the form of averages of functions and its range, but we know nothing else about it. What would be the best prior to choose? Heuristically, the best thing would be to find a probability distribution defined within the specific range that obeys all the constraints given by the averages, but would otherwise convey us with as least information as possible. This is a kind of Occam's razor. We want to stick to the data and assume nothing else to construct our prior distribution.

Knowing that entropy can be associated to lack of information in a probability distribution and that we want our inferred distribution to convey as least information as possible beyond the data, we could require that the entropy of this function, the lack of information when sampling from it, is a maximum. In order to do that, it is necessary to maximise $S[P]$ under the constraints given by the data by adding the averages with **Lagrange multipliers**. Let us put it into equations. Suppose the constraints are given as N averages in the form

$$\langle f_i(x) \rangle_x = \mu_i, \quad i = 1, \dots, N. \quad (3.2)$$

Then, we would write our Lagrangean as

$$\mathcal{L}[P(x)] = S[P(x)] + \lambda \left(\int dx P(x) - 1 \right) + \sum_i \lambda_i \left(\int dx P(x) f_i(x) - \mu_i \right), \quad (3.3)$$

and, as we are looking for the functional form of $P(x)$, require that the *functional* derivative with respect to $P(x)$ be zero

$$\frac{\delta \mathcal{L}[P(x)]}{\delta P(y)} = 0. \quad (3.4)$$

A not very rigorous definition of the functional derivative is given in terms of the Dirac delta as

$$\frac{\delta F[f(x)]}{\delta f(y)} = \lim_{\epsilon \rightarrow 0} \frac{1}{\epsilon} \{ F[f(x) + \epsilon \delta(x - y)] - F[x] \}. \quad (3.5)$$

This definition guarantees that the functional derivative follows the same usual rules of derivation as the normal derivatives. Using this expression to calculate the functional derivative of the Lagrangean function (3.3) yields as a solution the probability distribution

$$P(x) = \frac{1}{Z} \exp \left[- \sum_i \lambda_i f_i(x) \right], \quad (3.6)$$

with

$$Z = e^\lambda = \int dx \exp \left[- \sum_i \lambda_i f_i(x) \right], \quad (3.7)$$

the normalisation, which we are symbolising by Z remembering that in physics it corresponds to the **partition function**, as we have already seen.

It turns out that depending on which function averages we constrain in equations (3.2) we end up with a different parametric family defined by the parameters μ_i . We will see examples of this in the following chapters.

3.2 Kullback-Leibler Projection

There are many situations where it is convenient to approximate a function by a simpler one sacrificing some precision in the process. The same statement is true for probability distributions in some situations. There are many parametric families that have mathematical properties that make analytical calculations easier when dealing with them, the most eminent case being the Gaussian distributions to be studied in the next chapter. But what would be the best way to do that? As always, the answer depends on how we look at the problem. Usually, we need to define some kind of distance between probabilities distribution that we would like to minimise in our approximation. There are many different distance measures between probability distributions, but there is one which has some appeal when seen from the information theory point of view. This one is called the **Kullback-Leibler divergence** and it is defined as

$$D(P||Q) = \int dx P(x) \ln \frac{P(x)}{Q(x)}, \quad (3.8)$$

and is also known as the **cross-entropy** between the two distributions. The standard reference about Kullback-Leibler divergence is Kullback's own book [8], although it is not a basic or introductory book.

Once we have agreed about the distance measure, we want to make this distance as small as possible between our original distribution and the approximate one. There is a small detail to pay attention here. The Kullback-Leibler (KL) divergence is not symmetric, which means that the distance from P to Q is not the same as the distance from Q to P . We shall choose which one we want to minimise. Here we choose the approximation in such a way that in the above formula $P(x)$ stands for the exact distribution while $Q(x)$ stands for the approximate one. The rough justification is that we want the average of the logarithm to be taken in the exact distribution rather than in the approximate one.

The procedure now becomes very similar to the maximum entropy method. Indeed, we will extremise the KL divergence with respect to $Q(x)$ by taking its functional derivative and add constraints that define the parametric family to which $Q(x)$ belongs. Remembering that $Q(x)$ will be then completely defined by the values of its parameters μ_i , the beautiful and simple solution to the approximation problem in this case becomes

$$Q(x) = \mathcal{F}(x|\mu_1, \dots, \mu_N), \quad (3.9)$$

with

$$\mu_i = \langle f_i(x) \rangle_{P(x)}, \quad (3.10)$$

where we are now twisting our notation a little bit to indicate that the average will be taken in the original probability distribution.

Chapter 4

Gaussians

Gaussian distributions are the most used probability distributions. As we are going to see, they have very nice properties and are easy to handle analytically. They are defined completely by a mean vector, which we will also see as a column matrix, μ and a covariance matrix C as

$$\mathcal{N}(\mathbf{x}|\mu, C) = \frac{1}{\sqrt{(2\pi)^N |C|}} \exp \left[-\frac{1}{2}(\mathbf{x} - \mu)^T C^{-1}(\mathbf{x} - \mu) \right], \quad (4.1)$$

which is sometimes indicated in an abbreviated way as $\mathbf{x} \sim \mathcal{N}(\mu, C)$, where \mathcal{N} stands for **normal distribution**, which is an alternative name for the Gaussian.

One interesting fact about Gaussians is that they can be somehow derived from the maximum entropy principle as the parametric family that maximises the entropy with the only constraints being its mean value and its covariance. In other words, if everything you know about a probability distribution is its mean value and its covariance, its range also being completely unknown, the best guess according to the minimum entropy principle is a Gaussian!

Accordingly, as we have seen in the last chapter, if we want to approximate a distribution by a Gaussian minimising the KL divergence, all we need to do is to match the mean and the covariance of this Gaussian to the ones of the original distributions. Approximating distributions by Gaussians may be very convenient as Gaussians have the advantage of being integrable and, as we shall see in the following, its moments are easily calculated. In physics, for instance, a Gaussian approximation for a Lagrangean of a physical system is what is called the small oscillations approximation for a potential which is also equivalent to expand the potential to second order in its Taylor series.

Another interesting property of Gaussians is their closure under convolutions

$$\mathcal{N}(\mu_1, C_1) * \mathcal{N}(\mu_2, C_2) = \mathcal{N}(\mu_1 + \mu_2, C_1 + C_2). \quad (4.2)$$

Now, if we consider $\mathcal{N}(0, 0) = \delta(\mathbf{x})$, the Gaussians form an Abelian monoid under convolution, i.e., they form a semi-group with identity. It is a semi-group for the positive-definiteness of the covariance matrices does not allow for inverses. However, apart from being an interesting property, the monoid structure does not seem to have any application until now.

4.1 Moments

The moments of a Gaussian distribution can be analytically calculated using two very simple methods. The first one is algebraic, while the second one is graphical. Both of them are simple instances of similar methods used in Quantum Field Theory (QFT), namely **Feynmann diagrams**[9]. Let us describe the algebraic method first. Consider an N -dimensional vector $\mathbf{x} \sim \mathcal{N}(\mu, C)$. Suppose we want to calculate the moment, also called the **correlation function**,

$$\left\langle x_1^{k_1} x_2^{k_2} \cdots x_N^{k_N} \right\rangle_{\mathbf{x}}, \quad (4.3)$$

where the k_i 's are integers between 0 and N . The zero value for any k_i is just a way of saying that the corresponding variable does not appear in the correlation to be calculated. In order to do this, we will define a function $Z(\mathbf{J})$ given by

$$Z(\mathbf{J}) = \int d\mathbf{x} \exp \left[-\frac{1}{2}(\mathbf{x} - \mu)^T C^{-1}(\mathbf{x} - \mu) + \mathbf{J} \cdot \mathbf{x} \right], \quad (4.4)$$

where the components of \mathbf{J} are known in QFT as sources. This expression is easy to calculate and gives

$$Z(\mathbf{J}) = \sqrt{(2\pi)^N |C|} \exp \left(\frac{1}{2} \mathbf{J}^T C \mathbf{J} + \mathbf{J} \cdot \mu \right). \quad (4.5)$$

This expression is clearly equal to the normalisation of the Gaussian for the limit when the sources disappear, $\mathbf{J} = \mathbf{0}$. Now, the correlation we want

to calculate can be written as

$$\begin{aligned} \left\langle x_1^{k_1} x_2^{k_2} \cdots x_N^{k_N} \right\rangle_{\mathbf{x}} &= \frac{1}{Z(\mathbf{0})} \lim_{\mathbf{J} \rightarrow \mathbf{0}} \int d\mathbf{x} x_1^{k_1} x_2^{k_2} \cdots x_N^{k_N} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T C^{-1}(\mathbf{x}-\boldsymbol{\mu}) + \mathbf{J} \cdot \mathbf{x}} \\ &= \frac{1}{Z(\mathbf{0})} \lim_{\mathbf{J} \rightarrow \mathbf{0}} \frac{\partial^{k_1}}{\partial J_1^{k_1}} \cdots \frac{\partial^{k_N}}{\partial J_N^{k_N}} Z(\mathbf{J}). \end{aligned} \quad (4.6)$$

The nice feature is that the result depends only on the means μ_i 's and the covariance matrix C . This should be expected once we have already seen that a Gaussian is completely specified by the first two moments plus the maximum entropy principle.

The above result can also be cast in a graphical way. Those familiar with QFT can already see it. The method is very simple. The expression for the correlation becomes a summation of all graphs constructed in the following way. We associate each power of any variable x_i with one vertex labelled with the number i . Then we construct all possible graphs containing these vertices with zero edges, then with one edge and so on with the restriction that there is at most one edge attached to each vertex. The term encoded by each graph is simply a multiplication of μ_i for each isolated vertex with label i in the graph times C_{jk} for each edge connecting vertices labelled by j and k . For example, the correlation

$$\left\langle x_1^2 x_3 \right\rangle_{\mathbf{x}} = \mu_1^2 \mu_3 + 2\mu_1 C_{13} + \mu_3 C_{11}, \quad (4.7)$$

is depicted in its graphical representation in figure 4.1.

And as the theorems are the same here as in QFT up to a multiplicative constant as the expert reader already noticed, the following formula

$$\left\langle x_1^{k_1} x_2^{k_2} \cdots x_N^{k_N} \right\rangle_{\mathbf{x}}^c = \lim_{\mathbf{J} \rightarrow \mathbf{0}} \frac{\partial^{k_1}}{\partial J_1^{k_1}} \cdots \frac{\partial^{k_N}}{\partial J_N^{k_N}} \ln Z(\mathbf{J}), \quad (4.8)$$

is called the **connected correlation function**, for it can be calculated by summing only the corresponding **connected diagrams** appearing in the full correlation function, which means that its result is equal to the sum of diagrams where there are no isolated vertices.

These techniques are very powerful, using them it is possible to approximate by a series of diagrams the average of any function of \mathbf{x} by expanding it in a Taylor series, each power of any coordinate can simply be substituted by

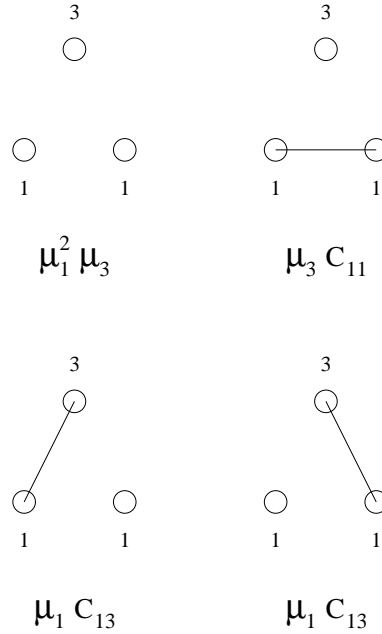


Figure 4.1: Graphical calculation of the moment $\langle x_1^2 x_3 \rangle_{\mathbf{x}}$.

its derivative giving the required result. For example, for the one-dimensional case we would have, expanding the function around zero,

$$\langle f(x) \rangle_x = \sum_n \frac{1}{n!} \left. \frac{d^n f(x)}{dx^n} \right|_{x=0} \langle x^n \rangle_x = \sum_n \frac{1}{n!} \left. \frac{d^n f(x)}{dx^n} \right|_{x=0} \frac{1}{Z(0)} \lim_{J \rightarrow 0} \frac{d^n}{dJ^n} Z(J), \quad (4.9)$$

which becomes an infinite series of diagrams that can be stopped at some convenient point.

4.2 Limit Theorems

Let us look at some aspects of limits of sums of random variables, which we will call loosely by the name of **limit theorems**. The **central limit theorem** states that the sum of N random variables x_1, \dots, x_N i.i.d., with mean μ and variance σ^2 each, in the limit of large N becomes a Gaussian distribution $\mathcal{N}(N\mu, N\sigma^2)$. This theorem can be generalised in many directions. We will give one of them here and present a highly non-rigorous proof

using the techniques we have learnt until now. The same procedure can be used to prove similar instances of the theorem

Consider, for instance, N random variables x_1, \dots, x_N independently but possibly not indentially distributed. Let us define the linear combination

$$S_N = \frac{1}{N} \sum_{i=1}^N \alpha_i x_i, \quad (4.10)$$

where the α_i 's are real coefficients. Its probability distribution is

$$\begin{aligned} \mathcal{P}(S_N) &= \left\langle \delta \left(S_N - \frac{1}{N} \sum_{i=1}^N \alpha_i x_i \right) \right\rangle_{x_1, \dots, x_N} \\ &= \sum_{x_1, \dots, x_N} \left[\prod_i \mathcal{P}(x_i) \right] \int \frac{dk}{2\pi} \exp \left[ik \left(S_N - \frac{1}{N} \sum_{i=1}^N \alpha_i x_i \right) \right] \\ &= \int \frac{dk}{2\pi} e^{ikS_N} \prod_i \sum_{x_i} \mathcal{P}(x_i) \exp \left(-\frac{ik\alpha_i x_i}{N} \right). \end{aligned} \quad (4.11)$$

Expanding the exponential to second order in N we have

$$\begin{aligned} \mathcal{P}(S_N) &= \int \frac{dk}{2\pi} e^{ikS_N} \prod_i \left(1 - \frac{ik\alpha_i \langle x_i \rangle}{N} - \frac{\alpha_i^2 k^2}{2N^2} \langle x_i^2 \rangle \right) \\ &= \int \frac{dk}{2\pi} e^{ikS_N} \exp \left[\sum_i \ln \left(1 - \frac{ik\alpha_i \langle x_i \rangle}{N} - \frac{\alpha_i^2 k^2}{2N^2} \langle x_i^2 \rangle \right) \right]. \end{aligned} \quad (4.12)$$

Using the Taylor expansion of the logarithm

$$\ln(1+x) = x - \frac{1}{2}x^2 + \frac{1}{3}x^3 - \dots, \quad (4.13)$$

up to second order, we then have

$$\begin{aligned} \mathcal{P}(S_N) &= \int \frac{dk}{2\pi} e^{ikS_N} \prod_i \left(1 - \frac{ik\alpha_i \langle x_i \rangle}{N} - \frac{\alpha_i^2 k^2}{2N^2} \langle x_i^2 \rangle \right) \\ &= \int \frac{dk}{2\pi} e^{ikS_N} \exp \left[-\frac{ik}{N} \sum_i \alpha_i \langle x_i \rangle - \frac{k^2}{2N^2} \sum_i \alpha_i^2 \langle x_i^2 \rangle \right. \\ &\quad \left. - \frac{1}{2} \sum_i \left(\frac{ik\alpha_i}{N} \langle x_i \rangle + \frac{\alpha_i^2 k^2}{2N^2} \langle x_i^2 \rangle \right)^2 \right], \end{aligned} \quad (4.14)$$

which with the definitions

$$\mu = \frac{1}{N} \sum_i \alpha_i \langle x_i \rangle, \quad (4.15)$$

$$\sigma^2 = \frac{1}{N^2} \sum_i \alpha_i^2 (\langle x_i^2 \rangle - \langle x_i \rangle^2), \quad (4.16)$$

and by dropping higher powers of N becomes

$$\begin{aligned} \mathcal{P}(S_N) &= \int \frac{dk}{2\pi} e^{-\frac{k^2 \sigma^2}{2} + ik(S_N - \mu)} \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[-\frac{(S_N - \mu)^2}{2\sigma^2} \right], \end{aligned} \quad (4.17)$$

which is a Gaussian $\mathcal{N}(\mu, \sigma^2)$. For $\alpha_i = 1$ and i.i.d. variables, we recover the usual central limit theorem. Of course there are loads of cheats in the above derivation, but the point is that if you are careful enough, this calculation can be carried out in several situations to give a quick answer to limits that appear in many applications.

Chapter 5

Gammas and Betas

Our next parametric families, not only those on this chapter but also the one in the next, can be considered as generalisations of the **binomial distribution**

$$b(k|n, p) = \binom{n}{k} p^k (1-p)^{n-k}, \quad (5.1)$$

where $0 \leq p \leq 1$ and k assume an integer value between zero and n . This is basically the probability of k heads and $n - k$ tails in n biased coin tossings where the probability of heads is p . The mean is np and the variance is $np(1-p)$.

This distribution will be our starting point to introduce the Beta and Gamma distributions in the next sections.

5.1 The Beta Distribution

Let us now look at the binomial distribution from another point of view. Instead of considering a distribution of k , let us view it as a distribution of p . As this variable is a probability, we have that $p \in [0, 1]$ and it is a real number. Let us generalise more. Consider now that n and k can be real numbers. Actually, let us define the real numbers $\alpha = k + 1$ and $\beta = n - k + 1$. It turns out that all we have to do to guarantee that this distribution is normalised is to use, instead of the binomial, its analytical continuation given by the one over the beta function

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)}, \quad (5.2)$$

and the resulting distribution is called the **Beta distribution** and is defined by

$$\mathcal{B}(x|\alpha, \beta) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1}, \quad (5.3)$$

where $0 \leq x \leq 1$. Its mean and variance are given by

$$\langle x \rangle = \frac{\alpha}{\alpha + \beta}, \quad (5.4)$$

$$\text{Var}(x) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}. \quad (5.5)$$

5.2 The Gamma Distribution

Now comes the most interesting story of this chapter. Given that a variable x is Beta distributed, what would be the distribution of $y = bx$, where $b \geq 0$? We know how to calculate this using deltas

$$\begin{aligned} \mathcal{P}(y) &= \langle \delta(y - bx) \rangle_x \\ &= \left\langle \frac{1}{b} \delta\left(x - \frac{y}{b}\right) \right\rangle_x \\ &= \frac{b^{-1}}{B(\alpha, \beta)} \left(\frac{y}{b}\right)^{\alpha-1} \left(1 - \frac{y}{b}\right)^{\beta-1}, \end{aligned} \quad (5.6)$$

where now the range of y is $[0, b]$. We will be interested in the limit when $b \rightarrow \infty$, which means that the support of our distribution will be the whole non-negative real axis. As we have already seen, the mean of this variable will be given by

$$\langle y \rangle_y = b \langle x \rangle_x = \frac{b\alpha}{\alpha + \beta}. \quad (5.7)$$

However, we have a scaling problem here. If we take the desired limit, this mean will blow up, something that is not very interesting. The best way to keep this average finite is then to scale β with b . Actually, let us take the exponent of $(1 - y/b)$ to be $b\beta - 1$ and keep α constant, without scaling with b . Then, in the required limit, the mean becomes just α/β and remains finite and the distribution can be written as

$$\mathcal{G}(y|\alpha, \beta) = \lim_{b \rightarrow \infty} \frac{b^{-1}}{B(\alpha, b\beta)} \left(\frac{y}{b}\right)^{\alpha-1} \left(1 - \frac{y}{b}\right)^{b\beta-1}. \quad (5.8)$$

For further reference, note that the variance also remains finite and becomes α/β^2 . To find the above distribution we need to calculate the limit. The first piece of the puzzle is

$$\left(1 - \frac{y}{b}\right)^{b\beta-1} \rightarrow e^{-\beta y}. \quad (5.9)$$

The second and last piece must be handled with care in order to not throw away any significant term. It is the term that arises from all the b dependencies and is given in the limit of large b by

$$\begin{aligned} \frac{\Gamma(\alpha + b\beta)}{\Gamma(b\beta)} \frac{1}{b^\alpha} &= \frac{\sqrt{\frac{2\pi}{\alpha+b\beta}} \left(\frac{\alpha+b\beta}{e}\right)^{\alpha+b\beta}}{\sqrt{\frac{2\pi}{b\beta}} \left(\frac{b\beta}{e}\right)^{b\beta}} \frac{1}{b^\alpha} \\ &= \frac{(\alpha + b\beta)^{\alpha+b\beta}}{(b\beta)^{b\beta}} \frac{1}{e^\alpha b^\alpha} \\ &= \beta^\alpha \left(1 + \frac{\alpha}{b\beta}\right)^{\alpha+b\beta} \frac{1}{e^\alpha}. \end{aligned} \quad (5.10)$$

The limit in parenthesis is easily recognisable as e^α when $b \rightarrow \infty$ and where the gamma functions were approximated using Stirling's formula. Putting everything together we have

$$\mathcal{G}(y|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\beta y}, \quad (5.11)$$

which is called the **Gamma distribution**. In some texts, β is called the **rate parameter** and α the **shape parameter**.

The sum of two independent Gamma distributed variables with the same rate parameter β and shape parameters α_1 and α_2 is again a Gamma with the same rate parameter but with shape parameter $\alpha_1 + \alpha_2$. Which means that a family of Gamma distributions with the same rate parameter is also a monoid under convolution with the identity given by the Dirac delta centered at zero. This is so because in the limit of a shape parameter being zero, the variance and the mean also goes to zero, characterising the delta. Note that both the mean and the variance of the resulting convolution are the sum of the mean and variance of the two individual Gammas. There is a nice way of proving this using Dirac deltas that goes this way. We want the probability distribution of $s = x_1 + x_2$ with x_1 and x_2 independently Gamma

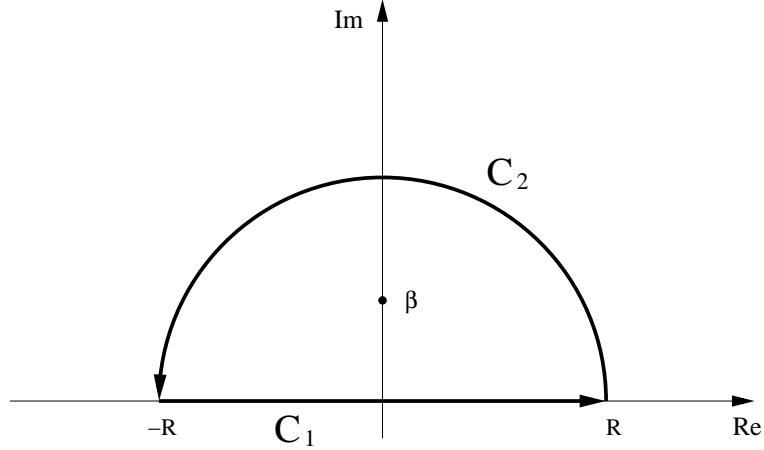


Figure 5.1: Integration path in the complex plane.

distributed with rate parameters β and shape parameters respectively α_1 and α_2 . Analogous to the case of the central limit theorem, equation (4.11), this can be written as

$$\begin{aligned} \mathcal{P}(s) &= \langle \delta(s - x_1 - x_2) \rangle_{x_1, x_2} \\ &= \frac{\beta^{\alpha_1 + \alpha_2}}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \int \frac{dk}{2\pi} e^{iks} \prod_{i=1,2} \int dx_i e^{-(ikx_i + \beta)x_i} x_i^{\alpha_i}. \end{aligned} \quad (5.12)$$

The integral in x_i can be obtained in any integral table (for instance, [10]). Plugging in the result we have

$$\begin{aligned} \mathcal{P}(s) &= \frac{\beta^{\alpha_1 + \alpha_2}}{2\pi} \int dk \frac{e^{iks}}{(ik + \beta)^{\alpha_1 + \alpha_2}} \\ &= \frac{\beta^{\alpha_1 + \alpha_2}}{2\pi i^{\alpha_1 + \alpha_2}} \int dk \frac{e^{iks}}{(k - i\beta)^{\alpha_1 + \alpha_2}}. \end{aligned} \quad (5.13)$$

The integral over k can easily be solved using **residues**[11]. If we close the path of integration with a semicircle in the upper half of the complex plane as in figure 5.1, the integral over the closed path $C_1 + C_2$ can be written as

$$I_R = \int_{-R}^R dk \frac{e^{iks}}{(k - i\beta)^{\alpha_1 + \alpha_2}} + I(C_2), \quad (5.14)$$

where $I(C_2)$ is the part of the integral that goes over the path C_2 . It can be shown that in the limit of $R \rightarrow \infty$, $I(C_2)$ becomes zero. The complete

integral has a pole on $i\beta$ and, although the order of the pole may not be an integer, it turns out that we can use an analytical continuation and pretend it is to carry on the calculation. The integral can then be calculated as

$$\begin{aligned}\mathcal{P}(s) &= \frac{\beta^{\alpha_1+\alpha_2}}{2\pi i^{\alpha_1+\alpha_2}} \lim_{R \rightarrow \infty} I_R \\ &= \frac{\beta^{\alpha_1+\alpha_2}}{2\pi i^{\alpha_1+\alpha_2}} 2\pi i \operatorname{Res}\left(\frac{e^{iks}}{(k-i\beta)^{\alpha_1+\alpha_2}}, i\beta\right),\end{aligned}\tag{5.15}$$

where, as we said, we will analytically continue the usual residue formula to real values by substituting the factorials in the differentiation formula by gamma functions

$$\begin{aligned}\operatorname{Res}\left(\frac{e^{iks}}{(k-i\beta)^{\alpha_1+\alpha_2}}, i\beta\right) &= \frac{1}{\Gamma(\alpha_1+\alpha_2)} \lim_{k \rightarrow i\beta} \frac{d^{\alpha_1+\alpha_2-1}}{dk^{\alpha_1+\alpha_2-1}} e^{iks} \\ &= \frac{(is)^{\alpha_1+\alpha_2-1}}{\Gamma(\alpha_1+\alpha_2)} e^{-\beta s},\end{aligned}\tag{5.16}$$

which when substituted in the formula for the distribution gives simply

$$\mathcal{P}(s) = \mathcal{G}(s|\alpha_1+\alpha_2, \beta).\tag{5.17}$$

Chapter 6

Dirichlets

Another generalisation of the binomial distribution is known by the name of **Dirichlet distribution**. In fact, it can be seen more as a direct generalisation of the Beta distribution. It is a multivariate distribution for an n -dimensional vector $\mathbf{x} = (x_1, x_2, \dots, x_n)$ which obeys the constraints

$$0 \leq x_i \leq 1, \quad (6.1)$$

$$\sum_{i=1}^n x_i = 1, \quad (6.2)$$

and is given by

$$\mathcal{D}(\mathbf{x}|\mathbf{u}) = \frac{1}{Z(\mathbf{u})} \prod_{i=1}^n x_i^{u_i-1}, \quad (6.3)$$

with $u_i > 0$ and normalisation

$$Z(\mathbf{u}) = \frac{\prod_{i=1}^n \Gamma(u_i)}{\Gamma(u_0)}, \quad (6.4)$$

where

$$u_0 = \sum_{i=1}^n u_i. \quad (6.5)$$

Again we have a distribution where the variables can be considered as probabilities. This means that the Dirichlet, and in the particular univariate case the Beta, are distributions of discrete distributions. This is a useful property as, for example, we can use them as priors for coefficients of probability mixtures, once we have already seen in section 1.2 that they can be seen exactly as probability distributions.

6.1 Normalisation

Although you will not find it in many books, obtaining the normalisation of the Dirichlet is not so trivial. In the following, I will show two methods to do it.

6.1.1 Method One: Complex Integration

Using a Dirac delta to enforce the constraints (6.1) we have

$$\begin{aligned} Z(\mathbf{u}) &= \int d\mathbf{x} \delta\left(\sum_i x_i - 1\right) \prod_i \theta(x_i) x_i^{u_i-1} \\ &= \int \frac{dk}{2\pi} e^{ik} \prod_i \int dx_i \theta(x_i) e^{-ikx_i} x_i^{u_i-1} \\ &= \int \frac{dk}{2\pi} e^{ik} \prod_i I(u_i, k), \end{aligned} \quad (6.6)$$

where $\theta(x)$ is the Heaviside step function and

$$I(u, k) = \int dx \theta(x) e^{-ikx} x^{u-1} = \int_0^\infty dx e^{-ikx} x^{u-1}. \quad (6.7)$$

Changing variables to $y = ikx$ we have

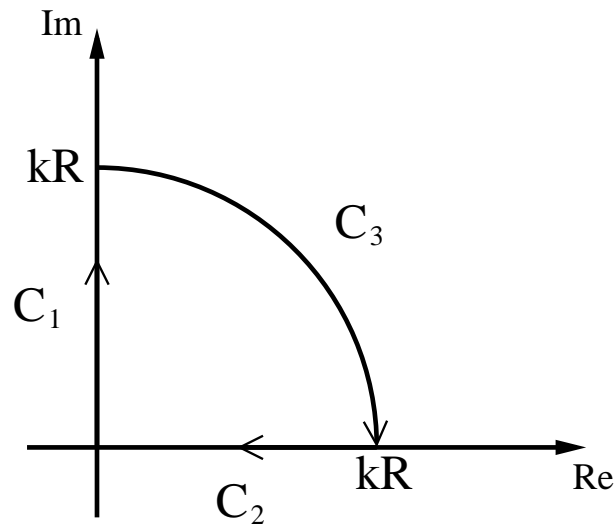
$$I(u, k) = \frac{1}{(ik)^u} \int_0^{ik\infty} dy e^{-y} y^{u-1}. \quad (6.8)$$

Let us analyse separately the cases $k > 0$ and $k < 0$. For $k > 0$, we use the integration path of figure 6.1 which is composed of three parts: C_1 , C_2 and C_3 . The integrand is

$$f(z) = e^{-z} z^{u-1}, \quad (6.9)$$

which does not have any poles inside the closed path $C = C_1 + C_2 + C_3$ and therefore

$$\int_C f(z) dz = \int_{C_1} f(z) dz + \int_{C_2} f(z) dz + \int_{C_3} f(z) dz = 0. \quad (6.10)$$

Figure 6.1: Integration path in the complex plane for $k > 0$.

For $R \rightarrow \infty$, the integral in C_3 goes to zero. The integral in C_2 is on \mathbb{R} and is just the integral representation of the Gamma function with a minus sign due to the path orientation. Therefore

$$I(u, k) = \frac{1}{(ik)^u} \int_0^{ik\infty} dy e^{-y} y^{u-1} = \frac{\Gamma(u)}{(ik)^u}, \quad (6.11)$$

with $k > 0$. For $k < 0$ the analysis is the same but with the path of figure 6.2 and the result is equivalent. Substituting in $Z(\mathbf{u})$ we have

$$Z(\mathbf{u}) = \frac{\prod_{i=1}^n \Gamma(u_i)}{2\pi i^{u_0}} A(u_0), \quad (6.12)$$

where

$$A(x) \equiv \int_{-\infty}^{\infty} dk \frac{e^{ik}}{k^x}. \quad (6.13)$$

Using the fact that

$$\frac{1}{k^x} = \frac{-1}{x-1} \frac{d}{dk} \frac{1}{k^{x-1}}, \quad (6.14)$$

we can write

$$A(x) = \frac{-1}{x-1} \int_{-\infty}^{\infty} dk e^{ik} \frac{d}{dk} \frac{1}{k^{x-1}} = \frac{i}{x-1} \int_{-\infty}^{\infty} dk e^{ik} \frac{1}{k^{x-1}}, \quad (6.15)$$

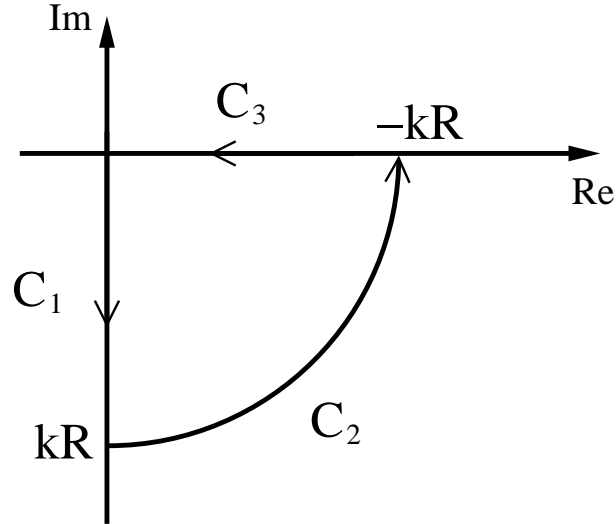


Figure 6.2: Integration path in the complex plane for $k < 0$.

where we used integration by parts in the last equality. Then we see that $A(x)$ satisfies the recurrence relationship

$$A(x) = \frac{i}{x-1} A(x-1), \quad (6.16)$$

and therefore we can identify

$$A(x) = \frac{c i^x}{\Gamma(x)}. \quad (6.17)$$

The value of the constant c is obtained by $A(1) = ic$. The integral in $A(1)$ is trivial and equals $2\pi i$. The final result is

$$A(x) = \frac{2\pi i^x}{\Gamma(x)}. \quad (6.18)$$

Substituting in $Z(\mathbf{u})$ again

$$Z(\mathbf{u}) = \frac{\prod_{i=1}^n \Gamma(u_i)}{\Gamma(u_0)}. \quad (6.19)$$

6.1.2 Method Two: Analytical Continuation

There is another, dirtier method to obtain the normalisation of the Dirichlet. Let us consider once again equations (6.6) and (6.7). We can analytically continue the derivative of a polynomial from integer to real powers by using the expression

$$(-1)^n \frac{d^n}{ds^n} \frac{1}{s} = \frac{n!}{s^{n+1}} \rightarrow (-1)^x \frac{d^x}{ds^x} \frac{1}{s} = \frac{\Gamma(x+1)}{s^{x+1}}, \quad (6.20)$$

which then we substitute in expression (6.7) and, using the definition for the derivatives of the Dirac delta given in chapter 2, we obtain the result

$$I(u) = \int \frac{ds}{i^u} \frac{1}{s} \frac{d^{u-1}}{dx^{u-1}} \delta(s-k) = \frac{1}{i^u} \frac{\Gamma(u)}{k^u}. \quad (6.21)$$

Plugging this expression back in the normalisation we have

$$Z(\mathbf{u}) = \frac{\prod_i \Gamma(u_i)}{2\pi i^{u_0}} \int dk \frac{e^{ik}}{k^{u_0}}, \quad (6.22)$$

and using another analytical continuation, this time for the complex integral in the form

$$\int_{-\infty}^{+\infty} dk \frac{e^{ik}}{k^n} = \frac{2\pi i^n}{(n-1)!} \rightarrow \int_{-\infty}^{+\infty} dk \frac{e^{ik}}{k^x} = \frac{2\pi i^x}{\Gamma(x)}, \quad (6.23)$$

we finally get

$$Z(\mathbf{u}) = \frac{\prod_i \Gamma(u_i)}{2\pi i^{u_0}} \frac{2\pi i^{u_0}}{\Gamma(u_0)}, \quad (6.24)$$

which gives once more the expected result

$$Z(\mathbf{u}) = \frac{\prod_{i=1}^n \Gamma(u_i)}{\Gamma(u_0)}. \quad (6.25)$$

As with the Gaussian, the Dirichlet distribution can be obtained in a very particular way from the maximum entropy principle as the distribution that maximises the entropy [12, 13] under constraints that fix the range of the variables

$$0 \leq x_i \leq 1, \quad \sum_i x_i = 1, \quad (6.26)$$

and the value of the n averages

$$\langle \ln x_i \rangle_x, \quad i = 1, \dots, n. \quad (6.27)$$

Indeed, this can easily be seen by writing its defining equation as an exponential in the form

$$\mathcal{D}(\mathbf{x}|\mathbf{u}) = \frac{1}{Z(\mathbf{u})} e^{\sum_i (u_i - 1) \ln x_i}, \quad (6.28)$$

which can be compared to the general solution of the maximum entropy principle, equation (3.6). The interesting thing is that the averages of the logarithms that we fixed are averages of what in information theory is interpreted as the surprise $\ln x_i$ in sampling a value with probability x_i from the discrete distribution defined by the vector \mathbf{x} . This point is also unexplored in the literature.

6.2 Moments

It is interesting to see how to calculate the moments of the Dirichlet. Consider the general expression

$$\left\langle \prod_i x_i^{\alpha_i} \right\rangle = \int \frac{dx}{Z(\mathbf{u})} \delta \left(\sum_i x_i - 1 \right) \prod_i \theta(x_i) x_i^{u_i - 1 + \alpha_i}, \quad (6.29)$$

which can be calculated based on the integrals of the last section as

$$\left\langle \prod_i x_i^{\alpha_i} \right\rangle = \frac{\prod_i \Gamma(u_i + \alpha_i)}{\Gamma(u_0 + \alpha_0)} \frac{\Gamma(u_0)}{\prod_i \Gamma(u_i)}, \quad (6.30)$$

with $\alpha_0 = \sum_i \alpha_i$.

Therefore, we have the means

$$\begin{aligned} \langle x_i \rangle &= \left[\prod_{j \neq i} \Gamma(u_j) \right] \frac{\Gamma(u_i + 1)}{\Gamma(u_0 + 1)} \frac{\Gamma(u_0)}{\prod_j \Gamma(u_j)} \\ &= \frac{\Gamma(u_i + 1)}{\Gamma(u_i)} \frac{\Gamma(u_0)}{\Gamma(u_0 + 1)} \\ &= \frac{u_i}{u_0}, \end{aligned} \quad (6.31)$$

obviously satisfying

$$\sum_i \langle x_i \rangle = 1. \quad (6.32)$$

Using

$$\begin{aligned} \langle x_i^2 \rangle &= \frac{\Gamma(u_i + 2)}{\Gamma(u_i)} \frac{\Gamma(u_0)}{\Gamma(u_0 + 1)} \\ &= \frac{u_i(u_i + 1)}{u_0(u_0 + 1)}, \end{aligned} \quad (6.33)$$

it is easy to calculate the variances

$$\begin{aligned} \text{Var}(x_i) &= \langle x_i^2 \rangle - \langle x_i \rangle^2 \\ &= \frac{u_i(u_i + 1)}{u_0(u_0 + 1)} - \frac{u_i^2}{u_0^2} \\ &= \frac{u_i^2 u_0 + u_i u_0 - u_i^2 u_0 - u_i^2}{u_0^2(u_0 + 1)} \\ &= \frac{u_i(u_0 - u_i)}{u_0^2(u_0 + 1)}. \end{aligned} \quad (6.34)$$

Another useful average, which appears in some applications, is

$$l_i = \left\langle \left(\prod_j x_j^{\alpha_j} \right) \ln x_i \right\rangle, \quad (6.35)$$

which can be calculated using an analytical continuation very popular in

statistical physics by the name of **replica trick**

$$\begin{aligned}
l_i &= \left[\frac{\partial}{\partial n} \left\langle \left(\prod_{j \neq i} x_j^{\alpha_j} \right) x_i^{\alpha_i + n} \right\rangle \right]_{n=0} \\
&= \left\{ \frac{\partial}{\partial n} \left[\frac{\prod_{j \neq i} \Gamma(u_j + \alpha_j) \Gamma(u_i + \alpha_i + n)}{\Gamma(u_0 + \alpha_0 + n)} \frac{\Gamma(u_0)}{\prod_j \Gamma(u_j)} \right] \right\}_{n=0} \\
&= \frac{\Gamma(u_0)}{\prod_j \Gamma(u_j)} \prod_{j \neq i} \Gamma(u_j + \alpha_j) \left[\frac{\Gamma'(u_i + \alpha_i + n)}{\Gamma(u_i + \alpha_i + n)} - \frac{\Gamma(u_i + \alpha_i + n) \Gamma'(u_0 + \alpha_0 + n)}{\Gamma^2(u_0 + \alpha_0 + n)} \right]_{n=0} \\
&= \frac{\Gamma(u_0)}{\prod_j \Gamma(u_j)} \frac{\prod_j \Gamma(u_j + \alpha_j)}{\Gamma(u_0 + \alpha_0)} \left[\frac{\Gamma'(u_i + \alpha_i)}{\Gamma(u_i + \alpha_i)} - \frac{\Gamma'(u_0 + \alpha_0)}{\Gamma(u_0 + \alpha_0)} \right] \\
&= \left\langle \prod_j x_j^{\alpha_j} \right\rangle [\psi(u_i + \alpha_i) - \psi(u_0 + \alpha_0)].
\end{aligned} \tag{6.36}$$

The trick has this name because in physics it can be seen as analysing n replicas of the original physical system. The actual **replica method** used in applications on physics of disordered systems is much more complex and full of subtleties than the simple trick that was its origin. More information on it can be obtained in the standard reference which also has many reprints of the original key papers [14].

Note how the relation

$$\left\langle \left(\prod_j x_j^{\alpha_j} \right) \ln x_i \right\rangle = \left\langle \prod_j x_j^{\alpha_j} \right\rangle [\psi(u_i + \alpha_i) - \psi(u_0 + \alpha_0)], \tag{6.37}$$

looks like an eigenvalue equation. It can be actually interpreted as a kind of generalisation of it, but there is not much development of this idea in the literature.

Bibliography

- [1] Sivia, D. and Skilling, J. *Data Analysis: A Bayesian Tutorial*. Oxford University Press, USA, 2 edition, July (2006).
- [2] Jaynes, E. T. *Probability Theory : The Logic of Science*. Cambridge University Press, April (2003).
- [3] Arfken, G. B. and Weber, H. J. *Mathematical Methods For Physicists*. Academic Press, June (2005).
- [4] Shannon, C. E. *The Bell System Technical Journal* **27**, 379–423 July (1948).
- [5] Shannon, C. E. *The Bell System Technical Journal* **27**, 623–656 October (1948).
- [6] Cover, T. M. and Thomas, J. *Elements of Information Theory*. John Wiley & Sons, New York, NY, (1991).
- [7] Reif, F. *Fundamentals of Statistical and Thermal Physics (Fundamentals of Physics)*. McGraw-Hill Higher Education, January (1965).
- [8] Kullback, S. *Information Theory and Statistics (Dover Books on Mathematics)*. Dover Publications, July (1997).
- [9] Ryder, L. H. *Quantum Field Theory*. Cambridge University Press, 2 edition, June (1996).
- [10] Gradshteyn, I. S. and Ryzhik, I. M. *Table of Integrals, Series, and Products*. Academic Press, USA, (1993).
- [11] Brown, J. B. and Churchill, R. V. *Complex Variables and Applications*. McGraw-Hill Higher Education, 7 edition, June (2003).

- [12] Vlad, M. O., Tsuchiya, M., Oefner, P., and Ross, J. *Phys. Rev. E* **65**, 011112(1)–011112(8) (2001).
- [13] Alamino, R. C. and Caticha, N. *Discrete and Continuous Dynamical Systems - Series B (DCDS-B)* **9**(1), 1–10 January (2008).
- [14] Mézard, M., Parisi, G., and Virasoro, M. *Spin Glass Theory and Beyond*. World Scientific Publishing Co., Singapore, (1987).

Index

- Bayes' Rule, 3
- Bayes' Theorem, 3
- Beta distribution, 24
- binomial distribution, 23
- central limit theorem, 20
- chain rule, 2
- conditional probability, 2
- connected correlation function, 19
- connected diagrams, 19
- continuous mixtures, 5
- convex combination, 4
- convolution, 10
- correlation function, 18
- cross-entropy, 15
- Dirac delta, 7
- Dirichlet distribution, 29
- entropy, 13
- Feynmann diagrams, 18
- functional, 13
- Gamma distribution, 25
- Gaussian distribution, 17
- Heaviside step function, 11
- Jacobian, 12
- Kullback-Leibler divergence, 15
- Lagrange multipliers, 14
- likelihood, 3
- limit theorems, 20
- maximum entropy, 13
- mixture model, 4
- normal distribution, 17
- Occam's razor, 14
- parametric family, 4
- parametric mixture, 5
- partition function, 3, 15
- posterior distribution, 3
- prior distribution, 3
- probability mixture, 4
- probability operator, 1
- Quantum Field Theory, 18
- rate parameter, 25
- replica method, 36
- replica trick, 36
- residue, 26
- shape parameter, 25