

Measure and integration

Teo Banica

DEPARTMENT OF MATHEMATICS, UNIVERSITY OF CERGY-PONTOISE, F-95000
CERGY-PONTOISE, FRANCE. teo.banica@gmail.com

2010 *Mathematics Subject Classification.* 28A20

Key words and phrases. Measure theory, Haar integration

ABSTRACT. This is an introduction to measure theory, integration and function spaces, with all the needed preliminaries included, and with some applications included as well. We first discuss some basic motivations, coming from discrete probability, that we develop in detail, as a preliminary to general measure theory. Then we discuss measure theory, integration and function spaces, all developed in a standard way, and with emphasis on the explicit computation of various integrals. Finally, we come back to probability, discrete and continuous, with a more advanced discussion, of quantum flavor.

Preface

You certainly know some calculus, in one and several variables, along with some basic algebra, geometry and probability. For most questions in science and engineering, that knowledge is enough, although learning new methods and tricks, from time to time, say from older colleagues that you will work with, will always improve your science.

This being said, why not learning some new methods and tricks right away, now that you are still studying. You would perhaps say, why bothering with this, I already had enough troubles with math, and learned enough of it, as to get started in life. But let me ask you a few questions, about your current knowledge, and judge yourself:

(1) The probability for a randomly picked $x \in \mathbb{R}$ to be $x = 1$ is, and you surely know that, $P = 0$. But what does this exactly mean, mathematically speaking?

(2) Along the same lines, but more concerning, everything that happens in the real life, happens with probability $P = 0$. And this is something that we must clarify.

(3) Even worse, in certain physics disciplines like quantum mechanics, or statistical mechanics, P is the only tool available, and $P = 0$, your daily nightmare.

(4) Another interesting thing that you surely know is that the reals are uncountable, $|\mathbb{R}| > |\mathbb{N}| = \infty$. So, how many possible ∞ beasts are there, in mathematics?

(5) What about arbitrarily small infinitesimals, how many $1/\infty$ beasts are there? And with the comment that these can be, technically speaking, very useful.

(6) Less philosophically now, you are surely a heavy user of $dx dy = dy dx$, called Fubini theorem, for integrating 2-variable functions. But, does Fubini always work?

(7) Also, when using polar coordinates, the formula is $dx dy = J dr dt$, with $J = r$. But no one ever fully proves this in class, so is this formula really correct, and why?

(8) You surely know about the Dirac δ_x function from electrostatics, or other physics. But, what is the precise mathematics behind this delta function?

(9) In fact, even mathematicians can pull out tricks based on δ_x . For instance they have a theory where the basic step function is differentiable, with derivative δ_0 .

(10) Still talking classical analysis, you might know from Fourier transform theory that functions can be of type L^1, L^2, L^∞ and so on, and clarifying all this matters.

(11) As even worse classical analysis, arbitrary functions can be non-integrable at all, at least in theory. And with these being non-negligible in quantum mechanics.

(12) In fact, talking worse things that can happen, these are wrong formulae, obtained via correct proofs. So, make sure that you know as much theory as possible.

In the hope that I convinced you that learning some more analysis would be a great investment, and the present book will be here for that. The above questions all fall under the banner “measure theory”, and we will discuss here measure theory, with an introduction to the subject, and answers, from partial to full, to most of the above questions.

This book was conceived as a complement to my calculus book [8], or to any other modern calculus book, which is typically weak on measure theory. Of course, if you are brave enough for learning full calculus directly, the old way, with measure theory included, go with some classical Cold War books, such as those of Rudin [75], [76].

Many thanks go to my students, who had to endure several beta versions of the material presented here, once for instance, long ago, with a measure theory class featuring a full proof for the Zorn Lemma. Thanks as well go to my cats, who seem to deal with everything, successfully, just by using fast binary logic. But we are not all that fast.

Cergy, June 2025

Teo Banica

Contents

Preface	3
Part I. Discrete laws	9
Chapter 1. Real numbers	11
1a. Real numbers	11
1b. Equations, roots	20
1c. Random numbers	26
1d. Probability zero	30
1e. Exercises	32
Chapter 2. Discrete laws	33
2a. Discrete laws	33
2b. Binomial laws	40
2c. Poisson limits	47
2d. Bell numbers	51
2e. Exercises	56
Chapter 3. Combinatorics	57
3a. Group theory	57
3b. Derangements	62
3c. Bessel laws	68
3d. Cumulants	74
3e. Exercises	80
Chapter 4. Central limits	81
4a. Central limits	81
4b. Normal laws	88
4c. Complex variables	94
4d. Wick formula	98
4e. Exercises	104

Part II. Measure theory	105
Chapter 5. Measure theory	107
5a. Metric spaces	107
5b. Compactness	112
5c. Measured spaces	118
5d. Integration theory	123
5e. Exercises	128
Chapter 6. Main theorems	129
6a. Preliminaries	129
6b. Riesz theorem	137
6c. Regular measures	146
6d. Lebesgue measure	148
6e. Exercises	152
Chapter 7. Integration theory	153
7a. Multiple integrals	153
7b. Change of variables	158
7c. Probability, revised	169
7d. Radon-Nikodym	174
7e. Exercises	176
Chapter 8. Contour integrals	177
8a. Curves, surfaces	177
8b. Contour integrals	183
8c. Green and Stokes	190
8d. Gauss, Maxwell	195
8e. Exercises	200
Part III. Function spaces	201
Chapter 9. Banach spaces	203
9a. Normed spaces	203
9b. Banach spaces	211
9c. Linear forms	215
9d. Duality results	220
9e. Exercises	224

Chapter 10. Hilbert spaces	225
10a. Linear algebra	225
10b. Hilbert spaces	231
10c. Bases, separability	237
10d. Linear operators	242
10e. Exercises	248
Chapter 11. Function spaces	249
11a. Distributions	249
11b. Complex analysis	254
11c. Fourier analysis	259
11d. Into the waves	266
11e. Exercises	272
Chapter 12. Linear operators	273
12a. Operator theory	273
12b. Spectral radius	279
12c. Normal operators	289
12d. Diagonalization	292
12e. Exercises	296
Part IV. Haar integration	297
Chapter 13. Random walks	299
13a. Random walks	299
13b. Catalan numbers	302
13c. Stieltjes inversion	307
13d. Finite graphs	312
13e. Exercises	320
Chapter 14. Haar integration	321
14a. Compact groups	321
14b. Haar integration	325
14c. Diagrams, easiness	331
14d. Weingarten formula	340
14e. Exercises	344
Chapter 15. Random matrices	345

15a. Spectral measures	345
15b. Gaussian matrices	351
15c. Wigner and Wishart	355
15d. Block modifications	362
15e. Exercises	368
Chapter 16. Free integration	369
16a. Free tori	369
16b. Free spheres	374
16c. Free rotations	377
16d. Fine structure	386
16e. Exercises	392
Bibliography	393
Index	397

Part I

Discrete laws

*Low lie
The fields of Athenry
Where once we watched
The small free birds fly*

CHAPTER 1

Real numbers

1a. Real numbers

Welcome to measure theory. What we will be doing in this book, namely measure theory and function spaces, is standard advanced undergraduate mathematics, sometimes erring on the graduate side, and can be best described as being a mixture of “advanced calculus” and “rigorous calculus”. That is, you are supposed to know reasonably well basic calculus, in one and several variables, learned from one of the many calculus books available, such as those of Lax and Terrell [67], [68], or why not mine [8], and looking for a continuation of that, more advanced of course, but more rigorous too.

I can hear you thinking “why bothering with more rigorous calculus, what I want to learn is more advanced calculus, that is what looks the most useful”. Good point, and if you are really into this, I can only recommend you, as a continuation of your calculus experience, to start reading some physics. You have here the famous course of Feynman [30], [31], [32], or the lovely books of Griffiths [41], [42], [43], or those of Weinberg [93], [94], [95], if you already know some physics. So go with them, for advanced calculus, physicists know best, that is their job, to come up via advanced calculus to all sorts of amazing formulae, and you will certainly learn many interesting things from them.

This being said, even if you are genuinely interested in physics, a bit of rigor, that you will not necessarily learn from Feynman, Griffiths or Weinberg, can help. To make my point, let me formulate the following question, which is something very natural:

QUESTION 1.1. *The probability for a randomly picked $x \in \mathbb{R}$ to be $x = 1$ is obviously $P = 0$. But what does this mean, mathematically speaking?*

And I hope that you agree with me, this is a very good question, for anyone having a bit of calculus background, and looking to learn more, regardless of the precise goals. For instance, don’t expect to understand any sort of truly advanced physics, which often relies on probabilistic techniques, without having such things perfectly understood.

We will discuss this question several times in this book, first at the end of the present chapter, with a quick but rather advanced explanation, and then on several occasions afterwards, once our knowledge in measure theory and probability improves.

Getting started now, let us first talk about numbers. The commonly accepted fact, by most humans, is that the set $\mathbb{N} = \{0, 1, 2, 3, \dots\}$ was invented by God. According however to certain mathematicians, God only invented the empty set \emptyset , and then \mathbb{N} naturally came afterwards, somehow by inventing itself, according to the following scheme:

$$\begin{aligned} |\emptyset| &= 0 \\ |\{\emptyset\}| &= 1 \\ |\{\emptyset, \{\emptyset\}\}| &= 2 \\ |\{\emptyset, \{\emptyset\}, \{\emptyset, \{\emptyset\}\}\}| &= 3 \\ &\vdots \end{aligned}$$

Along the same lines, also alternatively, according to certain physicists, God only invented the Big Bang, and everything including \mathbb{N} naturally came afterwards. But shall we trust all this modern science, better stick with good old traditional religion.

Once you have \mathbb{N} , solving $a + b = c$ naturally leads you to the set of all integers \mathbb{Z} . Then, once you have \mathbb{Z} , solving $ab = c$ naturally leads you to the set of all rationals \mathbb{Q} . So, this will be our starting point, the set of rationals \mathbb{Q} , defined as follows:

DEFINITION 1.2. *The rational numbers are the quotients $r = a/b$, with $a, b \in \mathbb{Z}$ and $b \neq 0$, identified according to the usual rule for quotients, namely:*

$$\frac{a}{b} = \frac{c}{d} \iff ad = bc$$

These quotients add, multiply and invert according to the following formulae:

$$\frac{a}{b} + \frac{c}{d} = \frac{ad + bc}{bd} \quad , \quad \frac{a}{b} \cdot \frac{c}{d} = \frac{ac}{bd} \quad , \quad \left(\frac{a}{b}\right)^{-1} = \frac{b}{a}$$

We denote the set of rational numbers by \mathbb{Q} , standing for “quotients”.

As a comment here, in the hope that you love fractions. This is what makes the difference between science and humanities, in science we all love fractions, and are particularly excited whenever we must do any computations with them, useful or unuseful.

In more advanced mathematical terms, the above operations, namely sum, product and inversion, tell us that \mathbb{Q} is a field, in the following sense:

DEFINITION 1.3. *A field is a set F with a sum operation $+$ and a product operation \times , subject to the following conditions:*

- (1) $a + b = b + a$, $a + (b + c) = (a + b) + c$, there exists $0 \in F$ such that $a + 0 = 0$, and any $a \in F$ has an inverse $-a \in F$, satisfying $a + (-a) = 0$.
- (2) $ab = ba$, $a(bc) = (ab)c$, there exists $1 \in F$ such that $a1 = a$, and any $a \neq 0$ has a multiplicative inverse $a^{-1} \in F$, satisfying $aa^{-1} = 1$.
- (3) *The sum and product are compatible via $a(b + c) = ab + ac$.*

Apparently, what we did so far, with our philosophical discussion regarding creation, $\diamond \rightarrow \mathbb{N} \rightarrow \mathbb{Z} \rightarrow \mathbb{Q}$, was to construct the simplest possible field, \mathbb{Q} . However, this is not exactly true, because by a strange twist of fate, the numbers $0, 1$, whose presence in a field is mandatory, $0, 1 \in F$, can form themselves a field, with addition as follows:

$$1 + 1 = 0$$

To be more precise, according to our field axioms, we certainly must have:

$$0 + 0 = 0 \times 0 = 0 \times 1 = 1 \times 0 = 0$$

$$0 + 1 = 1 + 0 = 1 \times 1 = 1$$

Thus, everything regarding the addition and multiplication of $0, 1$ is uniquely determined, except for the value of $1 + 1$. And here, you would say that we should normally set $1 + 1 = 2$, with $2 \neq 0$ being a new field element, but the point is that $1 + 1 = 0$ is something natural too, this being the addition modulo 2. And, what we get is a field:

$$\mathbb{F}_2 = \{0, 1\}$$

Let us summarize this finding, along with a bit more, obtained by suitably replacing our 2, used for addition, with an arbitrary prime number p , as follows:

THEOREM 1.4. *The following happen:*

- (1) \mathbb{Q} is the simplest field having the property $1 + \dots + 1 \neq 0$, in the sense that any field F having this property must contain it, $\mathbb{Q} \subset F$.
- (2) The property $1 + \dots + 1 \neq 0$ can hold or not, and if not, the smallest number of terms needed for having $1 + \dots + 1 = 0$ is a certain prime number p .
- (3) $\mathbb{F}_p = \{0, 1, \dots, p-1\}$, with p prime, is the simplest field having the property $1 + \dots + 1 = 0$, with p terms, in the sense that this implies $\mathbb{F}_p \subset F$.

PROOF. All this is basic number theory, the idea being as follows:

(1) This is clear, because $1 + \dots + 1 \neq 0$ tells us that we have an embedding $\mathbb{N} \subset F$, and then by taking inverses with respect to $+$ and \times we obtain $\mathbb{Q} \subset F$.

(2) Again, this is clear, because assuming $1 + \dots + 1 = 0$, with $p = ab$ terms, chosen minimal, we would have a formula as follows, which is a contradiction:

$$\underbrace{(1 + \dots + 1)}_{a \text{ terms}} \underbrace{(1 + \dots + 1)}_{b \text{ terms}} = 0$$

(3) This follows a bit as in (1), with the copy $\mathbb{F}_p \subset F$ consisting by definition of the various sums of type $1 + \dots + 1$, which must cycle modulo p , as shown by (2). \square

Getting back now to our philosophical discussion regarding creation, what we have in Theorem 1.4 is not exactly good news, suggesting that, on purely mathematical grounds, there is a certain rivalry between \mathbb{Q} and \mathbb{F}_p , as being the simplest field. So, which of them shall we study, as being created first? Not an easy question, and as answer, we have:

ANSWER 1.5. *Ignoring what pure mathematics might say, and trusting instead physics and chemistry, we will choose to trust in \mathbb{Q} , as being the simplest field.*

In short, welcome to analysis, and with this being something quite natural, analysis being the main topic of the present book. Moving ahead now, many things can be done with \mathbb{Q} , but getting straight to the point, one thing that fails is solving $x^2 = 2$:

THEOREM 1.6. *The field \mathbb{Q} does not contain a square root of 2:*

$$\sqrt{2} \notin \mathbb{Q}$$

In fact, among integers, only the squares, $n = m^2$ with $m \in \mathbb{N}$, have square roots in \mathbb{Q} .

PROOF. This is something very standard, the idea being as follows:

(1) In what regards $\sqrt{2}$, assuming that $r = a/b$ with $a, b \in \mathbb{N}$ prime to each other satisfies $r^2 = 2$, we have $a^2 = 2b^2$, and so $a \in 2\mathbb{N}$. But then by using again $a^2 = 2b^2$ we obtain $b \in 2\mathbb{N}$ as well, which contradicts our assumption $(a, b) = 1$.

(2) Along the same lines, any prime number $p \in \mathbb{N}$ has the property $\sqrt{p} \notin \mathbb{Q}$, with the proof here being as the above one for $p = 2$, by congruence and contradiction.

(3) More generally, our claim is that any $n \in \mathbb{N}$ which is not a square has the property $\sqrt{n} \notin \mathbb{Q}$. Indeed, we can argue here that our number decomposes as $n = p_1^{a_1} \dots p_k^{a_k}$, with p_1, \dots, p_k distinct primes, and our assumption that n is not a square tells us that one of the exponents $a_1, \dots, a_k \in \mathbb{N}$ must be odd. Moreover, by extracting all the obvious squares from n , we can in fact assume $a_1 = \dots = a_k = 1$. But with this done, we can set $p = p_1$, and the congruence argument from (2) applies, and gives $\sqrt{n} \notin \mathbb{Q}$, as desired. \square

In short, in order to advance with our mathematics, we are in need to introduce the field of real numbers \mathbb{R} . You would probably say that this is very easy, via decimal writing, like everyone does, but before doing that, let me ask you a few questions:

(1) Honestly, do you really like the addition of real numbers, using the decimal form? Let us take, as example, the following computation:

$$\begin{array}{r} 12.456\,783\,872 \\ + 27.536\,678\,377 \\ \hline \end{array}$$

This computation can surely be done, but, annoyingly, it must be done from right to left, instead of left to right, as we would prefer. I mean, personally I would be most interested in knowing first what happens at left, if the integer part is 39 or 40, but go do all the computation, starting from the right, in order to figure out that. In short, my feeling is that this addition algorithm, while certainly good, is a bit deceiving.

(2) What about multiplication. Here things become even more complicated, imagine for instance that Mars attacks, with δ -rays, which are something unknown to us, and

100,000 stronger than γ -rays, and which have paralyzed all our electronics, and that in order to protect Planet Earth, you must do the following multiplication by hand:

$$\begin{array}{r} 12.456\,783\,872 \\ \times 27.536\,678\,377 \\ \hline \end{array}$$

This does not look very inviting, doesn't it. In short, as before with the addition, there is a bit of a bug with all this, the algorithm being too complicated.

(3) Getting now to the problem that we were interested in, namely extracting the square root of 2, here the algorithm is as follows, not very inviting either:

$$\begin{aligned} 1.4^2 < 2 < 1.5^2 &\implies \sqrt{2} = 1.4\dots \\ 1.41^2 < 2 < 1.42^2 &\implies \sqrt{2} = 1.41\dots \\ 1.414^2 < 2 < 1.415^2 &\implies \sqrt{2} = 1.414\dots \\ 1.4142^2 < 2 < 1.4143^2 &\implies \sqrt{2} = 1.4142\dots \\ &\vdots \end{aligned}$$

In short, quite concerning all this, and don't count on such things, mathematics of the decimal form, if Mars attacks. Let us record these findings as follows:

FACT 1.7. *The real numbers $x \in \mathbb{R}$ can be certainly introduced via their decimal form, but with this, the field structure of \mathbb{R} remains something quite unclear.*

Well, it looks like we are a bit stuck. Fortunately, there is a clever solution to this, due to Dedekind. His definition for the real numbers is as follows:

DEFINITION 1.8. *The real numbers $x \in \mathbb{R}$ are formal cuts in the set of rationals,*

$$\mathbb{Q} = A_x \sqcup B_x$$

with such a cut being by definition subject to the following conditions:

$$p \in A_x, q \in B_x \implies p < q, \quad \inf B_x \notin B_x$$

These numbers add and multiply by adding and multiplying the corresponding cuts.

This might look quite original, but believe me, there is some genius behind this definition. As a first observation, we have an inclusion $\mathbb{Q} \subset \mathbb{R}$, obtained by identifying each rational number $r \in \mathbb{Q}$ with the obvious cut that it produces, namely:

$$A_r = \{p \in \mathbb{Q} \mid p \leq r\}, \quad B_r = \{q \in \mathbb{Q} \mid q > r\}$$

As a second observation, the addition and multiplication of real numbers, obtained by adding and multiplying the corresponding cuts, in the obvious way, is something very simple. To be more precise, in what regards the addition, the formula is as follows:

$$A_{x+y} = A_x + A_y$$

As for the multiplication, the formula here is similar, namely $A_{xy} = A_x A_y$, up to some issues with positives and negatives, which are quite easy to untangle, and with this being a good exercise. We can also talk about order between real numbers, as follows:

$$x \leq y \iff A_x \subset A_y$$

But let us perhaps leave more abstractions for later, and go back to more concrete things. As a first success of our theory, we can formulate the following theorem:

THEOREM 1.9. *The equation $x^2 = 2$ has two solutions over the real numbers, namely the positive solution, denoted $\sqrt{2}$, and its negative counterpart, which is $-\sqrt{2}$.*

PROOF. By using $x \rightarrow -x$, it is enough to prove that $x^2 = 2$ has exactly one positive solution $\sqrt{2}$. But this is clear, because $\sqrt{2}$ can only come from the following cut:

$$A_{\sqrt{2}} = \mathbb{Q}_- \sqcup \left\{ p \in \mathbb{Q}_+ \mid p^2 < 2 \right\} \quad , \quad B_{\sqrt{2}} = \left\{ q \in \mathbb{Q}_+ \mid q^2 > 2 \right\}$$

Thus, we are led to the conclusion in the statement. \square

More generally, the same method works in order to extract the square root \sqrt{r} of any number $r \in \mathbb{Q}_+$, or even of any number $r \in \mathbb{R}_+$, and we have the following result:

THEOREM 1.10. *The solutions of $ax^2 + bx + c = 0$ with $a, b, c \in \mathbb{R}$ are*

$$x_{1,2} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

provided that $b^2 - 4ac \geq 0$. In the case $b^2 - 4ac < 0$, there are no solutions.

PROOF. We can write our equation in the following way:

$$\begin{aligned} ax^2 + bx + c = 0 &\iff x^2 + \frac{b}{a}x + \frac{c}{a} = 0 \\ &\iff \left(x + \frac{b}{2a}\right)^2 - \frac{b^2}{4a^2} + \frac{c}{a} = 0 \\ &\iff \left(x + \frac{b}{2a}\right)^2 = \frac{b^2 - 4ac}{4a^2} \\ &\iff x + \frac{b}{2a} = \pm \frac{\sqrt{b^2 - 4ac}}{2a} \end{aligned}$$

Thus, we are led to the conclusion in the statement. \square

Summarizing, we have a nice abstract definition for the real numbers, that we can certainly do some mathematics with. As a first general result now, which is something very useful, and puts us back into real life, and science and engineering, we have:

THEOREM 1.11. *The real numbers $x \in \mathbb{R}$ can be written in decimal form,*

$$x = \pm a_1 \dots a_n . b_1 b_2 b_3 \dots$$

with $a_i, b_i \in \{0, 1, \dots, 9\}$, with the convention $\dots b999 \dots = \dots (b+1)000 \dots$

PROOF. This is something non-trivial, even for the rationals $x \in \mathbb{Q}$ themselves, which require some work in order to be put in decimal form, the idea being as follows:

(1) First of all, our precise claim is that any $x \in \mathbb{R}$ can be written in the form in the statement, with the integer $\pm a_1 \dots a_n$ and then each of the digits b_1, b_2, b_3, \dots providing the best approximation of x , at that stage of the approximation.

(2) Moreover, we have a second claim as well, namely that any expression of type $x = \pm a_1 \dots a_n . b_1 b_2 b_3 \dots$ corresponds to a real number $x \in \mathbb{R}$, and that with the convention $\dots b999 \dots = \dots (b+1)000 \dots$, the correspondence is bijective.

(3) In order to prove now these two assertions, our first claim is that we can restrict the attention to the case $x \in [0, 1)$, and with this meaning of course $0 \leq x < 1$, with respect to the order relation for the reals discussed in the above.

(4) Getting started now, let $x \in \mathbb{R}$, coming from a cut $\mathbb{Q} = A_x \sqcup B_x$. Since the set $A_x \cap \mathbb{Z}$ consists of integers, and is bounded from above by any element $q \in B_x$ of your choice, this set has a maximal element, that we can denote $[x]$:

$$[x] = \max(A_x \cap \mathbb{Z})$$

It follows from definitions that $[x]$ has the usual properties of the integer part, namely:

$$[x] \leq x < [x] + 1$$

Thus we have $x = [x] + y$ with $[x] \in \mathbb{Z}$ and $y \in [0, 1)$, and getting back now to what we want to prove, namely (1,2) above, it is clear that it is enough to prove these assertions for the remainder $y \in [0, 1)$. Thus, we have proved (3), and we can assume $x \in [0, 1)$.

(5) So, assume $x \in [0, 1)$. We are first looking for a best approximation from below of type $0.b_1$, with $b_1 \in \{0, \dots, 9\}$, and it is clear that such an approximation exists, simply by comparing x with the numbers $0.0, 0.1, \dots, 0.9$. Thus, we have our first digit b_1 , and then we can construct the second digit b_2 as well, by comparing x with the numbers $0.b_10, 0.b_11, \dots, 0.b_19$. And so on, which finishes the proof of our claim (1).

(6) In order to prove now the remaining claim (2), let us restrict again the attention, as explained in (4), to the case $x \in [0, 1)$. First, it is clear that any expression of type $x = 0.b_1 b_2 b_3 \dots$ defines a real number $x \in [0, 1]$, simply by declaring that the corresponding cut $\mathbb{Q} = A_x \sqcup B_x$ comes from the following set, and its complement:

$$A_x = \bigcup_{n \geq 1} \left\{ p \in \mathbb{Q} \mid p \leq 0.b_1 \dots b_n \right\}$$

(7) Thus, we have our correspondence between real numbers as cuts, and real numbers as decimal expressions, and we are left with the question of investigating the bijectivity of this correspondence. But here, the only bug that happens is that numbers of type $x = \dots b999\dots$, which produce reals $x \in \mathbb{R}$ via (6), do not come from reals $x \in \mathbb{R}$ via (5). So, in order to finish our proof, we must investigate such numbers.

(8) So, consider an expression of type $\dots b999\dots$. Going back to the construction in (6), we are led to the conclusion that we have the following equality:

$$A_{b999\dots} = B_{(b+1)000\dots}$$

Thus, at the level of the real numbers defined as cuts, we have:

$$\dots b999\dots = \dots (b+1)000\dots$$

But this solves our problem, because by identifying $\dots b999\dots = \dots (b+1)000\dots$ the bijectivity issue of our correspondence is fixed, and we are done. \square

The above theorem was of course quite difficult, but this is how things are. Alternatively, we have the following definition for the real numbers:

THEOREM 1.12. *The field of real numbers \mathbb{R} can be defined as well as the completion of \mathbb{Q} with respect to the usual distance on the rationals, namely*

$$d\left(\frac{a}{b}, \frac{c}{d}\right) = \left|\frac{a}{b} - \frac{c}{d}\right|$$

and with the operations on \mathbb{R} coming from those on \mathbb{Q} , via Cauchy sequences.

PROOF. There are several things going on here, the idea being as follows:

(1) Getting back to Definition 1.2, we know from there what the rational numbers are. But, as a continuation of the material there, we can talk about the distance between such rational numbers, as being given by the formula in the statement, namely:

$$d\left(\frac{a}{b}, \frac{c}{d}\right) = \left|\frac{a}{b} - \frac{c}{d}\right| = \frac{|ad - bc|}{|bd|}$$

(2) Very good, so let us get now into Cauchy sequences. We say that a sequence of rational numbers $\{r_n\} \subset \mathbb{Q}$ is Cauchy when the following condition is satisfied:

$$\forall \varepsilon > 0, \exists N \in \mathbb{N}, m, n \geq N \implies d(r_m, r_n) < \varepsilon$$

Here of course $\varepsilon \in \mathbb{Q}$, because we do not know yet what the real numbers are.

(3) With this notion in hand, the idea will be to define the reals $x \in \mathbb{R}$ as being the limits of the Cauchy sequences $\{r_n\} \subset \mathbb{Q}$. But since these limits are not known yet to exist to us, precisely because they are real, we must employ a trick. So, let us define instead the reals $x \in \mathbb{R}$ as being the Cauchy sequences $\{r_n\} \subset \mathbb{Q}$ themselves.

(4) The question is now, will this work. As a first observation, we have an inclusion $\mathbb{Q} \subset \mathbb{R}$, obtained by identifying each rational $r \in \mathbb{Q}$ with the constant sequence $r_n = r$. Also, we can sum and multiply our real numbers in the obvious way, namely:

$$(r_n) + (p_n) = (r_n + p_n) \quad , \quad (r_n)(p_n) = (r_n p_n)$$

We can also talk about the order between such reals, as follows:

$$(r_n) < (p_n) \iff \exists N, n \geq N \implies r_n < p_n$$

Finally, we can also solve equations of type $x^2 = 2$ over our real numbers, say by using our previous work on the decimal writing, which shows in particular that $\sqrt{2}$ can be approximated by rationals $r_n \in \mathbb{Q}$, by truncating the decimal writing.

(5) However, there is still a bug with our theory, because there are obviously more Cauchy sequences of rationals, than real numbers. In order to fix this, let us go back to the end of step (3) above, and make the following convention:

$$(r_n) = (p_n) \iff d(r_n, p_n) \rightarrow 0$$

(6) But, with this convention made, we have our theory. Indeed, the considerations in (4) apply again, with this change, and we obtain an ordered field \mathbb{R} , containing \mathbb{Q} . Moreover, the equivalence with the Dedekind cuts is something which is easy to establish, and we will leave this as an instructive exercise, and this gives all the results. \square

Very nice all this, so have have two equivalent definitions for the real numbers. Finally, getting back to the decimal writing approach, that can be recycled too, with some analysis know-how, and we have a third possible definition for the real numbers, as follows:

THEOREM 1.13. *The real numbers \mathbb{R} can be defined as well via the decimal form*

$$x = \pm a_1 \dots a_n . a_{n+1} a_{n+2} a_{n+3} \dots$$

with $a_i \in \{0, 1, \dots, 9\}$, with the usual convention for such numbers, namely

$$\dots a999 \dots = \dots (a + 1)000 \dots$$

and with the sum and multiplication coming by writing such numbers as

$$x = \pm \sum_{k \in \mathbb{Z}} a_k 10^{-k}$$

and then summing and multiplying, in the obvious way.

PROOF. This is something which looks quite intuitive, but which in practice, and we insist here, is not exactly beginner level, the idea with this being as follows:

(1) Let us first forget about the precise decimal writing in the statement, and define the real numbers $x \in \mathbb{R}$ as being formal sums as follows, with the sum being over integers

$k \in \mathbb{Z}$ assumed to be greater than a certain integer, $k \geq k_0$:

$$x = \pm \sum_{k \in \mathbb{Z}} a_k 10^{-k}$$

(2) Now by truncating, we can see that what we have here are certain Cauchy sequences of rationals, and with a bit more work, we conclude that the \mathbb{R} that we constructed is precisely the \mathbb{R} that we constructed in Theorem 1.12. Thus, we get the result.

(3) Alternatively, by getting back to Theorem 1.11 and its proof, we can argue, based on that, that the \mathbb{R} that we constructed coincides with the old \mathbb{R} from Definition 1.8, the one constructed via Dedekind cuts, and this gives again all the assertions. \square

1b. Equations, roots

Is the field \mathbb{R} that we constructed good enough, for doing analysis? Certainly yes for many questions, as we both know, but not so quick with conclusions, because we have:

FACT 1.14. *It would be desirable to have complex numbers, $z = a + ib$ with $a, b \in \mathbb{R}$ and with $i^2 = -1$, as for the following dreams to become true:*

- (1) *The polynomial $X^2 + 1$, and other polynomials $P \in \mathbb{R}[X]$ too, to have roots.*
- (2) *In particular, characteristic polynomials $P(x) = \det(A - x)$ to have roots.*
- (3) *And so, to have more matrices $A \in M_N(\mathbb{R})$ which are diagonalizable.*
- (4) *With these including beasts like $A = (df_i/dx_j)_{ij}$, with $f : \mathbb{R}^N \rightarrow \mathbb{R}^N$.*
- (5) *Or other analytic beasts, such as $A = (d^2 f/dx_i dx_j)_{ij}$, with $f : \mathbb{R}^N \rightarrow \mathbb{R}$.*

In short, and you get the point I hope, multivariable calculus begs for complex numbers. So, let us introduce the complex numbers \mathbb{C} . Many interesting things can be said here, and you certainly know all of them, with their summary being as follows:

THEOREM 1.15. *The complex numbers, $z = a + ib$ with $a, b \in \mathbb{R}$ and with i being a formal number satisfying $i^2 = -1$, form a field \mathbb{C} . Moreover:*

- (1) *We have a field embedding $\mathbb{R} \subset \mathbb{C}$, given by $a \rightarrow a + 0 \cdot i$.*
- (2) *Additively, we have $\mathbb{C} \simeq \mathbb{R}^2$, with $z = a + ib$ corresponding to (a, b) .*
- (3) *The length of vectors $r = |z|$, with $z = a + ib$, is given by $r = \sqrt{a^2 + b^2}$.*
- (4) *With $z = r(\cos t + i \sin t)$, the products $z = z' z''$ are given by $r = r' r''$, $t = t' + t''$.*
- (5) *We have $e^{it} = \cos t + i \sin t$, so we can write $z = r e^{it}$.*
- (6) *There are N solutions to the equation $z^N = 1$, called N -th roots of unity.*
- (7) *Any degree 2 equation with complex coefficients has both roots in \mathbb{C} .*

PROOF. We have a field, with $z^{-1} = (a - ib)/(a^2 + b^2)$, and regarding the rest:

- (1) This is clear.
- (2) Again, this is clear.
- (3) Again, this is clear. Observe also that we have $r^2 = z \bar{z}$, with $\bar{z} = a - ib$.

(4) We need here the formulae for the sines and cosines of sums, which are as follows, coming from some trigonometry, done the old way, with triangles in the plane:

$$\cos(s + t) = \cos s \cos t - \sin s \sin t$$

$$\sin(s + t) = \sin s \cos t + \cos s \sin t$$

Indeed, with these formulae in hand, we have the following computation, as desired:

$$\begin{aligned} & (\cos s + i \sin s)(\cos t + i \sin t) \\ &= (\cos s \cos t - \sin s \sin t) + i(\sin s \cos t + \cos s \sin t) \\ &= \cos(s + t) + i \sin(s + t) \end{aligned}$$

(5) In order to prove $e^{it} = \cos t + i \sin t$, consider the following function $f : \mathbb{R} \rightarrow \mathbb{C}$:

$$f(t) = \frac{\cos t + i \sin t}{e^{it}}$$

By using $\sin' = \cos$ and $\cos' = -\sin$, coming from the formulae in (4), we have:

$$\begin{aligned} f'(t) &= (e^{-it}(\cos t + i \sin t))' \\ &= -ie^{-it}(\cos t + i \sin t) + e^{-it}(-\sin t + i \cos t) \\ &= e^{-it}(-i \cos t + \sin t) + e^{-it}(-\sin t + i \cos t) \\ &= 0 \end{aligned}$$

We conclude that $f : \mathbb{R} \rightarrow \mathbb{C}$ is constant, equal to $f(0) = 1$, as desired.

(6) This is clear from (5), with $z = w^k$, with $w = e^{2\pi i/N}$ and $k = 0, 1, \dots, N-1$.

(7) This follows in the usual way, with $\sqrt{re^{it}} = \pm\sqrt{r}e^{it/2}$ at the end, using (5). \square

Quite remarkably, we have in fact the following result, generalizing what we know in degree 2, and telling us that with \mathbb{C} we are safe, no need to look for more:

THEOREM 1.16. *Any polynomial $P \in \mathbb{C}[X]$ decomposes as*

$$P = c(X - a_1) \dots (X - a_N)$$

with $c \in \mathbb{C}$ and with $a_1, \dots, a_N \in \mathbb{C}$.

PROOF. The problem is that of proving that our polynomial has at least one root, because afterwards we can proceed by recurrence. We prove this by contradiction. So, assume that P has no roots, and pick a number $z \in \mathbb{C}$ where $|P|$ attains its minimum:

$$|P(z)| = \min_{x \in \mathbb{C}} |P(x)| > 0$$

Since $Q(t) = P(z+t) - P(z)$ is a polynomial which vanishes at $t = 0$, this polynomial must be of the form $ct^k + \text{higher terms}$, with $c \neq 0$, and with $k \geq 1$ being an integer. We obtain from this that, with $t \in \mathbb{C}$ small, we have the following estimate:

$$P(z+t) \simeq P(z) + ct^k$$

Now let us write $t = rw$, with $r > 0$ small, and with $|w| = 1$. Our estimate becomes:

$$P(z + rw) \simeq P(z) + cr^k w^k$$

Now recall that we have assumed $P(z) \neq 0$. We can therefore choose $w \in \mathbb{T}$ such that cw^k points in the opposite direction to that of $P(z)$, and we obtain in this way:

$$|P(z + rw)| \simeq |P(z) + cr^k w^k| = |P(z)|(1 - |c|r^k)$$

Now by choosing $r > 0$ small enough, as for the error in the first estimate to be small, and overcome by the negative quantity $-|c|r^k$, we obtain from this:

$$|P(z + rw)| < |P(z)|$$

But this contradicts our definition of $z \in \mathbb{C}$, as a point where $|P|$ attains its minimum. Thus P has a root, and by recurrence it has N roots, as stated. \square

As a concrete question, let us try to solve a degree 3 equation, $aX^3 + bX^2 + cX + d = 0$. By linear transformations we can always assume $a = 1, b = 0$, and then it is convenient to write $c = 3p, d = 2q$. Thus, our equation becomes $x^3 + 3px + 2q = 0$, and regarding such equations, we have the following famous result, due to Cardano:

THEOREM 1.17. *For a normalized degree 3 equation, namely*

$$x^3 + 3px + 2q = 0$$

the discriminant is $\Delta = -108(p^3 + q^2)$. Assuming $p, q \in \mathbb{R}$ and $\Delta < 0$, the numbers

$$z = w \sqrt[3]{-q + \sqrt{p^3 + q^2}} + w^2 \sqrt[3]{-q - \sqrt{p^3 + q^2}}$$

with $w = 1, e^{2\pi i/3}, e^{4\pi i/3}$ are the solutions of our equation.

PROOF. There are several things going on here, as follows:

(1) The formula of Δ comes the theory of the discriminant. If you know that, good, and if you don't, no worries, because we will only need $p^3 + q^2 > 0$, in what follows.

(2) With z as in the statement, by using $(a + b)^3 = a^3 + b^3 + 3ab(a + b)$, we have:

$$\begin{aligned} z^3 &= \left(w \sqrt[3]{-q + \sqrt{p^3 + q^2}} + w^2 \sqrt[3]{-q - \sqrt{p^3 + q^2}} \right)^3 \\ &= -2q + 3 \sqrt[3]{-q + \sqrt{p^3 + q^2}} \cdot \sqrt[3]{-q - \sqrt{p^3 + q^2}} \cdot z \\ &= -2q + 3 \sqrt[3]{q^2 - p^3 - q^2} \cdot z \\ &= -2q - 3pz \end{aligned}$$

Thus, we are led to the conclusion in the statement. \square

In degree 4 now, we have the following result, which is famous as well:

THEOREM 1.18. *The roots of a normalized degree 4 equation, written as*

$$x^4 + 6px^2 + 4qx + 3r = 0$$

are as follows, with y satisfying the equation $(y^2 - 3r)(y - 3p) = 2q^2$,

$$x_1 = \frac{1}{\sqrt{2}} \left(-\sqrt{y - 3p} + \sqrt{-y - 3p + \frac{4q}{\sqrt{2y - 6p}}} \right)$$

$$x_2 = \frac{1}{\sqrt{2}} \left(-\sqrt{y - 3p} - \sqrt{-y - 3p + \frac{4q}{\sqrt{2y - 6p}}} \right)$$

$$x_3 = \frac{1}{\sqrt{2}} \left(\sqrt{y - 3p} + \sqrt{-y - 3p - \frac{4q}{\sqrt{2y - 6p}}} \right)$$

$$x_4 = \frac{1}{\sqrt{2}} \left(\sqrt{y - 3p} - \sqrt{-y - 3p - \frac{4q}{\sqrt{2y - 6p}}} \right)$$

and with y being computable via the Cardano formula.

PROOF. This is something quite tricky, the idea being as follows:

(1) To start with, let us write our equation in the following form:

$$x^4 = -6px^2 - 4qx - 3r$$

Now assume that we have a number y satisfying the following equation:

$$(y^2 - 3r)(y - 3p) = 2q^2$$

With this magic number y in hand, our equation takes the following form:

$$\begin{aligned} (x^2 + y)^2 &= x^4 + 2x^2y + y^2 \\ &= -6px^2 - 4qx - 3r + 2x^2y + y^2 \\ &= (2y - 6p)x^2 - 4qx + y^2 - 3r \\ &= (2y - 6p)x^2 - 4qx + \frac{2q^2}{y - 3p} \\ &= \left(\sqrt{2y - 6p} \cdot x - \frac{2q}{\sqrt{2y - 6p}} \right)^2 \end{aligned}$$

(2) Which looks very good, leading us to the following degree 2 equations:

$$x^2 + y + \sqrt{2y - 6p} \cdot x - \frac{2q}{\sqrt{2y - 6p}} = 0$$

$$x^2 + y - \sqrt{2y - 6p} \cdot x + \frac{2q}{\sqrt{2y - 6p}} = 0$$

Now let us write these two degree 2 equations in standard form, as follows:

$$x^2 + \sqrt{2y - 6p} \cdot x + \left(y - \frac{2q}{\sqrt{2y - 6p}} \right) = 0$$

$$x^2 - \sqrt{2y - 6p} \cdot x + \left(y + \frac{2q}{\sqrt{2y - 6p}} \right) = 0$$

(3) Regarding the first equation, the solutions there are as follows:

$$x_1 = \frac{1}{2} \left(-\sqrt{2y - 6p} + \sqrt{-2y - 6p + \frac{8q}{\sqrt{2y - 6p}}} \right)$$

$$x_2 = \frac{1}{2} \left(-\sqrt{2y - 6p} - \sqrt{-2y - 6p + \frac{8q}{\sqrt{2y - 6p}}} \right)$$

As for the second equation, the solutions there are as follows:

$$x_3 = \frac{1}{2} \left(\sqrt{2y - 6p} + \sqrt{-2y - 6p - \frac{8q}{\sqrt{2y - 6p}}} \right)$$

$$x_4 = \frac{1}{2} \left(\sqrt{2y - 6p} - \sqrt{-2y - 6p - \frac{8q}{\sqrt{2y - 6p}}} \right)$$

(4) Now by cutting a $\sqrt{2}$ factor from everything, this gives the formulae in the statement. As for the last claim, regarding the nature of y , this comes from Cardano. \square

We still have to compute the number y appearing in the above via Cardano, and the result here, adding to what we already have in Theorem 1.18, is as follows:

THEOREM 1.19 (continuation). *The value of y in the previous theorem is*

$$y = t + p + \frac{a}{t}$$

where the number t is given by the formula

$$t = \sqrt[3]{b + \sqrt{b^2 - a^3}}$$

with $a = p^2 + r$ and $b = 2p^2 - 3pr + q^2$.

PROOF. The legend has it that this is what comes from Cardano, but depressing and normalizing and solving $(y^2 - 3r)(y - 3p) = 2q^2$ makes it for too many operations, so the most pragmatic way is to simply check this equation. With y as above, we have:

$$\begin{aligned} y^2 - 3r &= t^2 + 2pt + (p^2 + 2a) + \frac{2pa}{t} + \frac{a^2}{t^2} - 3r \\ &= t^2 + 2pt + (3p^2 - r) + \frac{2pa}{t} + \frac{a^2}{t^2} \end{aligned}$$

With this in hand, we have the following computation:

$$\begin{aligned}
(y^2 - 3r)(y - 3p) &= \left(t^2 + 2pt + (3p^2 - r) + \frac{2pa}{t} + \frac{a^2}{t^2} \right) \left(t - 2p + \frac{a}{t} \right) \\
&= t^3 + (a - 4p^2 + 3p^2 - r)t + (2pa - 6p^3 + 2pr + 2pa) \\
&\quad + (3p^2a - ra - 4p^2a + a^2)\frac{1}{t} + \frac{a^3}{t^3} \\
&= t^3 + (a - p^2 - r)t + 2p(2a - 3p^2 + r) + a(a - p^2 - r)\frac{1}{t} + \frac{a^3}{t^3} \\
&= t^3 + 2p(-p^2 + 3r) + \frac{a^3}{t^3}
\end{aligned}$$

Now by using the formula of t in the statement, this gives:

$$\begin{aligned}
(y^2 - 3r)(y - 3p) &= b + \sqrt{b^2 - a^3} - 4p^2 + 6pr + \frac{a^3}{b + \sqrt{b^2 - a^3}} \\
&= b + \sqrt{b^2 - a^3} - 4p^2 + 6pr + b - \sqrt{b^2 - a^3} \\
&= 2b - 4p^2 + 6pr \\
&= 2(2p^2 - 3pr + q^2) - 4p^2 + 6pr \\
&= 2q^2
\end{aligned}$$

Thus, we are led to the conclusion in the statement. \square

In degree 5 and more, things become fairly complicated, and we have:

THEOREM 1.20. *There is no general formula for the roots of polynomials of degree $N = 5$ and higher, with the reason for this, coming from Galois theory, being that the group S_5 is not solvable. The simplest numeric example is $P = X^5 - X - 1$.*

PROOF. This is something quite tricky, the idea being as follows:

(1) Given a field F , assume that the roots of $P \in F[X]$ can be computed by using iterated roots, a bit as for the degree 2 equation, or the degree 3 and 4 equations. Then, algebraically speaking, this gives rise to a tower of fields as follows, with $F_0 = F$, and each F_{i+1} being obtained from F_i by adding a root, $F_{i+1} = F_i(x_i)$, with $x_i^{n_i} \in F_i$:

$$F_0 \subset F_1 \subset \dots \subset F_k$$

(2) In order for Galois theory to apply to this situation, we must make all the extensions normal, which amounts in replacing each $F_{i+1} = F_i(x_i)$ by its extension $K_i(x_i)$, with K_i extending F_i by adding a n_i -th root of unity. Thus, with this replacement, we can assume that the tower in (1) is normal, meaning that all Galois groups are cyclic.

(3) Now by Galois theory, at the level of the corresponding Galois groups we obtain a tower of groups as follows as follows, which is a resolution of the last group G_k , the Galois

group of P , in the sense of group theory, in the sense that all quotients are cyclic:

$$G_1 \subset G_2 \subset \dots \subset G_k$$

As a conclusion, Galois theory tells us that if the roots of a polynomial $P \in F[X]$ can be computed by using iterated roots, then its Galois group $G = G_k$ must be solvable.

(4) In the generic case, the conclusion is that Galois theory tells us that, in order for all polynomials of degree 5 to be solvable, via square roots, the group S_5 , which appears there as Galois group, must be solvable, in the sense of group theory. But this is wrong, because the alternating subgroup $A_5 \subset S_5$ is simple, and therefore not solvable.

(5) Finally, regarding the polynomial $P = X^5 - X - 1$, some elementary computations here, based on arithmetic over $\mathbb{F}_2, \mathbb{F}_3$, and involving various cycles of length 2, 3, 5, show that its Galois group is S_5 . Thus, we have our counterexample.

(6) Finally, let us mention that all this shows as well that a random polynomial of degree 5 or higher is not solvable by square roots, and with this being an elementary consequence of the main result from (4), via some standard analysis arguments. \square

There is a lot of further interesting theory that can be developed here, following Galois and others. For more on all this, we recommend any number theory book.

1c. Random numbers

All that we have been talking about so far, while certainly beautiful and potentially useful, was a bit abstract. So, what about true applications, say for the scientist or engineer using a computer for everything that is heavy mathematics?

So, let us talk now about integrating functions, which is a favorite pastime of our friends, the computers. As definition for the integral, which is something simple and straightforward, that our computer friend will certainly appreciate, we have:

THEOREM 1.21. *We have the Riemann integration formula,*

$$\int_a^b f(x)dx = (b-a) \times \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{k=1}^N f\left(a + \frac{b-a}{N} \cdot k\right)$$

which can serve as a definition for the integral, for the continuous functions.

PROOF. This is standard, by drawing rectangles. We have indeed the following formula, which can stand as a definition for the signed area below the graph of f :

$$\int_a^b f(x)dx = \lim_{N \rightarrow \infty} \sum_{k=1}^N \frac{b-a}{N} \cdot f\left(a + \frac{b-a}{N} \cdot k\right)$$

Thus, we are led to the formula in the statement. \square

Observe that the above formula suggests that $\int_a^b f(x)dx$ is the length of the interval $[a, b]$, namely $b - a$, times the average of f on the interval $[a, b]$. Thinking a bit, this is indeed something true, with no need for Riemann sums, coming directly from definitions, because area means side times average height. Thus, we can also formulate:

THEOREM 1.22. *The integral of a function $f : [a, b] \rightarrow \mathbb{R}$ is given by*

$$\int_a^b f(x)dx = (b - a) \times A(f)$$

where $A(f)$ is the average of f over the interval $[a, b]$.

PROOF. As explained above, this is clear from definitions, via some geometric thinking. Alternatively, this is something which certainly comes from Theorem 1.21. \square

Going ahead with more interpretations of the integral, we have:

THEOREM 1.23. *We have the Monte Carlo integration formula,*

$$\int_a^b f(x)dx = (b - a) \times \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{k=1}^N f(x_k)$$

with $x_1, \dots, x_N \in [a, b]$ being random.

PROOF. We recall from Theorem 1.21 that the idea is to use a formula as follows, with the points $x_1, \dots, x_N \in [a, b]$ being uniformly distributed:

$$\int_a^b f(x)dx = (b - a) \times \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{k=1}^N f(x_k)$$

But this works as well when the points $x_1, \dots, x_N \in [a, b]$ are randomly distributed, for somewhat obvious reasons, and this gives the result. \square

Observe that the Monte Carlo integration works better than Riemann integration, for instance when trying to improve the estimate, via $N \rightarrow N + 1$. Indeed, in the context of Riemann integration, assume that we managed to find an estimate as follows, which in practice requires computing N values of our function f , and making their average:

$$\int_a^b f(x)dx \simeq \frac{b - a}{N} \sum_{k=1}^N f\left(a + \frac{b - a}{N} \cdot k\right)$$

In order to improve this estimate, any extra computed value of our function $f(y)$ will be unuseful. For improving our formula, what we need are N extra values of our function,

$f(y_1), \dots, f(y_N)$, with the points y_1, \dots, y_N being precisely the midpoints of the previous division of $[a, b]$, so that we can write an improvement of our formula, as follows:

$$\int_a^b f(x)dx \simeq \frac{b-a}{2N} \sum_{k=1}^{2N} f\left(a + \frac{b-a}{2N} \cdot k\right)$$

With Monte Carlo, things are far more flexible. Assume indeed that we managed to find an estimate as follows, which again requires computing N values of our function:

$$\int_a^b f(x)dx \simeq \frac{b-a}{N} \sum_{k=1}^N f(x_k)$$

Now if we want to improve this, any extra computed value of our function $f(y)$ will be helpful, because we can set $x_{n+1} = y$, and improve our estimate as follows:

$$\int_a^b f(x)dx \simeq \frac{b-a}{N+1} \sum_{k=1}^{N+1} f(x_k)$$

And isn't this potentially useful, and powerful, when thinking at practically computing integrals, either by hand, or by using a computer. Let us record this finding as follows:

CONCLUSION 1.24. *Monte Carlo integration works better than Riemann integration, when it comes to computing as usual, by estimating, and refining the estimate.*

As another interesting feature of Monte Carlo integration, this works better than Riemann integration for functions having various symmetries, because Riemann integration can get “fooled” by these symmetries, while Monte Carlo remains strong.

As an example for this phenomenon, chosen to be quite drastic, let us attempt to integrate, via both Riemann and Monte Carlo, the following function $f : [0, \pi] \rightarrow \mathbb{R}$:

$$f(x) = \left| \sin(120x) \right|$$

The first few Riemann sums for this function are then as follows:

$$I_2(f) = \frac{\pi}{2}(|\sin 0| + |\sin 60\pi|) = 0$$

$$I_3(f) = \frac{\pi}{3}(|\sin 0| + |\sin 40\pi| + |\sin 80\pi|) = 0$$

$$I_4(f) = \frac{\pi}{4}(|\sin 0| + |\sin 30\pi| + |\sin 60\pi| + |\sin 90\pi|) = 0$$

$$I_5(f) = \frac{\pi}{5}(|\sin 0| + |\sin 24\pi| + |\sin 48\pi| + |\sin 72\pi| + |\sin 96\pi|) = 0$$

$$I_6(f) = \frac{\pi}{6}(|\sin 0| + |\sin 20\pi| + |\sin 40\pi| + |\sin 60\pi| + |\sin 80\pi| + |\sin 100\pi|) = 0$$

\vdots

Based on this evidence, we will conclude, obviously, that we have:

$$\int_0^\pi f(x)dx = 0$$

With Monte Carlo, however, such things cannot happen. Indeed, since there are finitely many points $x \in [0, \pi]$ having the property $\sin(120x) = 0$, a random point $x \in [0, \pi]$ will have the property $|\sin(120x)| > 0$, so Monte Carlo will give, at any $N \in \mathbb{N}$:

$$\int_0^\pi f(x)dx \simeq \frac{b-a}{N} \sum_{k=1}^N f(x_k) > 0$$

Again, this is something interesting, when practically computing integrals, either by hand, or by using a computer. So, let us record, as a complement to Conclusion 1.24:

CONCLUSION 1.25. *Monte Carlo integration is smarter than Riemann integration, because the symmetries of the function can fool Riemann, but not Monte Carlo.*

All this is good to know, and in practice now, we are led into the following question, coming from these Monte Carlo considerations, and from a myriad other reasons:

QUESTION 1.26. *How to produce random numbers?*

And good question this is. The first thought goes into producing first random digits $x \in \{0, 1, \dots, 9\}$, that we can combine afterwards into random real numbers, by using the decimal form. But, how to produce these random digits $x \in \{0, 1, \dots, 9\}$?

You would say, easy, just pick the digits of some reasonably complicated real number $r \in \mathbb{R}$. In practice, however, this is not easy at all, the reasons being as follows:

(1) To start with, any rational number $r \in \mathbb{Q}$, be that a very complicated one, will not do, because the decimal expansion of rationals is periodic.

(2) Then, we can try beasts as $\sqrt{2}$, or more generally, roots of polynomials $P \in \mathbb{Q}[X]$. But here, we potentially run again into troubles, because, say in the context of a Monte Carlo application, if our function f to be integrated happens to be related to P , using random numbers coming from solutions of $P = 0$ might flaw our computation.

(3) So, we are seemingly led into transcendental numbers, those which are not roots of polynomials $P \in \mathbb{Q}[X]$. But here, again, lots of problems waiting for us. For instance being transcendental is certainly not enough, because there are such numbers having only 0, 1 as digits. Also, the simplest such numbers are e, π , and we don't want to use these for integrating our functions f , which usually have something to do with e, π .

So, what to do? Not clear. Question 1.26 is actually one of the points where modern mathematics fails, and for telling you everything, what we scientists do is to use for our

simulations whatever random numbers are provided by the software. And, in case the output starts going completely wrong, pick random numbers from another batch.

1d. Probability zero

We have now a quite good knowledge of numbers, and randomness. So, let us go back to the question that we raised in the beginning of this chapter, namely:

QUESTION 1.27. *The probability for a randomly picked $x \in \mathbb{R}$ to be $x = 1$ is obviously $P = 0$. But what does this mean, mathematically speaking?*

And surprise here, this is something quite tricky. Indeed, the uniform measure on \mathbb{R} , that you first think of, when it comes to randomly pick numbers $x \in \mathbb{R}$, has mass ∞ , so is not a probability measure. Damn. So our formula in Question 1.27, namely $P(x = 1) = 0$, while being something undoubtedly true, and it would be foolish to deny this, and think otherwise, looks more like a “social science fact”, coming from our human methods of picking random numbers $x \in \mathbb{R}$, than something mathematical.

So, in order to answer our question, here we are into some deep thinking, both in social science, and in abstract mathematics. And the situation here is as follows:

(1) Humanities first. If you ask people on the street for random numbers $x \in \mathbb{R}$, it is most likely that most of them will choose small numbers, say $x \in [-1000, 1000]$, and basically no one will think at choices of type $x = 872, 563, 267, 127, 167, 176$.

(2) Of course, there is a similar phenomenon involving decimals too, but let us ignore this. Our finding is that P should come from some sort of bell-shaped curve around the origin 0, and that can only be, say by the Central Limit Theorem, a normal law.

(3) Regarding now the mathematics, this is quite clear. Once we have our normal law, the probability $P(x = 1)$ that we want to compute is obtained by integrating the density of this law between 1 and 1. And we obtain of course 0, no matter what $t > 0$ is.

As a conclusion to this, mystery solved, our answer to Question 1.27 being:

ANSWER 1.28. *The probability for a randomly picked $x \in \mathbb{R}$ to be $x = 1$ is*

$$P = \frac{1}{\sqrt{2\pi t}} \int_1^1 e^{-x^2/2t} dx$$

which means, by computing the integral, $P = 0$.

Well done, but that wasn't trivial, wasn't it. Which leads us now into a bit of a philosophy scare, so if dealing with such a simple question like Question 1.27 was non-trivial, what about more complicated questions of the same type, based on the obvious fact that “everything that happens in the real life, happens with probability $P = 0$ ”.

Very concerning all this, hope you agree with me. In short, we have to review now all our basic probability knowledge, by putting that on a fully rigorous basis, in order to never ever get annoyed by such $P = 0$ things, and have these fully understood, in a fully rigorous way. And good news, this is what we will be doing in this book, by systematically developing measure theory. So, let us record this finding, as follows:

FACT 1.29. *Everything that happens in the real life, happens with probability $P = 0$, and in order to clarify this, we must develop measure theory.*

Which is very nice, at least we know one thing, and have now a plan.

Before developing measure theory, however, let us further explore Question 1.27, with the aim of bringing even more mess to what we have, namely Answer 1.28. Indeed, there are in fact some bugs with that Answer 1.28, the rationale being as follows:

(1) If you ask people on the street for random numbers $x \in \mathbb{R}$, frankly, you will certainly have some answering $x = 1$. Thus, we have in fact $P(x = 1) > 0$.

(2) Looking back at our reasoning leading to Answer 1.28, the modeling error there was by ignoring what happens to the decimals. People don't like them, for sure.

(3) Thus, by simplifying, we must rather understand what happens when picking random integers $x \in \mathbb{Z}$. Or even better, by picking numbers $x \in \mathbb{N}$.

In view of this, let us formulate a more realistic version of Question 1.27, as follows:

QUESTION 1.30. *The probability for a randomly picked $x \in \mathbb{N}$ to be $x = 1$ is obviously $P > 0$. But what does this mean, mathematically speaking, and what exactly is P ?*

So, here we go again with modeling questions, belonging to both social science, and to mathematics. To be more precise, from a pure mathematics viewpoint, the first measure that comes to mind, namely the uniform measure on \mathbb{N} , has mass ∞ , and so is not a probability measure. Thus, wrong way, and we must throw in a bit of social science.

But social science tells us that, exactly as before in the continuous setting, people prefer small numbers, say $x \in [0, 1000]$, to beasts such as $x = 872, 563, 267, 127, 167, 176$. Thus, we can only expect to have a density decreasing from 0 to ∞ , and a bit of thinking, say based on the Poisson Limit Theorem, tells us that what we need is a Poisson law.

Time now to answer Question 1.30. Based on the above, we can formulate:

ANSWER 1.31. *The probability for a randomly picked $x \in \mathbb{N}$ to be $x = 1$ is*

$$P = \frac{1}{e^t} \int_1^\infty \sum_{k=0}^\infty \frac{t^k \delta_k}{k!}$$

which means, by computing the integral, $P = t/e^t$.

Of course, this was something a bit mathematical, with the $P > 0$ phenomenon being definitely explained, but with the actual figure $P = t/e^t$ still depending on a parameter $t > 0$, whose realistic value remains to be computed, based on social science considerations. But we will not get here into this latter question, which looks quite difficult.

As a conclusion to this, let us complement Fact 1.29 with:

FACT 1.32. *Some things happen in the real life with probability $P > 0$, but even these need, mathematically speaking, measure theory, for dealing with the Dirac masses.*

Long story short, no matter what we want to do advanced probability, or even basic probability, to be fully honest, we must develop measure theory first.

1e. Exercises

This was a quite crowded preliminary chapter, and as exercises, we have:

EXERCISE 1.33. *Learn more about fields, and their characteristic.*

EXERCISE 1.34. *Write down a self-contained approach to \mathbb{R} , using the decimal writing.*

EXERCISE 1.35. *Learn about what can be done with numeration bases $\neq 10$.*

EXERCISE 1.36. *Check all the details for the Cardano formulae in degree 3 and 4.*

EXERCISE 1.37. *Learn more about fields, and their Galois theory.*

EXERCISE 1.38. *Prove that e is irrational, then look up the fact that e is transcendental.*

EXERCISE 1.39. *Learn about random numbers, and how to produce them.*

EXERCISE 1.40. *Experiment a bit with Monte Carlo, on a computer.*

As bonus exercise, work more on the material from the last section.

CHAPTER 2

Discrete laws

2a. Discrete laws

We have seen in chapter 1 that even very simple philosophical questions regarding the real numbers \mathbb{R} require some knowledge of advanced probability, or at least of probability theory going beyond what our daily life experience teaches us. Time now to do this, and in this chapter and in the next one we will discuss the discrete probability measures. Then, in chapter 4 we will discuss the continuous probability measures.

As starting point, we have basic things such as playing with coins, dice, or cards. You surely know a bit about this, both from real life and from previous math classes, and for our purposes here, it is convenient to axiomatize things in the following way:

DEFINITION 2.1. *A discrete probability space is a set X , usually finite or countable, whose elements $x \in X$ are called events, together with a function*

$$P : X \rightarrow [0, \infty)$$

called probability function, which is subject to the condition

$$\sum_{x \in X} P(x) = 1$$

telling us that the overall probability for something to happen is 1.

As a first comment, our condition $\sum_{x \in X} P(x) = 1$ perfectly makes sense, and this even if X is uncountable, because the sum of positive numbers is always defined, as a number in $[0, \infty]$, and this no matter how many positive numbers we have.

As a second comment, we have chosen in the above not to assume that X is finite or countable, and this for instance because we want to be able to regard any probability function on \mathbb{N} as a probability function on \mathbb{R} , by setting $P(x) = 0$ for $x \notin \mathbb{N}$.

As a third comment, once we have a probability function $P : X \rightarrow [0, \infty)$ as above, with $P(x) \in [0, 1]$ telling us what the probability for an event $x \in X$ to happen is, we can compute what the probability for a set of events $Y \subset X$ to happen is, by setting:

$$P(Y) = \sum_{y \in Y} P(y)$$

But more on this, mathematical aspects of discrete probability theory, later, when further building on Definition 2.1. Before that, let us explore the basic examples, coming from the real life. And here, there are many things to be learned, as follows:

EXAMPLE 2.2. *Flipping coins.*

Here things are simple and clear, because when you flip a coin the corresponding discrete probability space, together with its probability measure, is as follows:

$$X = \{\text{heads, tails}\} \quad , \quad P(\text{heads}) = P(\text{tails}) = \frac{1}{2}$$

In the case where the coin is biased, as to land on heads with probability $2/3$, and on tails with probability $1/3$, the corresponding probability space is as follows:

$$X = \{\text{heads, tails}\} \quad , \quad P(\text{heads}) = \frac{2}{3} \quad , \quad P(\text{tails}) = \frac{1}{3}$$

More generally, given any number $p \in [0, 1]$, we have an abstract probability space as follows, where we have replaced heads and tails by win and lose:

$$X = \{\text{win, lose}\} \quad , \quad P(\text{win}) = p \quad , \quad P(\text{lose}) = 1 - p$$

Finally, things become more interesting when flipping a coin, biased or not, several times in a row. We will be back to this in a moment, with details.

EXAMPLE 2.3. *Throwing dice.*

Again, things here are simple and clear, because when you throw a die the corresponding probability space, together with its probability measure, is as follows:

$$X = \{1, \dots, 6\} \quad , \quad P(i) = \frac{1}{6} \quad , \quad \forall i$$

As before with coins, we can further complicate this by assuming that the die is biased, say landing on face i with probability $p_i \in [0, 1]$. In this case the corresponding probability space, together with its probability measure, is as follows:

$$X = \{1, \dots, 6\} \quad , \quad P(i) = p_i \quad , \quad p_i \geq 0 \quad , \quad \sum_i p_i = 1$$

Also as before with coins, things become more interesting when throwing a die several times in a row, or equivalently, when throwing several identical dice at the same time. In this latter case, with n identically biased dice, the probability space is as follows:

$$X = \{1, \dots, 6\}^n \quad , \quad P(i_1 \dots i_n) = p_{i_1} \dots p_{i_n} \quad , \quad p_i \geq 0 \quad , \quad \sum_i p_i = 1$$

Observe that the sum 1 condition in Definition 2.1 is indeed satisfied, and with this proving that our dice modeling is bug-free, due to the following computation:

$$\begin{aligned}
 \sum_{i \in X} P(i) &= \sum_{i_1, \dots, i_n} P(i_1 \dots i_n) \\
 &= \sum_{i_1, \dots, i_n} p_{i_1} \dots p_{i_n} \\
 &= \sum_{i_1} p_{i_1} \dots \sum_{i_n} p_{i_n} \\
 &= 1 \times \dots \times 1 \\
 &= 1
 \end{aligned}$$

We will be back to dice in a moment, with further details and computations.

EXAMPLE 2.4. *Playing cards.*

As a third and last example, which is more advanced, and also more dangerous, don't try this at home, let us play some poker. Consider a deck of 32 cards:

$$X = \{7, 8, 9, 10, J, Q, K, A\} \times \{\clubsuit, \diamond, \heartsuit, \spadesuit\}$$

Now let us pick 5 cards, and compute for instance the probability for having four of a kind. The total number of possibilities for a poker hand, 5 cards out of 32, is:

$$\binom{32}{5} = \frac{32 \cdot 31 \cdot 30 \cdot 29 \cdot 28}{2 \cdot 3 \cdot 4 \cdot 5} = 32 \cdot 31 \cdot 29 \cdot 7$$

On the other hand, for having four of a kind, we have to pick that kind, then 4 cards of that kind, and then the 5th card, so the number of possibilities is:

$$\binom{8}{1} \binom{4}{4} \times \binom{7}{1} \binom{4}{1} = 8 \cdot 7 \cdot 4$$

Thus, the probability of having four of a kind is:

$$P(\text{four of a kind}) = \frac{8 \cdot 7 \cdot 4}{32 \cdot 31 \cdot 29 \cdot 7} = \frac{1}{31 \cdot 29} = \frac{1}{899}$$

So far, so good, but you might argue, what if we model our problem as for our poker hand to be ordered, do we still get the same answer? In answer, sure yes, but let us check this. The probability for having four of a kind, computed in this new way, is then:

$$P(\text{four of a kind}) = \frac{8 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 28}{32 \cdot 31 \cdot 30 \cdot 29 \cdot 28} = \frac{1}{31 \cdot 29} = \frac{1}{899}$$

To be more precise, here on the bottom $32 \cdot 31 \cdot 30 \cdot 29 \cdot 28$ stands for the total number of possibilities for an ordered poker hand, 5 out of 32, and on top, exercise for you to figure out what the above numbers 8, 5, then $4 \cdot 3 \cdot 2$, and 28, stand for.

Getting back now to theory, in the general context of Definition 2.1, we can see that what we have there is very close to the biased die, from Example 2.3. Indeed, in the general context of Definition 2.1, we can say that what happens is that we have a die with $|X|$ faces, which is biased such that it lands on face i with probability $P(i)$. Which is quite interesting, so as a conclusion, let us record this finding as follows:

CONCLUSION 2.5. *Discrete probability can be understood as being about throwing a general die, having an arbitrary number of faces, and which is arbitrarily biased too.*

Moving ahead now, let us go back to the context of Definition 2.1, which is the most convenient one, technically speaking. As usual in probability, we are mainly interested in winning. But, winning what? In case we are dealing with a usual die, what we win is what the die says, and on average, what we win is the following quantity:

$$E = \frac{1 + 2 + 3 + 4 + 5 + 6}{6} = 3.5$$

In case we are dealing with the biased die in Example 2.3, again what we win is what the die says, and on average, what we win is the following quantity:

$$E = \sum_i i \times p_i$$

With this understood, what about coins? Here, before doing any computation, we have to assign some numbers to our events, and a standard choice here is as follows:

$$f : \{\text{heads}, \text{tails}\} \rightarrow \mathbb{R} \quad , \quad f(\text{heads}) = 1 \quad , \quad f(\text{tails}) = 0$$

With this choice made, what we can expect to win is the following quantity:

$$\begin{aligned} E(f) &= f(\text{heads}) \times P(\text{heads}) + f(\text{tails}) \times P(\text{tails}) \\ &= 1 \times \frac{1}{2} + 0 \times \frac{1}{2} \\ &= \frac{1}{2} \end{aligned}$$

Of course, in the case where the coin is biased, this computation will lead to a different outcome. And also, with a different convention for f , we will get a different outcome too. Moreover, we can combine if we want these two degrees of flexibility.

In short, you get the point. In order to do some math, in the context of Definition 2.1, we need a random variable $f : X \rightarrow \mathbb{R}$, and the math will consist in computing the expectation of this variable, $E(f) \in \mathbb{R}$. Alternatively, in order to do some business in the context of Definition 2.1, we need some form of “money”, and our random variable $f : X \rightarrow \mathbb{R}$ will stand for that money, and then $E(f) \in \mathbb{R}$, for the average gain.

Let us axiomatize this situation as follows:

DEFINITION 2.6. *A random variable on a probability space X is a function*

$$f : X \rightarrow \mathbb{R}$$

and the expectation of such a random variable is the quantity

$$E(f) = \sum_{x \in X} f(x)P(x)$$

which is best thought as being the average gain, when the game is played.

Here the word “game” refers to the probability space interpretation from Conclusion 2.5. Indeed, in that context, with our discrete set of events X being thought of as corresponding to a generalized die, and by thinking of f as representing some sort of money, the above quantity $E(f)$ is what we win, on average, when playing the game.

We have already seen some good illustrations for Definition 2.6, so time now to get into more delicate aspects. Imagine that you want to set up some sort of business, with your variable $f : X \rightarrow \mathbb{R}$. You are of course mostly interested in the expectation $E(f) \in \mathbb{R}$, but passed that, the way this expectation comes in matters too. For instance:

(1) When your variable is constant, $f = c$, you certainly have $E(f) = c$, and your business will run smoothly, with not so many surprises on the way.

(2) On the opposite, for a complicated variable satisfying $E(f) = c$, your business will be more bumpy, with wins or loses on the way, depending on your skills.

In short, and extrapolating now from business to mathematics, physics, chemistry and everything else, we must complement Definition 2.6 with something finer, regarding the “quality” of the expectation $E(f) \in \mathbb{R}$ appearing there. And the first thought here, which is the correct one, goes to the following number, called variance of our variable:

$$\begin{aligned} V(f) &= E((f - E(f))^2) \\ &= E(f^2) - E(f)^2 \end{aligned}$$

However, let us not stop here. For a total control of your business, be that of financial, mathematical, physical or chemical type, you will certainly want to know more about your variable $f : X \rightarrow \mathbb{R}$. Which leads us into general moments, constructed as follows:

DEFINITION 2.7. *The moments of a variable $f : X \rightarrow \mathbb{R}$ are the numbers*

$$M_k = E(f^k)$$

which satisfy $M_0 = 1$, then $M_1 = E(f)$, and then $V(f) = M_2 - M_1^2$.

And, good news, with this we have all the needed tools in our bag for doing some good business. To put things in a very compacted way, M_0 is about foundations, M_1 is about running some business, M_2 is about running that business well, and M_3 and higher are advanced level, about ruining all the competing businesses.

As a further piece of basic probability, coming this time as a theorem, we have:

THEOREM 2.8. *Given a random variable $f : X \rightarrow \mathbb{R}$, if we define its law as being*

$$\mu = \sum_{x \in X} P(x) \delta_{f(x)}$$

regarded as probability measure on \mathbb{R} , then the moments are given by the formula

$$E(f^k) = \int_{\mathbb{R}} y^k d\mu(y)$$

with the usual convention that each Dirac mass integrates up to 1.

PROOF. There are several things going on here, the idea being as follows:

(1) To start with, given a random variable $f : X \rightarrow \mathbb{R}$, we can certainly talk about its law μ , as being the formal linear combination of Dirac masses in the statement. Our claim is that this is a probability measure on \mathbb{R} , in the sense of Definition 2.1. Indeed, the weight of each point $y \in \mathbb{R}$ is the following quantity, which is positive, as it should:

$$d\mu(y) = \sum_{f(x)=y} P(x)$$

Moreover, the total mass of this measure is 1, as it should, due to:

$$\begin{aligned} \sum_{y \in \mathbb{R}} d\mu(y) &= \sum_{y \in \mathbb{R}} \sum_{f(x)=y} P(x) \\ &= \sum_{x \in X} P(x) \\ &= 1 \end{aligned}$$

Thus, we have indeed a probability measure on \mathbb{R} , in the sense of Definition 2.1.

(2) Still talking basics, let us record as well the following alternative formula for the law, which is clear from definitions, and that we will often use, in what follows:

$$\mu = \sum_{y \in \mathbb{R}} P(f = y) \delta_y$$

(3) Now let us compute the moments of f . With the usual convention that each Dirac mass integrates up to 1, as mentioned in the statement, we have:

$$\begin{aligned} E(f^k) &= \sum_{x \in X} P(x) f(x)^k \\ &= \sum_{y \in \mathbb{R}} y^k \sum_{f(x)=y} P(x) \\ &= \int_{\mathbb{R}} y^k d\mu(y) \end{aligned}$$

Thus, we are led to the conclusions in the statement. \square

The above theorem is quite interesting, because we can see here a relation with integration, as we know it from calculus. In view of this, it is tempting to further go this way, by formulating the following definition, which is something purely mathematical:

DEFINITION 2.9. *Given a set X , which can be finite, countable, or even uncountable, a discrete probability measure on it is a linear combination as follows,*

$$\mu = \sum_{x \in X} \lambda_x \delta_x$$

with the coefficients $\lambda_i \in \mathbb{R}$ satisfying $\lambda_i \geq 0$ and $\sum_i \lambda_i = 1$. For $f : X \rightarrow \mathbb{R}$ we set

$$\int_X f(x) d\mu(x) = \sum_{x \in X} \lambda_x f(x)$$

with the convention that each Dirac mass integrates up to 1.

Observe that, with this, we are now into pure mathematics. However, and we insist on this, it is basic probability, as developed before, which is behind all this. Now by staying abstract for a bit more, with Definition 2.9 in hand, we can recover our previous basic probability notions, from Definition 2.1 and from Theorem 2.8, as follows:

THEOREM 2.10. *With the above notion of discrete probability measure in hand:*

- (1) *A discrete probability space is simply a space X , with a discrete probability measure on it ν . In this picture, the probability function is $P(x) = d\nu(x)$.*
- (2) *Each random variable $f : X \rightarrow \mathbb{R}$ has a law, which is a discrete probability measure on \mathbb{R} . This law is given by $\mu = f_*\nu$, push-forward of ν by f .*

PROOF. This might look a bit scary, but is in fact a collection of trivialities, coming straight from definitions, the details being as follows:

(1) Nothing much to say here, with our assertion being plainly clear, just by comparing Definition 2.1 and Definition 2.9. As a interesting comment, however, in the general

context of Definition 2.9, a probability measure $\mu = \sum_{x \in X} \lambda_x \delta_x$ as there depends only on the following function, called density of our probability measure:

$$\varphi : X \rightarrow \mathbb{R} \quad , \quad \varphi(x) = \lambda_x$$

And, with this notion in hand, our equation $P(x) = d\nu(x)$ simply says that the probability function P is the density of ν . Which is something which is good to know.

(2) Pretty much the same story here, with our first assertion being clear, just by comparing Theorem 2.8 and Definition 2.9. As for the second assertion, consider more generally a probability space (X, ν) , and a function $f : X \rightarrow Y$. We can then construct a probability measure $\mu = f_*\nu$ on Y , called push-forward of ν by f , as follows:

$$\nu = \sum_{x \in X} \lambda_x \delta_x \implies \mu = \sum_{y \in Y} \left(\sum_{f(x)=y} \lambda_x \right) \delta_y$$

Alternatively, at the level of the corresponding measures of the parts $Z \subset Y$, we have the following abstract formula, which looks more conceptual:

$$\mu(Z) = \nu(f^{-1}(Z))$$

In any case, one way or another we can talk about push-forward measures $\mu = f_*\nu$, and in the case of a random variable $f : X \rightarrow \mathbb{R}$, we obtain in this way the law of f . \square

Very nice all this, and needless to say, welcome to measure theory. In what follows we will rather go back to probability theory developed in the old way, as in the beginning of the present chapter, and keep developing that material, because we still have many interesting things to be learned. But, let us keep Definition 2.9 and Theorem 2.10, which are quite interesting, somewhere in our head. We will be back to these later.

2b. Binomial laws

Let us go back now to real-life probability, and try to understand what independence, which can only be something quite simple and intuitive, really means. The most basic occurrence of independence is when throwing a die several times in a row. In order to discuss this, let us first convene for the following rules for the game:

RULES 2.11. *Rolling the die is played with the following rules:*

- (1) *Every time it is 1, 2, 3, 4, 5, your partner wins \$1 from you.*
- (2) *And every time it is 6, you win \$5 from your partner.*

Of course, this is something very particular, but we will complicate the rules later on, no worries for that. Now with these rules agreed on, we have the following result:

THEOREM 2.12. *When rolling a die k times what you can win are quantities of type $6w - k$, with $w = 0, 1, \dots, k$, with the probability for this to happen being:*

$$P(6w - k) = \frac{5^{k-w}}{6^k} \binom{k}{w}$$

Geometrically, your winning curve starts with probability $(5/6)^k$ of losing k , then increases, up to some point, and then decreases, up to probability $1/6^k$ of winning $5k$.

PROOF. There are several things going on here, the idea being as follows:

(1) When rolling the die k times, what will happen is that you will win w times and lose l times, with $k = w + l$. And in this situation, your profit will be, as stated:

$$\begin{aligned} \$ &= 5w - l \\ &= 5w - (k - w) \\ &= 6w - k \end{aligned}$$

(2) As for the probability for this to happen, this is the total number of possibilities for you to win w times, which is $5^{k-w} \binom{k}{w}$, because this amounts in choosing the w times when you will win, among k , then multiplying by 5^{k-w} possibilities, at places where your partner wins, and finally dividing by the total number of possibilities, which is 6^k :

$$P(6w - k) = \frac{5^{k-w}}{6^k} \binom{k}{w}$$

(3) As usual when doing complicated math, let us doublecheck all this, matter of being sure that we did not mess up our counting. First, the sum of all probabilities involved must be 1, and 1 that sum is, as shown by the following computation:

$$\begin{aligned} \sum_{w=0}^k P(6w - k) &= \sum_{w=0}^k \frac{5^{k-w}}{6^k} \binom{k}{w} 5^{k-w} \\ &= \frac{1}{6^k} \sum_{w=0}^k \binom{k}{w} \\ &= \frac{1}{6^k} \sum_{w=0}^k \binom{k}{w} 1^w 5^{k-w} \\ &= \frac{1}{6^k} (1 + 5)^k \\ &= \frac{1}{6^k} \times 6^k \\ &= 1 \end{aligned}$$

(4) Let us triplecheck this as well. Obviously, Rules 2.11 do not favor you, nor your partner, so on average, you should win 0. And this is the case indeed, because:

$$\begin{aligned}
\sum_{w=0}^k P(6w - k) \times (6w - k) &= \frac{1}{6^k} \sum_{w=0}^k 5^{k-w} \binom{k}{w} (6w - k) \\
&= \frac{1}{6^k} \sum_{w=0}^k 5^{k-w} \binom{k}{w} 5w - \frac{1}{6^k} \sum_{w=0}^k 5^{k-w} \binom{k}{w} (k - w) \\
&= \frac{5}{6^k} \sum_{w=0}^k 5^{k-w} \binom{k}{w} w - \frac{1}{6^k} \sum_{w=0}^k 5^{k-w} \binom{k}{w} (k - w) \\
&= \frac{5k}{6^k} \sum_{w=0}^k 5^{k-w} \binom{k-1}{w-1} - \frac{k}{6^k} \sum_{w=0}^k 5^{k-w} \binom{k-1}{w} \\
&= \frac{5k}{6^k} (1+5)^{k-1} - \frac{5k}{6^k} (1+5)^{k-1} \\
&= 0
\end{aligned}$$

(5) This last computation was hot, wasn't it, but in math, triplechecks are mandatory. In any case theorem proved, and the final conclusions in the statement are clear too. \square

Quite interestingly, Theorem 2.12 is best seen, both at the level of the statement, and of the proof, from the viewpoint of your partner. Let us record this, as follows:

THEOREM 2.13. *When rolling a die k times what you can win are quantities of type $5k - 6l$, with $l = 0, 1, \dots, k$, with the probability for this to happen being:*

$$P(5k - 6l) = \frac{5^l}{6^k} \binom{k}{l}$$

Geometrically, your winning curve starts with probability $(5/6)^k$ of losing k , then increases, up to some point, and then decreases, up to probability $1/6^k$ of winning $5k$.

PROOF. As before, when rolling the die k times, you will win w times and lose l times, with $k = w + l$. And in this situation, your profit will be, as stated:

$$\begin{aligned}
\$ &= 5w - l \\
&= 5(k - l) - l \\
&= 5k - 6l
\end{aligned}$$

As for the rest, we already know all this from Theorem 2.12, but the point is that the proof of Theorem 2.12 becomes slightly simpler when using l instead of w . \square

Now with Theorem 2.13 in hand, it is quite clear that the basic $1/6 - 5/6$ probabilities at dice can be repaced with something of type $p - (1 - p)$, with $p \in [0, 1]$ being arbitrary. We are led in this way to the following notions, which are quite general:

DEFINITION 2.14. *Given $p \in [0, 1]$, the Bernoulli law of parameter p is given by:*

$$P(\text{win}) = p \quad , \quad P(\text{lose}) = 1 - p$$

More generally, the k -th binomial law of parameter p , with $k \in \mathbb{N}$, is given by

$$P(s) = p^s(1-p)^{k-s} \binom{k}{s}$$

with the Bernoulli law appearing at $k = 1$, with $s = 1, 0$ here standing for win and lose.

Let us try now to understand the relation between the Bernoulli and binomial laws. Indeed, we know that the Bernoulli laws produce the binomial laws, simply by iterating the game, from 1 throw to $k \in \mathbb{N}$ throws. Obviously, what matters in all this is the “independence” of our coin throws, so let us record this finding, as follows:

THEOREM 2.15. *The following happen, in the context of the biased coin game:*

- (1) *The Bernoulli laws μ_{ber} produce the binomial laws μ_{bin} , by iterating the game $k \in \mathbb{N}$ times, via the independence of the throws.*
- (2) *We have in fact $\mu_{bin} = \mu_{ber}^{*k}$, with $*$ being the convolution operation for real probability measures, given by $\delta_x * \delta_y = \delta_{x+y}$, and linearity.*

PROOF. Obviously, this is something a bit informal, but let us prove this as stated, and we will come back later to it, with precise definitions, general theorems and everything. In what regards the first assertion, nothing to be said there, this is what life teaches us. As for the second assertion, the formula $\mu_{bin} = \mu_{ber}^{*k}$ there certainly looks like mathematics, so job for us to figure out what this exactly means. And, this can be done as follows:

(1) The first idea is to encapsulate the data from Definition 2.14 into the probability measures associated to the Bernoulli and binomial laws. For the Bernoulli law, the corresponding measure is as follows, with the δ symbols standing for Dirac masses:

$$\mu_{ber} = (1-p)\delta_0 + p\delta_1$$

As for the binomial law, here the measure is as follows, constructed in a similar way, you get the point I hope, again with the δ symbols standing for Dirac masses:

$$\mu_{bin} = \sum_{s=0}^k p^s(1-p)^{k-s} \binom{k}{s} \delta_s$$

(2) Getting now to independence, the point is that, as we will soon discover abstractly, the mathematics there is that of the following formula, with $*$ standing for the convolution operation for the real measures, which is given by $\delta_x * \delta_y = \delta_{x+y}$ and linearity:

$$\mu_{bin} = \underbrace{\mu_{ber} * \dots * \mu_{ber}}_{k \text{ terms}}$$

(3) To be more precise, this latter formula does hold indeed, as a straightforward application of the binomial formula, the formal proof being as follows:

$$\begin{aligned}
\mu_{ber}^{*k} &= ((1-p)\delta_0 + p\delta_1)^{*k} \\
&= \sum_{s=0}^k p^s (1-p)^{k-s} \binom{k}{s} \delta_0^{*(k-s)} * \delta_1^{*s} \\
&= \sum_{s=0}^k p^s (1-p)^{k-s} \binom{k}{s} \delta_s \\
&= \mu_{bin}
\end{aligned}$$

(4) Summarizing, save for some uncertainties regarding what independence exactly means, mathematically speaking, and more on this in a moment, theorem proved. \square

Getting to formal mathematical work now, let us start with the following straightforward definition, inspired by what happens for coins, dice and cards:

DEFINITION 2.16. *We say that two variables $f, g : X \rightarrow \mathbb{R}$ are independent when*

$$P(f = x, g = y) = P(f = x)P(g = y)$$

happens, for any $x, y \in \mathbb{R}$.

As already mentioned, this is something very intuitive, inspired by what happens for coins, dice and cards. As a first result now regarding independence, we have:

THEOREM 2.17. *Assuming that $f, g : X \rightarrow \mathbb{R}$ are independent, we have:*

$$E(fg) = E(f)E(g)$$

More generally, we have the following formula, for the mixed moments,

$$E(f^k g^l) = E(f^k)E(g^l)$$

and the converse holds, in the sense that this formula implies the independence of f, g .

PROOF. We have indeed the following computation, using the independence of f, g :

$$\begin{aligned}
E(f^k g^l) &= \sum_{xy} x^k y^l P(f = x, g = y) \\
&= \sum_{xy} x^k y^l P(f = x)P(g = y) \\
&= \sum_x x^k P(f = x) \sum_y y^l P(g = y) \\
&= E(f^k)E(g^l)
\end{aligned}$$

As for the last assertion, this is clear too, because having the above computation work, for any $k, l \in \mathbb{N}$, amounts in saying that the independence formula for f, g holds. \square

Regarding now the convolution operation, motivated by what we found before, in Theorem 2.15, let us start with the following abstract definition:

DEFINITION 2.18. *Given a space X with a sum operation $+$, we can define the convolution of any two discrete probability measures on it,*

$$\mu = \sum_i a_i \delta_{x_i} \quad , \quad \nu = \sum_j b_j \delta_{y_j}$$

as being the discrete probability measure given by the following formula:

$$\mu * \nu = \sum_{ij} a_i b_j \delta_{x_i + y_j}$$

That is, the convolution operation $$ is defined by $\delta_x * \delta_y = \delta_{x+y}$, and linearity.*

As a first observation, our operation is well-defined, with $\mu * \nu$ being indeed a discrete probability measure, because the weights are positive, $a_i b_j \geq 0$, and their sum is:

$$\sum_{ij} a_i b_j = \sum_i a_i \sum_j b_j = 1 \times 1 = 1$$

Also, the above definition agrees with what we did before with coins, and Bernoulli and binomial laws. We have in fact the following general result:

THEOREM 2.19. *Assuming that $f, g : X \rightarrow \mathbb{R}$ are independent, we have*

$$\mu_{f+g} = \mu_f * \mu_g$$

where $$ is the convolution of real probability measures.*

PROOF. We have indeed the following straightforward verification, based on the independence formula from Definition 2.16, and on Definition 2.18:

$$\begin{aligned} \mu_{f+g} &= \sum_{x \in \mathbb{R}} P(f + g = x) \delta_x \\ &= \sum_{y, z \in \mathbb{R}} P(f = y, g = z) \delta_{y+z} \\ &= \sum_{y, z \in \mathbb{R}} P(f = y) P(g = z) \delta_y * \delta_z \\ &= \left(\sum_{y \in \mathbb{R}} P(f = y) \delta_y \right) * \left(\sum_{z \in \mathbb{R}} P(g = z) \delta_z \right) \\ &= \mu_f * \mu_g \end{aligned}$$

Thus, we are led to the conclusion in the statement. □

Before going further, let us attempt as well to find a proof of Theorem 2.19, based on the moment characterization of independence, from Theorem 2.17. For this purpose, we will need the following standard fact, which is of certain theoretical interest:

THEOREM 2.20. *The sequence of moments*

$$M_k = \int_{\mathbb{R}} x^k d\mu(x)$$

uniquely determines the law.

PROOF. Indeed, assume that the law of our variable is as follows:

$$\mu = \sum_i \lambda_i \delta_{x_i}$$

The sequence of moments is then given by the following formula:

$$M_k = \sum_i \lambda_i x_i^k$$

But it is then standard calculus to recover the numbers $\lambda_i, x_i \in \mathbb{R}$, and so the measure μ , out of the sequence of numbers M_k . Indeed, assuming that the numbers x_i are $0 < x_1 < \dots < x_n$ for simplifying, in the $k \rightarrow \infty$ limit we have the following formula:

$$M_k \sim \lambda_n x_n^k$$

Thus, we got the parameters $\lambda_n, x_n \in \mathbb{R}$ of our measure μ , and then by subtracting them and doing an obvious recurrence, we get the other parameters $\lambda_i, x_i \in \mathbb{R}$ as well. We will leave the details here as an instructive exercise, and come back to this problem later in this book, with more advanced and clever methods for dealing with it. \square

Getting back now to our philosophical question above, namely recovering Theorem 2.19 via moment technology, we can now do this, the result being as follows:

THEOREM 2.21. *Assuming that $f, g : X \rightarrow \mathbb{R}$ are independent, the measures*

$$\mu_{f+g} \quad , \quad \mu_f * \mu_g$$

have the same moments, and so, they coincide.

PROOF. We have the following computation, using the independence of f, g :

$$\begin{aligned} M_k(f+g) &= E((f+g)^k) \\ &= \sum_r \binom{k}{r} E(f^r g^{k-r}) \\ &= \sum_r \binom{k}{r} M_r(f) M_{k-r}(g) \end{aligned}$$

On the other hand, we have as well the following computation:

$$\begin{aligned}
 \int_X x^k d(\mu_f * \mu_g)(x) &= \int_{X \times X} (x + y)^k d\mu_f(x) d\mu_g(y) \\
 &= \sum_r \binom{k}{r} \int_X x^r d\mu_f(x) \int_X y^{k-r} d\mu_g(y) \\
 &= \sum_r \binom{k}{r} M_r(f) M_{k-r}(g)
 \end{aligned}$$

Thus, job done, and theorem proved, or rather Theorem 2.19 reproved. \square

Getting back now to the basic theory of independence, here is now a second result, coming as a continuation of Theorem 2.19, which is something more advanced:

THEOREM 2.22. *Assuming that $f, g : X \rightarrow \mathbb{R}$ are independent, we have*

$$F_{f+g} = F_f F_g$$

where $F_f(x) = E(e^{ixf})$ is the Fourier transform.

PROOF. We have the following computation, using Theorem 2.19:

$$\begin{aligned}
 F_{f+g}(x) &= \int_X e^{ixz} d\mu_{f+g}(z) \\
 &= \int_X e^{ixz} d(\mu_f * \mu_g)(z) \\
 &= \int_{X \times X} e^{ix(z+t)} d\mu_f(z) d\mu_g(t) \\
 &= \int_X e^{ixz} d\mu_f(z) \int_X e^{ixt} d\mu_g(t) \\
 &= F_f(x) F_g(x)
 \end{aligned}$$

Thus, we are led to the conclusion in the statement. \square

Of course, you might wonder what $i \in \mathbb{C}$ has to do with all this. Good point, this is something quite subtle, and more on this later, trust me in the meantime.

2c. Poisson limits

As a continuation of the above, and at a more advanced level, let us discuss now the key limiting result in discrete probability, which is the Poisson Limit Theorem (PLT). This result involves the Poisson laws p_t , constructed as follows:

DEFINITION 2.23. *The Poisson law of parameter 1 is the following measure,*

$$p_1 = \frac{1}{e} \sum_{k \geq 0} \frac{\delta_k}{k!}$$

and the Poisson law of parameter $t > 0$ is the following measure,

$$p_t = e^{-t} \sum_{k \geq 0} \frac{t^k}{k!} \delta_k$$

with the letter “p” standing for Poisson.

We are using here, as usual, some simplified notations for these laws. Observe that our laws have indeed mass 1, as they should, due to the following key formula:

$$e^t = \sum_{k \geq 0} \frac{t^k}{k!}$$

We will see in the moment why these measures appear a bit everywhere, in the discrete context, the reasons for this coming from the Poisson Limit Theorem (PLT), which is closely related to our previous investigations regarding the Bernoulli and binomial laws.

For the moment, let us first develop some general theory, for these Poisson laws. We first have the following result, regarding their mean and variance:

PROPOSITION 2.24. *The mean and variance of p_t are given by:*

$$E = t \quad , \quad V = t$$

In particular for the Poisson law p_1 we have $E = 1, V = 1$.

PROOF. We have two computations to be performed, as follows:

(1) Regarding the mean, this can be computed as follows:

$$\begin{aligned} E &= e^{-t} \sum_{k \geq 0} \frac{t^k}{k!} \cdot k \\ &= e^{-t} \sum_{k \geq 1} \frac{t^k}{(k-1)!} \\ &= e^{-t} \sum_{l \geq 0} \frac{t^{l+1}}{l!} \\ &= t e^{-t} \sum_{l \geq 0} \frac{t^l}{l!} \\ &= t \end{aligned}$$

(2) For the variance, we first compute the second moment, as follows:

$$\begin{aligned}
M_2 &= e^{-t} \sum_{k \geq 0} \frac{t^k}{k!} \cdot k^2 \\
&= e^{-t} \sum_{k \geq 1} \frac{t^k k}{(k-1)!} \\
&= e^{-t} \sum_{l \geq 0} \frac{t^{l+1}(l+1)}{l!} \\
&= te^{-t} \sum_{l \geq 0} \frac{t^l l}{l!} + te^{-t} \sum_{l \geq 0} \frac{t^l}{l!} \\
&= te^{-t} \sum_{l \geq 1} \frac{t^l}{(l-1)!} + t \\
&= t^2 e^{-t} \sum_{m \geq 0} \frac{t^m}{m!} + t \\
&= t^2 + t
\end{aligned}$$

Thus the variance is $V = M_2 - E^2 = (t^2 + t) - t^2 = t$, as claimed. \square

At the theoretical level now, we first have the following result:

THEOREM 2.25. *We have the following formula, for any $s, t > 0$,*

$$p_s * p_t = p_{s+t}$$

so the Poisson laws form a convolution semigroup.

PROOF. By using $\delta_k * \delta_l = \delta_{k+l}$ and the binomial formula, we obtain:

$$\begin{aligned}
p_s * p_t &= e^{-s} \sum_k \frac{s^k}{k!} \delta_k * e^{-t} \sum_l \frac{t^l}{l!} \delta_l \\
&= e^{-s-t} \sum_n \delta_n \sum_{k+l=n} \frac{s^k t^l}{k! l!} \\
&= e^{-s-t} \sum_n \frac{\delta_n}{n!} \sum_{k+l=n} \frac{n!}{k! l!} s^k t^l \\
&= e^{-s-t} \sum_n \frac{(s+t)^n}{n!} \delta_n \\
&= p_{s+t}
\end{aligned}$$

Thus, we are led to the conclusion in the statement. \square

Next in line, we have the following result, which is fundamental as well:

THEOREM 2.26. *The Poisson laws appear as formal exponentials*

$$p_t = \sum_k \frac{t^k (\delta_1 - \delta_0)^{*k}}{k!}$$

with respect to the convolution of measures $*$.

PROOF. By using the binomial formula, the measure on the right is:

$$\begin{aligned} \mu &= \sum_k \frac{t^k}{k!} \sum_{r+s=k} (-1)^s \frac{k!}{r!s!} \delta_r \\ &= \sum_k t^k \sum_{r+s=k} (-1)^s \frac{\delta_r}{r!s!} \\ &= \sum_r \frac{t^r \delta_r}{r!} \sum_s \frac{(-1)^s t^s}{s!} \\ &= \frac{1}{e^t} \sum_r \frac{t^r \delta_r}{r!} \\ &= p_t \end{aligned}$$

Thus, we are led to the conclusion in the statement. □

Regarding now the Fourier transform computation, this is as follows:

THEOREM 2.27. *The Fourier transform of p_t is given by*

$$F_{p_t}(y) = \exp((e^{iy} - 1)t)$$

for any $t > 0$.

PROOF. We have indeed the following computation:

$$\begin{aligned} F_{p_t}(y) &= e^{-t} \sum_k \frac{t^k}{k!} F_{\delta_k}(y) \\ &= e^{-t} \sum_k \frac{t^k}{k!} e^{iky} \\ &= e^{-t} \sum_k \frac{(e^{iy}t)^k}{k!} \\ &= \exp(-t) \exp(e^{iy}t) \\ &= \exp((e^{iy} - 1)t) \end{aligned}$$

Thus, we obtain the formula in the statement. □

Observe that the above result provides us with an alternative proof for Theorem 2.25, due to the fact that the logarithm of the Fourier transform is linear in t .

We can now establish the Poisson Limit Theorem, as follows:

THEOREM 2.28 (PLT). *We have the following convergence, in moments,*

$$\left(\left(1 - \frac{t}{n} \right) \delta_0 + \frac{t}{n} \delta_1 \right)^{*n} \rightarrow p_t$$

for any $t > 0$.

PROOF. Let us denote by ν_n the measure under the convolution sign. We have the following computation, for the Fourier transform of the limit:

$$\begin{aligned} F_{\delta_r}(y) = e^{iry} &\implies F_{\nu_n}(y) = \left(1 - \frac{t}{n} \right) + \frac{t}{n} e^{iy} \\ &\implies F_{\nu_n^{*n}}(y) = \left(\left(1 - \frac{t}{n} \right) + \frac{t}{n} e^{iy} \right)^n \\ &\implies F_{\nu_n^{*n}}(y) = \left(1 + \frac{(e^{iy} - 1)t}{n} \right)^n \\ &\implies F(y) = \exp((e^{iy} - 1)t) \end{aligned}$$

Thus, we obtain indeed the Fourier transform of p_t , as desired. \square

Many further things can be said here, mixing Bernoulli laws and variables, binomial laws and variables, and Poisson laws and variables, and in particular clarifying the speculations from chapter 1. For all this, we recommend any specialized probability book.

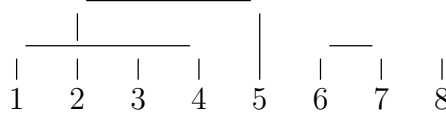
In what concerns us, the above will do. We have Theorem 2.28, which is something nice and mathematical, and reasonably connected to real-life probability, in the obvious way, and this is what we will need, for further developing our mathematics.

2d. Bell numbers

Let us discuss now the moments of the Poisson laws. As we will soon discover, things are quite tricky here, in relation with a lot of subtle mathematics, and understanding this type of mathematics will be actually a main theme of discussion, for this book.

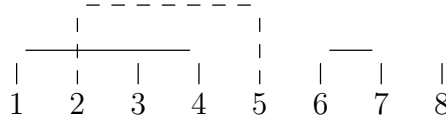
Let us start with something nice, and seemingly unrelated, namely:

DEFINITION 2.29. We denote by $P(k)$ the set of partitions of $\{1, \dots, k\}$, with these partitions $\pi \in P(k)$ being most conveniently being drawn as diagrams,



with the strings joining the numbers belonging to the same block of π . That is, the above diagram represents the partition $\{1, \dots, 8\} = \{1, 3, 4\} \cup \{2, 5\} \cup \{6, 7\} \cup \{8\}$.

Observe that there is a bit of care to be taken with this convention, in respect to the crossings. We can either proceed as above, with the $\{2, 5\}$ block being represented “under” the block $\{1, 3, 4\}$, or use different types of strings, as for instance:



Both conventions are good, and we will be mostly using the first one, that from Definition 2.29. Now, let us study these partitions. And here, surprise, instead of pulling a theorem, as you would expect, we must formulate something quite modest, as follows:

PROPOSITION 2.30. The Bell numbers $B_k = |P(k)|$ satisfy the recurrence relation

$$B_{k+1} = \sum_s \binom{k}{s} B_{k-s}$$

with initial data $B_0 = 1$, $B_1 = 1$, and are numerically as follows:

$$1, 1, 2, 5, 15, 52, 203, 877, 4140, 21147, 115975, 678570, \dots$$

However, there is no mathematical formula for B_k .

PROOF. There are several things going on here, the idea being as follows:

(1) Experiments first, before anything, let us compute a few Bell numbers. Obviously $B_1 = 1$, and then we have $B_2 = 2$, the partitions being as follows:

$$|| \quad , \quad \sqcap$$

Next, we have $B_3 = 5$, the partitions at $k = 3$ being as follows:

$$||| \quad , \quad \sqcap| \quad , \quad \sqcap \quad , \quad |\sqcap \quad , \quad \sqcap\sqcap$$

At $k = 4$ now, things become more complex, and it is better to trick. We can count the partitions up to permutations of the corresponding diagrams, and with this convention made, here are the relevant partitions and their multiplicities, leading to $B_4 = 15$:

$$|||| \times 1 \quad , \quad \sqcap||| \times 6 \quad , \quad \sqcap\sqcap \times 3 \quad , \quad |\sqcap| \times 4 \quad , \quad \sqcap\sqcap\sqcap \times 1$$

The same method works at $k = 5$, with the block distributions and multiplicities, which are simpler to draw than partitions, being as follows, leading to $B_5 = 52$:

$$11111 \rightarrow 1, \quad 2111 \rightarrow 10, \quad 221 \rightarrow 15, \quad 311 \rightarrow 10, \quad 32 \rightarrow 10, \quad 41 \rightarrow 5, \quad 5 \rightarrow 1$$

As for the case $k = 6$, where $B_6 = 203$, we will leave this as an instructive exercise.

(2) Let us try now to find a recurrence for these Bell numbers. Since a partition of $\{1, \dots, k+1\}$ appears by choosing s partners for 1, among the k numbers available, and then partitioning the $k-s$ elements left, we have the following formula:

$$B_{k+1} = \sum_s \binom{k}{s} B_{k-s}$$

Observe that this formula forces us to talk about $B_0 = 1$, as done in the statement.

(3) As for the last assertion, regarding the non-computability of the Bell numbers, take this as a physics fact. Mankind has tried to find a formula for these numbers, had not found anything, and we are reporting here this finding, which is of course rock-solid. \square

In relation now with the Poisson laws, we first have the following result:

THEOREM 2.31. *The moments of p_1 are the Bell numbers,*

$$M_k(p_1) = |P(k)|$$

where $P(k)$ is the set of partitions of $\{1, \dots, k\}$.

PROOF. The moments of p_1 are given by the following formula:

$$M_k = \frac{1}{e} \sum_s \frac{s^k}{s!}$$

We have the following recurrence formula for these moments:

$$\begin{aligned}
M_{k+1} &= \frac{1}{e} \sum_s \frac{(s+1)^{k+1}}{(s+1)!} \\
&= \frac{1}{e} \sum_s \frac{(s+1)^k}{s!} \\
&= \frac{1}{e} \sum_s \frac{s^k}{s!} \left(1 + \frac{1}{s}\right)^k \\
&= \frac{1}{e} \sum_s \frac{s^k}{s!} \sum_r \binom{k}{r} s^{-r} \\
&= \sum_r \binom{k}{r} \cdot \frac{1}{e} \sum_s \frac{s^{k-r}}{s!} \\
&= \sum_r \binom{k}{r} M_{k-r}
\end{aligned}$$

Let us try now to find a recurrence for the Bell numbers:

$$B_k = |P(k)|$$

A partition of $\{1, \dots, k+1\}$ appears by choosing r neighbors for 1, among the k numbers available, and then partitioning the $k-r$ elements left. Thus, we have:

$$B_{k+1} = \sum_r \binom{k}{r} B_{k-r}$$

Thus, the numbers M_k satisfy the same recurrence as the numbers B_k . Regarding now the initial values, for the moments of p_1 , these are:

$$M_0 = 1 \quad , \quad M_1 = 1$$

Indeed, the formula $M_0 = 1$ is clear, and the formula $M_1 = 1$ follows from:

$$M_1 = \frac{1}{e} \sum_s \frac{s}{s!} = \frac{1}{e} \times e = 1$$

Now by using the above recurrence we obtain from this:

$$M_2 = \sum_r \binom{1}{r} M_{k-r} = 1 + 1 = 2$$

Thus, we can say that the initial values for the moments are:

$$M_1 = 1 \quad , \quad M_2 = 2$$

As for the Bell numbers, here the initial values are as follows:

$$B_1 = 1 \quad , \quad B_2 = 2$$

Thus the initial values coincide, and so these numbers are equal, as stated. \square

More generally now, we have the following result:

THEOREM 2.32. *The moments of p_t are given by*

$$M_k(p_t) = \sum_{\pi \in P(k)} t^{|\pi|}$$

where $|\cdot|$ is the number of blocks.

PROOF. Observe first that the formula in the statement generalizes the one in Theorem 2.31, because at $t = 1$ we obtain, as we should:

$$M_k(p_1) = \sum_{\pi \in P(k)} 1^{|\pi|} = |P(k)| = B_k$$

In general now, the moments of p_t with $t > 0$ are given by:

$$M_k = e^{-t} \sum_s \frac{t^s s^k}{s!}$$

We have the following recurrence formula for the moments of p_t :

$$\begin{aligned} M_{k+1} &= e^{-t} \sum_s \frac{t^{s+1} (s+1)^{k+1}}{(s+1)!} \\ &= e^{-t} \sum_s \frac{t^{s+1} (s+1)^k}{s!} \\ &= e^{-t} \sum_s \frac{t^{s+1} s^k}{s!} \left(1 + \frac{1}{s}\right)^k \\ &= e^{-t} \sum_s \frac{t^{s+1} s^k}{s!} \sum_r \binom{k}{r} s^{-r} \\ &= \sum_r \binom{k}{r} \cdot e^{-t} \sum_s \frac{t^{s+1} s^{k-r}}{s!} \\ &= t \sum_r \binom{k}{r} M_{k-r} \end{aligned}$$

As for the initial values of these moments, we know from Proposition 2.24 and its proof that these are given by the following formulae:

$$M_1 = t \quad , \quad M_2 = t + t^2$$

On the other hand, consider the numbers in the statement, namely:

$$S_k = \sum_{\pi \in P(k)} t^{|\pi|}$$

Since a partition of $\{1, \dots, k+1\}$ appears by choosing r neighbors for 1, among the k numbers available, and then partitioning the $k-r$ elements left, we have:

$$S_{k+1} = t \sum_r \binom{k}{r} S_{k-r}$$

As for the initial values of these numbers, these are as follows:

$$S_1 = t \quad , \quad S_2 = t + t^2$$

Thus the initial values coincide, so these numbers are the moments, as stated. \square

2e. Exercises

We had a lot on interesting theory in this chapter, and as exercises, we have:

EXERCISE 2.33. *Compute all the probabilities at poker.*

EXERCISE 2.34. *Learn more about the Dirac masses, from the physicists.*

EXERCISE 2.35. *Meditate on the variance V , and its probabilistic meaning.*

EXERCISE 2.36. *Meditate on M_3, M_4, \dots , and their probabilistic meaning.*

EXERCISE 2.37. *Develop some mathematics for the formal exponential formula for p_t .*

EXERCISE 2.38. *Learn more about the PLT, and its probabilistic consequences.*

EXERCISE 2.39. *Work out, using calculus, the asymptotics of the Bell numbers.*

EXERCISE 2.40. *Fill in the missing details for the moments of p_t , with $t > 0$.*

As bonus exercise, set up a small business, with the theory that you learned here.

CHAPTER 3

Combinatorics

3a. Group theory

We have seen so far that, in what regards the discrete laws, passed a lot of general theory, that you should of course master well, and don't hesitate here to learn and relearn more, and do more exercises, it all comes to combinatorics. In fact, and believe me here, at the advanced level, probability and combinatorics are the same thing.

With this idea in mind, let us examine more in detail the relation between probability and combinatorics. As a starting point, we have the following definition:

DEFINITION 3.1. *A group is a set G with a multiplication operation $(g, h) \rightarrow gh$, which must satisfy the following conditions:*

- (1) *Associativity: we have $(gh)k = g(hk)$, for any $g, h, k \in G$.*
- (2) *Unit: there is an element $1 \in G$ such that $g1 = 1g = g$, for any $g \in G$.*
- (3) *Inverses: for any $g \in G$ there is $g^{-1} \in G$ such that $gg^{-1} = g^{-1}g = 1$.*

When the multiplication is commutative, $gh = hg$, we say that G is abelian.

As a first observation, with this notion in hand, we can now properly talk about the convolution of measures, provided that our probability space is a group G , or at least a semigroup, and with the best results being obtained when G is an abelian group, with a basic example here being $G = \mathbb{R}$, or $G = \mathbb{R}^N$. Let us record this finding as follows:

PROPOSITION 3.2. *When our probability space is an abelian group, we have:*

$$(\delta_x * \delta_y) * \delta_z = \delta_x * (\delta_y * \delta_z)$$

$$\delta_x * \delta_0 = \delta_0 * \delta_x = \delta_x$$

$$\delta_x * \delta_{-x} = \delta_{-x} * \delta_x = \delta_0$$

$$\delta_x * \delta_y = \delta_y * \delta_x$$

When our probability space is a non-abelian group, only the first three formulae hold.

PROOF. This is just a reformulation of the group axioms, from Definition 3.1, in the abelian case, and then in general, in terms of Dirac masses, and with the group elements denoted x, y, z, \dots , and the unit denoted, as usual in the abelian case, by 0. \square

However, it is not about such things, which remain quite abstract, that we would like to talk about, in this chapter. In order to get to truly exciting connections between probability and combinatorics, let us get back to Definition 3.1, as stated, and see where that definition brings us. That is, let us develop some group theory, somehow with focus on combinatorial aspects, and for connections with probability, we will see later.

Getting started with our program, let us first discuss the finite abelian groups, which are the simplest ones. As basic examples, we have the cyclic groups:

PROPOSITION 3.3. *The following constructions produce the same group, denoted \mathbb{Z}_N , which is finite and abelian, and is called cyclic group of order N :*

- (1) \mathbb{Z}_N is the set of remainders modulo N , with operation $+$.
- (2) $\mathbb{Z}_N \subset \mathbb{T}$ is the group of N -th roots of unity, with operation \times .

PROOF. Here the equivalence between (1) and (2) is obvious. More complicated, however, is the question of finding the good philosophy and notation for this group. In what concerns us, we will use the notation (1), and often prefer the interpretation (2). \square

We can construct further examples by making products between various cyclic groups \mathbb{Z}_N . Quite remarkably, we obtain in this way all finite abelian groups:

THEOREM 3.4. *The finite abelian groups are precisely the products of cyclic groups:*

$$G = \mathbb{Z}_{N_1} \times \dots \times \mathbb{Z}_{N_k}$$

Moreover, there are technical extensions of this result, going beyond the finite case.

PROOF. This is something quite tricky, the idea being as follows:

(1) In order to prove our result, assume that G is finite and abelian. For any prime number $p \in \mathbb{N}$, let us define $G_p \subset G$ to be the subset of elements having as order a power of p . Equivalently, this subset $G_p \subset G$ can be defined as follows:

$$G_p = \left\{ g \in G \mid \exists k \in \mathbb{N}, g^{p^k} = 1 \right\}$$

(2) It is then routine to check, based on definitions, that each G_p is a subgroup. Our claim now is that we have a direct product decomposition as follows:

$$G = \prod_p G_p$$

(3) Indeed, by using the fact that our group G is abelian, we have a morphism as follows, with the order of the factors when computing $\prod_p g_p$ being irrelevant:

$$\prod_p G_p \rightarrow G \quad , \quad (g_p) \rightarrow \prod_p g_p$$

Moreover, it is routine to check that this morphism is both injective and surjective, via some simple manipulations, so we have our group decomposition, as in (2).

(4) Thus, we are left with proving that each component G_p decomposes as a product of cyclic groups, having as orders powers of p , as follows:

$$G_p = \mathbb{Z}_{p^{r_1}} \times \dots \times \mathbb{Z}_{p^{r_s}}$$

But this is something that can be checked by recurrence on $|G_p|$, via some routine computations, and we are led to the conclusion in the statement. See Lang [64]. \square

Moving away now from finite or abelian groups, the next thought goes to matrices. So, let us call $GL_N(\mathbb{C})$ the group formed by the $N \times N$ matrices having nonzero determinant, with GL standing for “general linear”. By further imposing the condition $\det A = 1$ we obtain a subgroup $SL_N(\mathbb{C})$, with SL standing for “special linear”, and then we can talk as well about the real versions of these groups, and also intersect everything with the group of unitary matrices U_N . We obtain in this way 8 groups, as follows:

THEOREM 3.5. *We have groups of invertible matrices as follows,*

$$\begin{array}{ccccc}
 & & GL_N(\mathbb{R}) & \longrightarrow & GL_N(\mathbb{C}) \\
 & \nearrow & \uparrow & & \nearrow \\
 O_N & \longrightarrow & & \longrightarrow & U_N \\
 \uparrow & & \uparrow & & \uparrow \\
 & & SL_N(\mathbb{R}) & \longrightarrow & SL_N(\mathbb{C}) \\
 & \nearrow & \uparrow & & \nearrow \\
 SO_N & \longrightarrow & & \longrightarrow & SU_N
 \end{array}$$

with S standing here for “special”, meaning having determinant 1.

PROOF. This is clear indeed from the above discussion. As a comment, we can talk in fact about $GL_N(F)$ and $SL_N(F)$, once we have a ground field F , but in what regards the corresponding orthogonal and unitary groups, things here are more complicated. \square

As a continuation, let us discuss now the finite non-abelian groups. As basic examples, we have the symmetric group S_N , and its various subgroups. Let us record here:

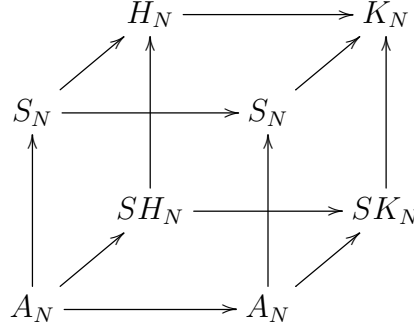
PROPOSITION 3.6. *We have finite non-abelian groups, as follows:*

- (1) S_N , the group of permutations of $\{1, \dots, N\}$.
- (2) $A_N \subset S_N$, the permutations having signature 1.
- (3) $D_N \subset S_N$, the group of symmetries of the regular N -gon.

PROOF. The fact that we have indeed groups is clear from definitions, and the non-abelianity of these groups is clear as well, provided of course that in each case N is chosen big enough, and with exercise for you to work out all this, with full details. \square

For constructing further examples of finite non-abelian groups, the best is to “look up”, by regarding S_N as being the permutation group of the N coordinate axes of \mathbb{R}^N . Indeed, this suggests looking at the symmetry groups of all sorts of geometric beasts inside \mathbb{R}^N , or even \mathbb{C}^N , and we end with a whole menagerie of groups, as follows:

THEOREM 3.7. *We have groups of unitary matrices as follows,*



for the most finite, and non-abelian, called complex reflection groups.

PROOF. The above statement is of course something informal, and here are explanations on all this, including definitions for all the groups involved:

(1) To start with, S_N is the symmetric group S_N that we know, but regarded now as permutation group of the N coordinate axes of \mathbb{R}^N , and so as subgroup $S_N \subset O_N$.

(2) Similarly, A_N is the alternating group A_N that we know, but coming now geometrically, as $A_N = S_N \cap SO_N$, with the intersection being computed inside O_N .

(3) Regarding $H_N \subset O_N$, this is a famous group, called hyperoctahedral group, appearing as the symmetry group of the hypercube $\square_N \subset \mathbb{R}^N$.

(4) Regarding $K_N \subset U_N$, this is the complex analogue of H_N , consisting of the unitary matrices $U \in U_N$ having exactly one nonzero entry, on each row and each column.

(5) We have as well on our diagram the groups SH_N, SK_N , with S standing as usual for “special”, that is, consisting of the matrices in H_N, K_N having determinant 1.

(6) In what regards now the diagram itself, sure I can see that S_N, A_N appear twice, but nothing can be done here, after thinking a bit, at how the diagram works.

(7) Let us mention too that the groups \mathbb{Z}_N, D_N have their place here, in N -dimensional geometry, but not exactly on our diagram, as being the symmetry groups of the oriented cycle, and unoriented cycle, with vertices at the simplex $X_N = \{e_i\} \subset \mathbb{R}^N$.

(8) Finally, in what regards finiteness, non-abelianity, and also the name “complex reflection groups”, many things to be checked here, left to you as an exercise. \square

Very nice all this. As a conclusion, which is something quite interesting, we have:

CONCLUSION 3.8. *Most groups are best seen as being groups of matrices,*

$$G \subset GL_N(\mathbb{C})$$

and even as groups of unitary matrices, $G \subset U_N$, in most cases.

So long for abstract group theory, and we will be back to this, on numerous occasions, in what follows. Going now towards probability, with somehow the above conclusion in mind, let us start with the following definition, putting everything on a solid basis:

DEFINITION 3.9. *A unitary representation of a compact group G is a morphism*

$$u : G \rightarrow U_N \quad , \quad g \rightarrow u_g$$

which can be faithful or not. The character of such a representation is the function

$$\chi_u : G \rightarrow \mathbb{C} \quad , \quad g \rightarrow \text{Tr}(u_g)$$

where Tr is the usual, unnormalized trace of the $N \times N$ matrices.

At the level of examples, most of the compact groups that we met so far naturally appear as subgroups $G \subset U_N$. In this case, the embedding $G \subset U_N$ is of course a representation, called fundamental representation. Many other examples of representations can be constructed, by using various operations for the representations, and we will be back to this, with details, in chapter 14 below, when systematically discussing group theory.

Getting now to the point, the character $\chi_u : G \rightarrow \mathbb{C}$ of a representation $u : G \rightarrow U_N$ can be regarded as a random variable, and the advanced theory of compact groups, which is something that we will discuss later, in chapter 14 below, leads us into computing the law of χ_u . Moreover, this problem is of particular interest when u is the fundamental representation of a matrix group $G \subset U_N$, so we are led into the following question:

QUESTION 3.10. *Given a compact group $G \subset U_N$, what is the law of*

$$\chi : G \rightarrow \mathbb{C} \quad , \quad g \rightarrow \text{Tr}(g)$$

with respect to the uniform, mass 1 measure on G ?

Observe that this latter question is something elementary, save for some details with the uniform measure on G , whose existence is something non-trivial. But, in what follows we will mostly deal with this question for the finite groups G , where the measure in question is simply the counting one. So, as a conclusion to this, pick $G \subset U_N$ finite, solve Question 3.10 for it, and the advanced theory of finite groups guarantees that you are into very interesting mathematics. We will see in what follows that this is indeed the case.

3b. Derangements

Getting back now to probability, we would like to investigate Question 3.10 for the various reflection groups from Theorem 3.7. For this purpose, we will need:

PROPOSITION 3.11. *We have the following formula,*

$$\left| \left(\bigcup_i A_i \right)^c \right| = |A| - \sum_i |A_i| + \sum_{i < j} |A_i \cap A_j| - \sum_{i < j < k} |A_i \cap A_j \cap A_k| + \dots$$

called inclusion-exclusion principle.

PROOF. This is quite clear, by thinking a bit, as follows:

- (1) In order to count $(\cup_i A_i)^c$, we certainly have to start with $|A|$.
- (2) Then, we obviously have to remove each $|A_i|$, and so remove $\sum_i |A_i|$.
- (3) But then, we have to put back each $|A_i \cap A_j|$, and so put back $\sum_{i < j} |A_i \cap A_j|$.
- (4) Then, we must remove each $|A_i \cap A_j \cap A_k|$, so remove $\sum_{i < j < k} |A_i \cap A_j \cap A_k|$.
- \vdots
- (5) And so on, which leads to the formula in the statement. □

We can now formulate the following key result, regarding the symmetric group S_N , making the link with the Poisson laws, that we met in chapter 2:

THEOREM 3.12. *The probability for a random $\sigma \in S_N$ to have no fixed points is*

$$P \simeq \frac{1}{e}$$

in the $N \rightarrow \infty$ limit, where $e = 2.7182\dots$ is the usual constant from analysis. More generally, the main character of S_N , which counts the fixed points, and is given by

$$\chi = \sum_i \sigma_{ii}$$

via the standard embedding $S_N \subset O_N$, follows the Poisson law p_1 , in the $N \rightarrow \infty$ limit. Even more generally, the truncated characters of S_N , given by

$$\chi_t = \sum_{i=1}^{[tN]} \sigma_{ii}$$

with $t \in (0, 1]$, follow the Poisson laws p_t , in the $N \rightarrow \infty$ limit.

PROOF. Many things going on here, the idea being as follows:

(1) In order to prove the first assertion, which is the key, and probably the most puzzling one, we will use the inclusion-exclusion principle. Let us set:

$$S_N^k = \left\{ \sigma \in S_N \mid \sigma(k) = k \right\}$$

The set of permutations having no fixed points, called derangements, is then:

$$X_N = \left(\bigcup_k S_N^k \right)^c$$

Now the inclusion-exclusion principle tells us that we have:

$$\begin{aligned} |X_N| &= \left| \left(\bigcup_k S_N^k \right)^c \right| \\ &= |S_N| - \sum_k |S_N^k| + \sum_{k < l} |S_N^k \cap S_N^l| - \dots + (-1)^N \sum_{k_1 < \dots < k_N} |S_N^{k_1} \cup \dots \cup S_N^{k_N}| \\ &= N! - N(N-1)! + \binom{N}{2}(N-2)! - \dots + (-1)^N \binom{N}{N}(N-N)! \\ &= \sum_{r=0}^N (-1)^r \binom{N}{r} (N-r)! \end{aligned}$$

Thus, the probability that we are interested in, for a random permutation $\sigma \in S_N$ to have no fixed points, is given by the following formula:

$$P = \frac{|X_N|}{N!} = \sum_{r=0}^N \frac{(-1)^r}{r!}$$

Since on the right we have the expansion of $1/e$, this gives the result.

(2) Let us construct now the main character of S_N , as in the statement. The permutation matrices being given by $\sigma_{ij} = \delta_{i\sigma(j)}$, we have the following formula:

$$\chi(\sigma) = \sum_i \delta_{\sigma(i)i} = \# \left\{ i \in \{1, \dots, N\} \mid \sigma(i) = i \right\}$$

In order to establish now the asymptotic result in the statement, regarding these characters, we must prove the following formula, for any $r \in \mathbb{N}$, in the $N \rightarrow \infty$ limit:

$$P(\chi = r) \simeq \frac{1}{r!e}$$

We already know, from (1), that this formula holds at $r = 0$. In the general case now, we have to count the permutations $\sigma \in S_N$ having exactly r points. Now since having

such a permutation amounts in choosing r points among $1, \dots, N$, and then permuting the $N - r$ points left, without fixed points allowed, we have:

$$\begin{aligned} \# \left\{ \sigma \in S_N \mid \chi(\sigma) = r \right\} &= \binom{N}{r} \# \left\{ \sigma \in S_{N-r} \mid \chi(\sigma) = 0 \right\} \\ &= \frac{N!}{r!(N-r)!} \# \left\{ \sigma \in S_{N-r} \mid \chi(\sigma) = 0 \right\} \\ &= N! \times \frac{1}{r!} \times \frac{\# \left\{ \sigma \in S_{N-r} \mid \chi(\sigma) = 0 \right\}}{(N-r)!} \end{aligned}$$

By dividing everything by $N!$, we obtain from this the following formula:

$$\frac{\# \left\{ \sigma \in S_N \mid \chi(\sigma) = r \right\}}{N!} = \frac{1}{r!} \times \frac{\# \left\{ \sigma \in S_{N-r} \mid \chi(\sigma) = 0 \right\}}{(N-r)!}$$

Now by using the computation at $r = 0$, that we already have, from (1), it follows that with $N \rightarrow \infty$ we have the following estimate:

$$P(\chi = r) \simeq \frac{1}{r!} \cdot P(\chi = 0) \simeq \frac{1}{r!} \cdot \frac{1}{e}$$

Thus, we obtain as limiting measure the Poisson law of parameter 1, as stated.

(3) Finally, let us construct the truncated characters of S_N , as in the statement. As before in the case $t = 1$, we have the following computation, coming from definitions:

$$\chi_t(\sigma) = \sum_{i=1}^{[tN]} \delta_{\sigma(i)i} = \# \left\{ i \in \{1, \dots, [tN]\} \mid \sigma(i) = i \right\}$$

Also before in the proofs of (1) and (2), we obtain by inclusion-exclusion that:

$$\begin{aligned} P(\chi_t = 0) &= \frac{1}{N!} \sum_{r=0}^{[tN]} (-1)^r \sum_{k_1 < \dots < k_r < [tN]} |S_N^{k_1} \cap \dots \cap S_N^{k_r}| \\ &= \frac{1}{N!} \sum_{r=0}^{[tN]} (-1)^r \binom{[tN]}{r} (N-r)! \\ &= \sum_{r=0}^{[tN]} \frac{(-1)^r}{r!} \cdot \frac{[tN]!(N-r)!}{N!([tN]-r)!} \end{aligned}$$

Now with $N \rightarrow \infty$, we obtain from this the following estimate:

$$P(\chi_t = 0) \simeq \sum_{r=0}^{[tN]} \frac{(-1)^r}{r!} \cdot t^r \simeq e^{-t}$$

More generally, by counting the permutations $\sigma \in S_N$ having exactly r fixed points among $1, \dots, [tN]$, as in the proof of (2), we obtain:

$$P(\chi_t = r) \simeq \frac{t^r}{r!e^t}$$

Thus, we obtain in the limit a Poisson law of parameter t , as stated. \square

As a first comment, we can approach the above problems as well directly, by using:

PROPOSITION 3.13. *The arbitrary integrals over the symmetric group S_N are given, modulo linearity, by the formula*

$$\int_{S_N} g_{i_1 j_1} \cdots g_{i_k j_k} = \begin{cases} \frac{(N - |\ker i|)!}{N!} & \text{if } \ker i = \ker j \\ 0 & \text{otherwise} \end{cases}$$

where $\ker i$ denotes as usual the partition of $\{1, \dots, k\}$ whose blocks collect the equal indices of i , and where $|\cdot|$ denotes the number of blocks.

PROOF. The first assertion follows from the Stone-Weierstrass theorem, because the standard coordinates g_{ij} separate the points of S_N , and so we have:

$$\langle g_{ij} \rangle = C(S_N)$$

Regarding now the second assertion, according to the definition of the matrix coordinates g_{ij} , the integrals in the statement are given by:

$$\int_{S_N} g_{i_1 j_1} \cdots g_{i_k j_k} = \frac{1}{N!} \# \left\{ \sigma \in S_N \mid \sigma(j_1) = i_1, \dots, \sigma(j_k) = i_k \right\}$$

Now observe that the existence of $\sigma \in S_N$ as above requires $i_m = i_n \iff j_m = j_n$. Thus, the above integral vanishes when the following condition is satisfied:

$$\ker i \neq \ker j$$

Regarding now the case $\ker i = \ker j$, if we denote by $b \in \{1, \dots, k\}$ the number of blocks of this partition $\ker i = \ker j$, we have $N - b$ points to be sent bijectively to $N - b$ points, and so $(N - b)!$ solutions, and the integral is $\frac{(N - b)!}{N!}$, as claimed. \square

We can now recover the computation of the asymptotic law of χ_t , as follows:

THEOREM 3.14. *For the symmetric group $S_N \subset O_N$, regarded as a compact group of matrices, $S_N \subset O_N$, via the standard permutation matrices, the truncated character*

$$\chi_t(g) = \sum_{i=1}^{[tN]} g_{ii}$$

counts the number of fixed points among $\{1, \dots, [tN]\}$, and its law with respect to the counting measure becomes, with $N \rightarrow \infty$, a Poisson law of parameter t .

PROOF. We already know this, from Theorem 3.12, but we can use as well the integration formula in Proposition 3.13. Indeed, with S_{kb} being the Stirling numbers, counting the partitions of $\{1, \dots, k\}$ having exactly b blocks, we have the following formula:

$$\begin{aligned} \int_{S_N} \chi_t^k &= \sum_{i_1, \dots, i_k=1}^{[tN]} \int_{S_N} g_{i_1 i_1} \cdots g_{i_k i_k} \\ &= \sum_{\pi \in P(k)} \frac{[tN]!}{([tN] - |\pi|)!} \cdot \frac{(N - |\pi|)!}{N!} \\ &= \sum_{b=1}^{[tN]} \frac{[tN]!}{([tN] - b)!} \cdot \frac{(N - b)!}{N!} \cdot S_{kb} \end{aligned}$$

In particular with $N \rightarrow \infty$ we obtain the following formula:

$$\lim_{N \rightarrow \infty} \int_{S_N} \chi_t^k = \sum_{b=1}^k S_{kb} t^b$$

But this is the k -th moment of the Poisson law p_t , and so we are done. \square

As another result now regarding S_N , here is a useful related formula:

THEOREM 3.15. *We have the law formula*

$$\text{law}(g_{11} + \dots + g_{ss}) = \frac{s!}{N!} \sum_{p=0}^s \frac{(N-p)!}{(s-p)!} \cdot \frac{(\delta_1 - \delta_0)^{*p}}{p!}$$

where g_{ij} are the standard coordinates of $S_N \subset O_N$.

PROOF. We have the following moment formula, where m_f is the number of permutations of $\{1, \dots, N\}$ having exactly f fixed points in the set $\{1, \dots, s\}$:

$$\int_{S_N} (u_{11} + \dots + u_{ss})^k = \frac{1}{N!} \sum_{f=0}^s m_f f^k$$

Thus the law in the statement, say ν_{sN} , is the following average of Dirac masses:

$$\nu_{sN} = \frac{1}{N!} \sum_{f=0}^s m_f \delta_f$$

Now observe that the permutations contributing to m_f are obtained by choosing f points in the set $\{1, \dots, s\}$, then by permuting the remaining $N - f$ points in $\{1, \dots, n\}$ in such a way that there is no fixed point in $\{1, \dots, s\}$. But these latter permutations are

counted as follows: we start with all permutations, we subtract those having one fixed point, we add those having two fixed points, and so on. We obtain in this way:

$$\begin{aligned}
\nu_{sN} &= \frac{1}{N!} \sum_{f=0}^s \binom{s}{f} \left(\sum_{k=0}^{s-f} (-1)^k \binom{s-f}{k} (N-f-k)! \right) \delta_f \\
&= \sum_{f=0}^s \sum_{k=0}^{s-f} (-1)^k \frac{1}{N!} \cdot \frac{s!}{f!(s-f)!} \cdot \frac{(s-f)!(N-f-k)!}{k!(s-f-k)!} \delta_f \\
&= \frac{s!}{N!} \sum_{f=0}^s \sum_{k=0}^{s-f} \frac{(-1)^k (N-f-k)!}{f!k!(s-f-k)!} \delta_f
\end{aligned}$$

We can proceed as follows, by using the new index $p = f + k$:

$$\begin{aligned}
\nu_{sN} &= \frac{s!}{N!} \sum_{p=0}^s \sum_{k=0}^p \frac{(-1)^k (N-p)!}{(p-k)!k!(s-p)!} \delta_{p-k} \\
&= \frac{s!}{N!} \sum_{p=0}^s \frac{(N-p)!}{(s-p)!p!} \sum_{k=0}^p (-1)^k \binom{p}{k} \delta_{p-k} \\
&= \frac{s!}{N!} \sum_{p=0}^s \frac{(N-p)!}{(s-p)!} \cdot \frac{(\delta_1 - \delta_0)^{*p}}{p!}
\end{aligned}$$

Here $*$ is convolution of real measures, and the assertion follows. \square

Observe that the above formula is finer than most of our previous formulae, which were asymptotic, because it is valid at any $N \in \mathbb{N}$. We can use this formula as follows:

THEOREM 3.16. *Let g_{ij} be the standard coordinates of $C(S_N)$.*

- (1) $u_{11} + \dots + u_{ss}$ with $s = o(N)$ is a projection of trace s/N .
- (2) $u_{11} + \dots + u_{ss}$ with $s = tN + o(N)$ is Poisson of parameter t .

PROOF. We can use indeed the formula in Theorem 3.15, as follows:

- (1) With s fixed and $N \rightarrow \infty$ we have the following estimate:

$$\begin{aligned}
law(u_{11} + \dots + u_{ss}) &= \sum_{p=0}^s \frac{(N-p)!}{N!} \cdot \frac{s!}{(s-p)!} \cdot \frac{(\delta_1 - \delta_0)^{*p}}{p!} \\
&= \delta_0 + \frac{s}{N} (\delta_1 - \delta_0) + O(N^{-2})
\end{aligned}$$

But the law on the right is that of a projection of trace s/N , as desired.

(2) We have a law formula of the following type:

$$\text{law}(u_{11} + \dots + u_{ss}) = \sum_{p=0}^s c_p \cdot \frac{(\delta_1 - \delta_0)^{*p}}{p!}$$

The coefficients c_p can be estimated by using the Stirling formula, as follows:

$$\begin{aligned} c_p &= \frac{(tN)!}{N!} \cdot \frac{(N-p)!}{(tN-p)!} \\ &\simeq \frac{(tN)^{tN}}{N^N} \cdot \frac{(N-p)^{N-p}}{(tN-p)^{tN-p}} \\ &= \left(\frac{tN}{tN-p} \right)^{tN-p} \left(\frac{N-p}{N} \right)^{N-p} \left(\frac{tN}{N} \right)^p \\ &\simeq e^p e^{-pt^p} \\ &= t^p \end{aligned}$$

We can now compute the Fourier transform with respect to a variable y :

$$F(\text{law}(u_{11} + \dots + u_{ss})) \simeq \sum_{p=0}^s t^p \cdot \frac{(e^y - 1)^p}{p!} = e^{t(e^y - 1)}$$

But this being the Fourier transform of p_t , this gives the second assertion. \square

3c. Bessel laws

In order to further extend our results regarding the Poisson laws, we will need a number of standard probabilistic preliminaries. We have the following notion:

DEFINITION 3.17. *Associated to any discrete positive measure ν on \mathbb{C} , not necessarily of mass 1, is the probability measure*

$$p_\nu = \lim_{n \rightarrow \infty} \left(\left(1 - \frac{t}{n} \right) \delta_0 + \frac{1}{n} \nu \right)^{*n}$$

where $t = \text{mass}(\nu)$, called *compound Poisson law*.

As a basic example, the measure $\nu = t\delta_1$ with $t > 0$ produces the Poisson laws, via the PLT. The following standard result allows us to detect compound Poisson laws:

PROPOSITION 3.18. *For $\nu = \sum_{i=1}^s t_i \delta_{z_i}$ with $t_i > 0$ and $z_i \in \mathbb{C}$, we have*

$$F_{p_\nu}(y) = \exp \left(\sum_{i=1}^s t_i (e^{iyz_i} - 1) \right)$$

where F denotes the Fourier transform.

PROOF. Let η_n be the measure in Definition 3.17, under the convolution sign:

$$\eta_n = \left(1 - \frac{t}{n}\right) \delta_0 + \frac{1}{n} \nu$$

We have then the following computation:

$$\begin{aligned} F_{\eta_n}(y) = \left(1 - \frac{t}{n}\right) + \frac{1}{n} \sum_{i=1}^s t_i e^{iyz_i} &\implies F_{\eta_n^{*n}}(y) = \left(\left(1 - \frac{t}{n}\right) + \frac{1}{n} \sum_{i=1}^s t_i e^{iyz_i}\right)^n \\ &\implies F_{p_\nu}(y) = \exp\left(\sum_{i=1}^s t_i (e^{iyz_i} - 1)\right) \end{aligned}$$

Thus, we have obtained the formula in the statement. \square

We have as well the following result, providing an alternative to Definition 3.17, and which will be our formulation here of the Compound Poisson Limit Theorem:

THEOREM 3.19 (CPLT). *For $\nu = \sum_{i=1}^s t_i \delta_{z_i}$ with $t_i > 0$ and $z_i \in \mathbb{C}$, we have*

$$p_\nu = \text{law} \left(\sum_{i=1}^s z_i \alpha_i \right)$$

where the variables α_i are Poisson (t_i), independent.

PROOF. Let α be the sum of Poisson variables in the statement, namely:

$$\alpha = \sum_{i=1}^s z_i \alpha_i$$

By using our standard Fourier transform formulae, we have:

$$\begin{aligned} F_{\alpha_i}(y) = \exp(t_i(e^{iyz_i} - 1)) &\implies F_{z_i \alpha_i}(y) = \exp(t_i(e^{iyz_i} - 1)) \\ &\implies F_\alpha(y) = \exp\left(\sum_{i=1}^s t_i (e^{iyz_i} - 1)\right) \end{aligned}$$

Thus we have indeed the same formula as in Proposition 3.18, as desired. \square

Getting now to the examples, let us start with the following definition:

DEFINITION 3.20. *The Bessel law of level $s \in \mathbb{N} \cup \{\infty\}$ and parameter $t > 0$ is*

$$b_t^s = p_{t\varepsilon_s}$$

with ε_s being the uniform measure on the s -th roots of unity.

Of particular interest are the cases $s = 1, 2, \infty$, where we obtain the Poisson laws p_t , and then certain measures b_t, B_t , called real and purely complex Bessel laws:

$$b_t^1 = p_t \quad , \quad b_t^2 = b_t \quad , \quad b_t^\infty = B_t$$

As a basic result on the Bessel laws, generalizing those about p_t , we have:

THEOREM 3.21. *The Fourier transform of b_t^s is given by*

$$\log F_t^s(z) = t(\exp_s z - 1)$$

where $\exp_s z$ is the level s exponential function, given by the formula

$$\exp_s z = \sum_{k=0}^{\infty} \frac{z^{sk}}{(sk)!}$$

so in particular the measures b_t^s have the property $b_t^s * b_{t'}^s = b_{t+t'}^s$.

PROOF. We know from Theorem 3.19 that b_t^s appears as follows, with a_1, \dots, a_s being independent, each of them following the Poisson law of parameter t/s , and $w = e^{2\pi i/s}$:

$$b_t^s = \text{law} \left(\sum_{k=1}^s w^k a_k \right)$$

We have the following computation, for the corresponding Fourier transform:

$$\begin{aligned} \log F(z) &= \sum_{k=1}^s \log F_{a_k}(w^k z) \\ &= \sum_{k=1}^s \frac{t}{s} (\exp(w^k z) - 1) \\ &= t \left(\left(\frac{1}{s} \sum_{k=1}^s \exp(w^k z) \right) - 1 \right) \\ &= t(\exp_s z - 1) \end{aligned}$$

Thus, we are led to the conclusions in the statement. □

Let us study now the density of b_t^s . We have here the following result:

THEOREM 3.22. *We have the formula*

$$b_t^s = e^{-t} \sum_{p_1=0}^{\infty} \dots \sum_{p_s=0}^{\infty} \frac{1}{p_1! \dots p_s!} \left(\frac{t}{s} \right)^{p_1 + \dots + p_s} \delta \left(\sum_{k=1}^s w^k p_k \right)$$

where $w = e^{2\pi i/s}$, and the δ symbol is a Dirac mass.

PROOF. The Fourier transform of the measure on the right is given by:

$$\begin{aligned} F(z) &= e^{-t} \sum_{p_1=0}^{\infty} \dots \sum_{p_s=0}^{\infty} \frac{1}{p_1! \dots p_s!} \left(\frac{t}{s}\right)^{p_1+\dots+p_s} \exp\left(\sum_{k=1}^s w^k p_k z\right) \\ &= e^{-t} \sum_{r=0}^{\infty} \left(\frac{t}{s}\right)^r \sum_{\Sigma p_i=r} \frac{\exp\left(\sum_{k=1}^s w^k p_k z\right)}{p_1! \dots p_s!} \end{aligned}$$

We multiply now by e^t , and we compute the derivative with respect to t :

$$\begin{aligned} (e^t F(z))' &= \sum_{r=1}^{\infty} \frac{r}{s} \left(\frac{t}{s}\right)^{r-1} \sum_{\Sigma p_i=r} \frac{\exp\left(\sum_{k=1}^s w^k p_k z\right)}{p_1! \dots p_s!} \\ &= \frac{1}{s} \sum_{r=1}^{\infty} \left(\frac{t}{s}\right)^{r-1} \sum_{\Sigma p_i=r} \sum_{l=1}^s \frac{\exp\left(\sum_{k=1}^s w^k p_k z\right)}{p_1! \dots p_{l-1}! (p_l-1)! p_{l+1}! \dots p_s!} \end{aligned}$$

By using the variable $u = r - 1$, we obtain from this the following formula:

$$\begin{aligned} (e^t F(z))' &= \frac{1}{s} \sum_{u=0}^{\infty} \left(\frac{t}{s}\right)^u \sum_{\Sigma q_i=u} \sum_{l=1}^s \frac{\exp\left(w^l z + \sum_{k=1}^s w^k q_k z\right)}{q_1! \dots q_s!} \\ &= \left(\frac{1}{s} \sum_{l=1}^s \exp(w^l z)\right) \left(\sum_{u=0}^{\infty} \left(\frac{t}{s}\right)^u \sum_{\Sigma q_i=u} \frac{\exp\left(\sum_{k=1}^s w^k q_k z\right)}{q_1! \dots q_s!}\right) \\ &= (\exp_s z)(e^t F(z)) \end{aligned}$$

But this gives $\log F = t(\exp_s z - 1)$, as in Theorem 3.21, as desired. \square

In analogy now to what we did before involving S_N , let us start with:

PROPOSITION 3.23. *Consider the hyperoctahedral group $H_N \subset O_N$, consisting of the symmetries of the standard hypercube in \mathbb{R}^N .*

- (1) H_N is the symmetry group of the N coordinate axes of \mathbb{R}^N .
- (2) H_N consists of the permutation-like matrices over $\{-1, 0, 1\}$.
- (3) We have the cardinality formula $|H_N| = 2^N N!$.
- (4) We have a crossed product decomposition $H_N = S_N \rtimes \mathbb{Z}_2^N$.
- (5) We have a wreath product decomposition $H_N = \mathbb{Z}_2 \wr S_N$.

PROOF. This is routine algebra, or perhaps geometry, as follows:

(1) This is indeed clear from definitions.

(2) Each of the permutations $\sigma \in S_N$ of the N coordinate axes of \mathbb{R}^N can be further “decorated” by a sign vector $\varepsilon \in \{\pm 1\}^N$, consisting of the possible ± 1 flips which can be applied to each coordinate axis, at the arrival. In matrix terms, this gives the result.

(3) We have indeed, using (2), the formula $|H_N| = |S_N| \cdot |\mathbb{Z}_2^N| = N! \cdot 2^N$.

(4) We know from (3) that at the level of cardinalities we have $|H_N| = |S_N \times \mathbb{Z}_2^N|$, and with a bit more work, we obtain that we have $H_N = S_N \rtimes \mathbb{Z}_2^N$, as claimed.

(5) This is simply a reformulation of (4), in terms of wreath products. \square

By doing now character computations, a bit as for S_N , we are led to:

THEOREM 3.24. *For the hyperoctahedral group $H_N \subset O_N$, the law of the truncated character $\chi = g_{11} + \dots + g_{ss}$ with $s = [tN]$ is, in the $N \rightarrow \infty$ limit, the measure*

$$b_t = e^{-t} \sum_{r=-\infty}^{\infty} \delta_r \sum_{p=0}^{\infty} \frac{(t/2)^{|r|+2p}}{(|r|+p)!p!}$$

which is the real Bessel law of parameter $t > 0$.

PROOF. We regard H_N as being the symmetry group of the graph $I_N = \{I^1, \dots, I^N\}$ formed by n segments. The diagonal coefficients are then given by:

$$u_{ii}(g) = \begin{cases} 0 & \text{if } g \text{ moves } I^i \\ 1 & \text{if } g \text{ fixes } I^i \\ -1 & \text{if } g \text{ returns } I^i \end{cases}$$

Let us denote by F_g, R_g the number of segments among $\{I^1, \dots, I^s\}$ which are fixed, respectively returned by an element $g \in H_N$. With this notation, we have:

$$u_{11} + \dots + u_{ss} = F_g - R_g$$

We denote by P_N probabilities computed over H_N . The density of the law of the variable $u_{11} + \dots + u_{ss}$ at a point $r \geq 0$ is then given by the following formula:

$$D(r) = P_N(F_g - R_g = r) = \sum_{p=0}^{\infty} P_N(F_g = r + p, R_g = p)$$

Thus, the asymptotic density can be computed as follows:

$$\begin{aligned} \lim_{N \rightarrow \infty} D(r) &= \lim_{N \rightarrow \infty} \sum_{p=0}^{\infty} (1/2)^{r+2p} \binom{r+2p}{r+p} P_N(F_g + R_g = r + 2p) \\ &= \sum_{p=0}^{\infty} (1/2)^{r+2p} \binom{r+2p}{r+p} \frac{t^{r+2p}}{e^t (r+2p)!} \\ &= e^{-t} \sum_{p=0}^{\infty} \frac{(t/2)^{r+2p}}{(r+p)!p!} \end{aligned}$$

Together with $D(-r) = D(r)$, this gives the formula in the statement. As for the last assertion, this follows from the Fourier transform formula found above. \square

In order now to unify and extend our various results, we will need:

DEFINITION 3.25. *The complex reflection group $H_N^s \subset U_N$ is the group of permutations of N copies of the s -simplex. Equivalently, we have*

$$H_N^s = M_N(\mathbb{Z}_s \cup \{0\}) \cap U_N$$

telling us that H_N^s consists of the permutation-type matrices with s -th roots of unity as entries. Also equivalently, we have the formula $H_N^s = \mathbb{Z}_s \wr S_N$.

To be more precise, the equivalence between the various viewpoints on H_N^s comes as in Proposition 3.23, which corresponds to the case $s = 2$. In fact, we already met in Theorem 3.7 some of these groups, namely S_N, H_N, K_N , appearing at $s = 1, 2, \infty$.

Getting now to probability computations, we have here the following result:

THEOREM 3.26. *For the complex reflection group H_N^s we have, with $N \rightarrow \infty$:*

$$\chi_t \sim b_t^s$$

Moreover, the asymptotic moments of this variable are the numbers

$$M_k(b_t^s) = \sum_{\pi \in P^s(k)} t^{|\pi|}$$

where $P^s(k)$ are the partitions of $\{1, \dots, k\}$ satisfying $\# \circ = \# \bullet (s)$, in each block.

PROOF. This is something quite technical, the idea being as follows:

(1) At $s = 1$ the reflection group is $H_N^1 = S_N$, the Bessel law is the Poisson law, $b_t^1 = p_t$, and the formula $\chi_t \sim p_t$ with $N \rightarrow \infty$ is something that we know. As for the moment formula, where $P^1 = P$, this is something that we know too.

(2) At $s = 2$ the reflection group is $H_N^2 = H_N$, the Bessel law is $b_t^2 = b_t$, and the formula $\chi_t \sim b_t$ with $N \rightarrow \infty$ is something that we know. As for the moment formula, where $P^2 = P_{\text{even}}$, this is something more technical, which can be established too.

(3) At $s = \infty$ the reflection group is $H_N^\infty = K_N$, the Bessel law is $b_t^\infty = B_t$, and the formula $\chi_t \sim B_t$ with $N \rightarrow \infty$ is something that can be proved as for S_N, H_N . As for the moment formula, where $P^\infty = P_{\text{even}}$, this can be established too.

(4) In the general case, $s \in \mathbb{N} \cup \{\infty\}$, the formula $\chi_t \sim b_t^s$ with $N \rightarrow \infty$ can be established like for S_N, H_N , and the moment formula is something more technical. For details on all this, and for the whole story, you can have a look at my book [9]. \square

3d. Cumulants

We have seen a lot of interesting combinatorics in this chapter, but this is not the end of the story. Following Rota, let us formulate now the following definition:

DEFINITION 3.27. *Associated to any real probability measure $\mu = \mu_f$ is the following modification of the logarithm of the Fourier transform $F_\mu(\xi) = E(e^{i\xi f})$,*

$$K_\mu(\xi) = \log E(e^{i\xi f})$$

called cumulant-generating function. The Taylor coefficients $k_n(\mu)$ of this series, given by

$$K_\mu(\xi) = \sum_{n=1}^{\infty} k_n(\mu) \frac{\xi^n}{n!}$$

are called cumulants of the measure μ . We also use the notations k_f, K_f for these cumulants and their generating series, where f is a variable following the law μ .

In other words, the cumulants are the coefficients of $\log F_\mu$, up to some normalizations. As a first observation, the sequence of cumulants k_1, k_2, k_3, \dots appears as a modification of the sequence of moments M_1, M_2, M_3, \dots , the numerics being as follows:

PROPOSITION 3.28. *The sequence of cumulants k_1, k_2, k_3, \dots appears as a modification of the sequence of moments M_1, M_2, M_3, \dots , and uniquely determines μ . We have*

$$k_1 = M_1$$

$$k_2 = -M_1^2 + M_2$$

$$k_3 = 2M_1^3 - 3M_1M_2 + M_3$$

$$k_4 = -6M_1^4 + 12M_1^2M_2 - 3M_2^2 - 4M_1M_3 + M_4$$

$$\vdots$$

in one sense, and in the other sense we have

$$M_1 = k_1$$

$$M_2 = k_1^2 + k_2$$

$$M_3 = k_1^3 + 3k_1k_2 + k_3$$

$$M_4 = k_1^4 + 6k_1^2k_2 + 3k_2^2 + 4k_1k_3 + k_4$$

$$\vdots$$

with in both cases the correspondence being polynomial, with integer coefficients.

PROOF. We know from Definition 3.27 that the cumulants are given by:

$$\log E(e^{\xi f}) = \sum_{s=1}^{\infty} k_s(f) \frac{\xi^s}{s!}$$

By exponentiating, we obtain from this the following formula:

$$E(e^{\xi f}) = \exp \left(\sum_{s=1}^{\infty} k_s(f) \frac{\xi^s}{s!} \right)$$

Now by looking at the terms of order 1, 2, 3, 4, this gives the above formulae. \square

The interest in cumulants comes from the fact that $\log F_\mu$, and so the cumulants $k_n(\mu)$ too, linearize the convolution. To be more precise, we have the following result:

THEOREM 3.29. *The cumulants have the following properties:*

- (1) $k_n(cf) = c^n k_n(f)$.
- (2) $k_1(f+d) = k_1(f) + d$, and $k_n(f+d) = k_n(f)$ for $n > 1$.
- (3) $k_n(f+g) = k_n(f) + k_n(g)$, if f, g are independent.

PROOF. Here (1) and (2) are both clear from definitions, because we have:

$$\begin{aligned} K_{cf+d}(\xi) &= \log E(e^{\xi(cf+d)}) \\ &= \log[e^{\xi d} \cdot E(e^{\xi cf})] \\ &= \xi d + K_f(c\xi) \end{aligned}$$

As for (3), this follows from the fact that the Fourier transform $F_f(\xi) = E(e^{i\xi f})$ satisfies the following formula, whenever f, g are independent random variables:

$$F_{f+g}(\xi) = F_f(\xi)F_g(\xi)$$

Indeed, by applying the logarithm, we obtain the following formula:

$$\log F_{f+g}(\xi) = \log F_f(\xi) + \log F_g(\xi)$$

With the change of variables $\xi \rightarrow -i\xi$, we obtain the following formula:

$$K_{f+g}(\xi) = K_f(\xi) + K_g(\xi)$$

Thus, at the level of coefficients, we obtain $k_n(f+g) = k_n(f) + k_n(g)$, as claimed. \square

At the level of examples now, we have the following key result:

THEOREM 3.30. *For a compound Poisson law p_ν we have*

$$k_n(p_\nu) = M_n(\nu)$$

valid for any integer $n \geq 1$.

PROOF. We can assume, by using a continuity argument, that our measure ν is discrete, as follows, with $t_i > 0$ and $z_i \in \mathbb{R}$, and with the sum being finite:

$$\nu = \sum_i t_i \delta_{z_i}$$

By using now the Fourier transform formula for p_ν , we obtain:

$$\begin{aligned} K_{p_\nu}(\xi) &= \log F_{p_\nu}(-i\xi) \\ &= \log \exp \left[\sum_i t_i (e^{\xi z_i} - 1) \right] \\ &= \sum_i t_i \sum_{n \geq 1} \frac{(\xi z_i)^n}{n!} \\ &= \sum_{n \geq 1} \frac{\xi^n}{n!} \sum_i t_i z_i^n \\ &= \sum_{n \geq 1} \frac{\xi^n}{n!} M_n(\nu) \end{aligned}$$

Thus, we are led to the conclusion in the statement. \square

Getting back to theory now, the sequence of cumulants k_1, k_2, k_3, \dots appears as a modification of the sequence of moments M_1, M_2, M_3, \dots , and understanding the relation between moments and cumulants will be our next task. Let us start with:

DEFINITION 3.31. *The Möbius function of any lattice, and so of P , is given by*

$$\mu(\pi, \nu) = \begin{cases} 1 & \text{if } \pi = \nu \\ -\sum_{\pi \leq \tau < \nu} \mu(\pi, \tau) & \text{if } \pi < \nu \\ 0 & \text{if } \pi \not\leq \nu \end{cases}$$

with the construction being performed by recurrence.

As an illustration here, for $P(2) = \{||, \sqcap\}$, we have by definition:

$$\mu(||, ||) = \mu(\sqcap, \sqcap) = 1$$

Also, $|| < \sqcap$, with no intermediate partition in between, so we obtain:

$$\mu(||, \sqcap) = -\mu(||, ||) = -1$$

Finally, we have $\sqcap \not\leq ||$, and so we have as well the following formula:

$$\mu(\sqcap, ||) = 0$$

The main interest in the Möbius function comes from the Möbius inversion formula, which in linear algebra terms can be stated and proved as follows:

THEOREM 3.32. *We have the following implication,*

$$f(\pi) = \sum_{\nu \leq \pi} g(\nu) \implies g(\pi) = \sum_{\nu \leq \pi} \mu(\nu, \pi) f(\nu)$$

valid for any two functions $f, g : P(n) \rightarrow \mathbb{C}$.

PROOF. Consider the adjacency matrix of P , given by the following formula:

$$A_{\pi\nu} = \begin{cases} 1 & \text{if } \pi \leq \nu \\ 0 & \text{if } \pi \not\leq \nu \end{cases}$$

Our claim is that the inverse of this matrix is the Möbius matrix of P , given by:

$$M_{\pi\nu} = \mu(\pi, \nu)$$

Indeed, the above matrix A is upper triangular, and when trying to invert it, we are led to the recurrence in Definition 3.31, so to the Möbius matrix M . Thus we have:

$$M = A^{-1}$$

Thus, in practice, we are led to the inversion formula in the statement. \square

With these ingredients in hand, let us go back to probability. We first have:

DEFINITION 3.33. *We define quantities $M_\pi(f), k_\pi(f)$, depending on partitions*

$$\pi \in P(k)$$

by starting with $M_n(f), k_n(f)$, and using multiplicativity over the blocks.

To be more precise, the convention here is that for the one-block partition $1_n \in P(n)$, the corresponding moment and cumulant are the usual ones, namely:

$$M_{1_n}(f) = M_n(f) \quad , \quad k_{1_n}(f) = k_n(f)$$

Then, for an arbitrary partition $\pi \in P(k)$, we decompose this partition into blocks, having sizes b_1, \dots, b_s , and we set, by multiplicativity over blocks:

$$M_\pi(f) = M_{b_1}(f) \dots M_{b_s}(f) \quad , \quad k_\pi(f) = k_{b_1}(f) \dots k_{b_s}(f)$$

With this convention, following Rota and others, we can now formulate a key result, fully clarifying the relation between moments and cumulants, as follows:

THEOREM 3.34. *We have the moment-cumulant formulae*

$$M_n(f) = \sum_{\nu \in P(n)} k_\nu(f) \quad , \quad k_n(f) = \sum_{\nu \in P(n)} \mu(\nu, 1_n) M_\nu(f)$$

or, equivalently, we have the moment-cumulant formulae

$$M_\pi(f) = \sum_{\nu \leq \pi} k_\nu(f) \quad , \quad k_\pi(f) = \sum_{\nu \leq \pi} \mu(\nu, \pi) M_\nu(f)$$

where μ is the Möbius function of $P(n)$.

PROOF. There are several things going on here, the idea being as follows:

(1) According to our conventions above, the first set of formulae is equivalent to the second set of formulae. Also, due to the Möbius inversion formula, in the second set of formulae, the two formulae there are in fact equivalent. Thus, the 4 formulae in the statement are all equivalent. In what follows we will focus on the first 2 formulae.

(2) Let us first work out some examples. At $n = 1, 2, 3$ the moment formula gives the following equalities, which are in tune with the findings from Proposition 3.28:

$$M_1 = k_{|} = k_1$$

$$M_2 = k_{||} + k_{\sqcap} = k_1^2 + k_2$$

$$M_3 = k_{|||} + k_{\sqcap|} + k_{\sqcap\sqcap} + k_{|\sqcap} + k_{\sqcap\sqcap} = k_1^3 + 3k_1k_2 + k_3$$

At $n = 4$ now, which is a case which is of particular interest for certain considerations to follow, the computation is as follows, again in tune with Proposition 3.28:

$$\begin{aligned} M_4 &= k_{||||} + \underbrace{(k_{\sqcap||} + \dots)}_{6 \text{ terms}} + \underbrace{(k_{\sqcap\sqcap} + \dots)}_{3 \text{ terms}} + \underbrace{(k_{\sqcap\sqcap|} + \dots)}_{4 \text{ terms}} + k_{\sqcap\sqcap\sqcap} \\ &= k_1^4 + 6k_1^2k_2 + 3k_2^2 + 4k_1k_3 + k_4 \end{aligned}$$

As for the cumulant formula, at $n = 1, 2, 3$ this gives the following formulae for the cumulants, again in tune with the findings from Proposition 3.28:

$$k_1 = M_{|} = M_1$$

$$k_2 = (-1)M_{||} + M_{\sqcap} = -M_1^2 + M_2$$

$$k_3 = 2M_{|||} + (-1)M_{\sqcap|} + (-1)M_{\sqcap\sqcap} + (-1)M_{|\sqcap} + M_{\sqcap\sqcap} = 2M_1^3 - 3M_1M_2 + M_3$$

Finally, at $n = 4$, after computing the Möbius function of $P(4)$, we obtain the following formula for the fourth cumulant, again in tune with Proposition 3.28:

$$\begin{aligned} k_4 &= (-6)M_{||||} + 2\underbrace{(M_{\sqcap||} + \dots)}_{6 \text{ terms}} + (-1)\underbrace{(M_{\sqcap\sqcap} + \dots)}_{3 \text{ terms}} + (-1)\underbrace{(M_{\sqcap\sqcap|} + \dots)}_{4 \text{ terms}} + M_{\sqcap\sqcap\sqcap} \\ &= -6M_1^4 + 12M_1^2M_2 - 3M_2^2 - 4M_1M_3 + M_4 \end{aligned}$$

(3) Time now to get to work, and prove the result. As mentioned above, the formulae in the statement are all equivalent, and it is enough to prove the first one, namely:

$$M_n(f) = \sum_{\nu \in P(n)} k_{\nu}(f)$$

In order to do this, we use the very definition of the cumulants, namely:

$$\log E(e^{\xi f}) = \sum_{s=1}^{\infty} k_s(f) \frac{\xi^s}{s!}$$

By exponentiating, we obtain from this the following formula:

$$E(e^{\xi f}) = \exp \left(\sum_{s=1}^{\infty} k_s(f) \frac{\xi^s}{s!} \right)$$

(4) Let us first compute the function on the left. This is easily done, as follows:

$$E(e^{\xi f}) = E \left(\sum_{n=0}^{\infty} \frac{(\xi f)^n}{n!} \right) = \sum_{n=0}^{\infty} M_n(f) \frac{\xi^n}{n!}$$

(5) Regarding now the function on the right, this is given by:

$$\begin{aligned} \exp \left(\sum_{s=1}^{\infty} k_s(f) \frac{\xi^s}{s!} \right) &= \sum_{p=0}^{\infty} \frac{\left(\sum_{s=1}^{\infty} k_s(f) \frac{\xi^s}{s!} \right)^p}{p!} \\ &= \sum_{p=0}^{\infty} \frac{1}{p!} \sum_{s_1=1}^{\infty} k_{s_1}(f) \frac{\xi^{s_1}}{s_1!} \cdots \sum_{s_p=1}^{\infty} k_{s_p}(f) \frac{\xi^{s_p}}{s_p!} \\ &= \sum_{p=0}^{\infty} \frac{1}{p!} \sum_{s_1=1}^{\infty} \cdots \sum_{s_p=1}^{\infty} k_{s_1}(f) \cdots k_{s_p}(f) \frac{\xi^{s_1+\dots+s_p}}{s_1! \cdots s_p!} \end{aligned}$$

But the point now is that all this leads us into partitions. Indeed, we are summing over indices $s_1, \dots, s_p \in \mathbb{N}$, which can be thought of as corresponding to a partition of $n = s_1 + \dots + s_p$. So, let us rewrite our sum, as a sum over partitions. For this purpose, recall that the number of partitions $\nu \in P(n)$ having blocks of sizes s_1, \dots, s_p is:

$$\binom{n}{s_1, \dots, s_p} = \frac{n!}{p_1! \cdots p_s!}$$

Also, when resumming over partitions, there will be a $p!$ factor as well, coming from the permutations of s_1, \dots, s_p . Thus, our sum can be rewritten as follows:

$$\begin{aligned} \exp \left(\sum_{s=1}^{\infty} k_s(f) \frac{\xi^s}{s!} \right) &= \sum_{n=0}^{\infty} \sum_{p=0}^{\infty} \frac{1}{p!} \sum_{s_1+\dots+s_p=n} k_{s_1}(f) \cdots k_{s_p}(f) \frac{\xi^n}{s_1! \cdots s_p!} \\ &= \sum_{n=0}^{\infty} \frac{\xi^n}{n!} \sum_{p=0}^{\infty} \frac{1}{p!} \sum_{s_1+\dots+s_p=n} \binom{n}{s_1, \dots, s_p} k_{s_1}(f) \cdots k_{s_p}(f) \\ &= \sum_{n=0}^{\infty} \frac{\xi^n}{n!} \sum_{\nu \in P(n)} k_{\nu}(f) \end{aligned}$$

(6) We are now in position to conclude. According to (3,4,5), we have:

$$\sum_{n=0}^{\infty} M_n(f) \frac{\xi^n}{n!} = \sum_{n=0}^{\infty} \frac{\xi^n}{n!} \sum_{\nu \in P(n)} k_{\nu}(f)$$

Thus, we have the following formula, valid for any $n \in \mathbb{N}$:

$$M_n(f) = \sum_{\nu \in P(n)} k_\nu(f)$$

We are therefore led to the conclusions in the statement. \square

Very nice all this, good work that we did in this chapter. We will be back to this, relation between combinatorics and probability, on many occasions, in what follows.

3e. Exercises

Here are some exercises on the above, all combinatorics, of various flavors:

EXERCISE 3.35. *Learn more about finite abelian groups, and Pontrjagin duality.*

EXERCISE 3.36. *Learn more about group representations, and their characters.*

EXERCISE 3.37. *We learned how to compute e , using S_N . What about π ?*

EXERCISE 3.38. *Learn about Bessel functions, of the first and second kind.*

EXERCISE 3.39. *Clarify all the details, for our computation involving H_N .*

EXERCISE 3.40. *Clarify all the details, for our computation involving H_N^s .*

EXERCISE 3.41. *Compute the Möbius matrix for $P(n)$, at $n = 2, 3, 4$.*

EXERCISE 3.42. *Compute more cumulants, for some measures of your choice.*

As bonus exercise, learn some group theory. All beautiful things, and very useful.

CHAPTER 4

Central limits

4a. Central limits

We have seen that some interesting theory can be developed for the discrete measures, notably with a lot of exciting results regarding the Poisson laws, and their versions. Normally this would be the end of what we have to say, in the discrete case, and our job now would be that of getting into the continuous case, first with measure theory systematically developed, and then with some applications of this, to continuous probability.

However, we cannot leave basic probability without talking, in one form or another, about central limits. You have certainly heard about bell-shaped curves, and perhaps even observed them in physics or chemistry class, because any routine measurement leads to such curves. Mathematically, here is the question that we would like to solve:

QUESTION 4.1. *Given random variables f_1, f_2, f_3, \dots , say taken discrete, which are i.i.d., centered, and with common variance $t > 0$, do we have*

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n f_i \sim g_t$$

in the $n \rightarrow \infty$ limit, for some bell-shaped density g_t ? And, what is the formula of g_t ?

Observe that this question perfectly makes sense, with the probability theory that we know, by assuming that our random variables f_1, f_2, f_3, \dots are discrete, as said above. As for the $1/\sqrt{n}$ factor, there is certainly need for a normalization factor there, as for things to have a chance to converge, and the good factor is $1/\sqrt{n}$, as shown by:

PROPOSITION 4.2. *In order for a sum of the following type to have a chance to converge, with f_1, f_2, f_3, \dots being i.i.d., centered, and with common variance $t > 0$,*

$$S = \sum_{i=1}^n f_i$$

we must normalize this sum by a $1/\sqrt{n}$ factor, as in Question 4.1.

PROOF. The idea here is to look at the moments of S . Since all variables f_i are centered, $E(f_i) = 0$, so is their sum, $E(S) = 0$, and no contradiction here. However,

when looking at the variance of S , which equals the second moment, due to $E(S) = 0$, things become interesting, due to the following computation:

$$\begin{aligned}
 V(S) &= E(S^2) \\
 &= E\left(\sum_{ij} f_i f_j\right) \\
 &= \sum_{ij} E(f_i f_j) \\
 &= \sum_i E(f_i^2) + \sum_{i \neq j} E(f_i) E(f_j) \\
 &= \sum_i E(f_i^2) \\
 &= nt
 \end{aligned}$$

Thus, we are in need a normalization factor α , in order for our sum to have a chance to converge. But, repeating the computation with S replaced by αS gives:

$$V(\alpha S) = \alpha^2 nt$$

Thus, the good normalization factor is $\alpha = 1/\sqrt{n}$, as claimed. \square

So far, so good, we have a nice problem above, and time now to make a plan, in order to solve it. With the tools that we have, from chapters 2-3, here is such a plan:

PLAN 4.3. *In order to solve our central limiting question, we have to:*

- (1) *Apply Fourier and let $n \rightarrow \infty$, as to compute the Fourier transform of g_t .*
- (2) *Do some combinatorics and calculus, as to compute the moments of g_t .*
- (3) *Recover g_t out of its moments, again via combinatorics and calculus.*

Which sounds very good, but since we will be navigating here in troubled waters, with g_t being continuous, as opposed to the discrete densities that we are used to, perhaps time to ask the cat, before embarking on our trip. And good news, cat is already onboard, hired to catch mice on our ship, so let's ask him what he has to say. And cat says:

CAT 4.4. *Capitain, that g_t we are looking for is something two-dimensional, and all this normally comes after multivariable calculus, well done. But we can still go ahead, as long as there's enough booze for you guys, and mice for me, things fine.*

Nice advice, especially for the last part, but in what regards the first part, I must admit that I am completely puzzled. What are that two dimensions cat is talking about? Hope he has not confused mice with booze. So thanks cat, but please don't touch human food during our trip, mice only for you, and let's leave now, time to go.

So, back to our Plan 4.3, let us start with (1) there. Things are quickly done here, by using the linearization results for convolution from chapter 2, which lead to:

THEOREM 4.5. *Given discrete variables f_1, f_2, f_3, \dots , which are i.i.d., centered, and with common variance $t > 0$, we have*

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n f_i \sim g_t$$

with $n \rightarrow \infty$, with g_t being the law having $F(x) = e^{-tx^2/2}$ as Fourier transform.

PROOF. There are several things going on here, the idea being as follows:

(1) Observe first that in terms of moments, the Fourier transform of an arbitrary random variable $f : X \rightarrow \mathbb{R}$ is given by the following formula:

$$\begin{aligned} F_f(x) &= E(e^{ixf}) \\ &= E\left(\sum_{k=0}^{\infty} \frac{(ixf)^k}{k!}\right) \\ &= \sum_{k=0}^{\infty} \frac{(ix)^k E(f^k)}{k!} \\ &= \sum_{k=0}^{\infty} \frac{i^k M_k(f)}{k!} x^k \end{aligned}$$

(2) In particular, in the case of a centered variable, $E(f) = 0$, as those that we are interested in, the Fourier transform formula that we get is as follows:

$$F_f(x) = 1 - \frac{M_2(f)}{2} \cdot x^2 - i \frac{M_3(f)}{6} \cdot x^3 + \dots$$

Moreover, by further assuming that the Fourier variable is small, $x \simeq 0$, the Fourier transform formula that we get, that we will use in what follows, becomes:

$$F_f(x) = 1 - \frac{M_2(f)}{2} \cdot x^2 + O(x^2)$$

(3) In addition to this, we will also need to know what happens to the Fourier transform when rescaling. But the formula here is very easy to find, as follows:

$$\begin{aligned} F_{\alpha f}(x) &= E(e^{ix\alpha f}) \\ &= E(e^{i\alpha x f}) \\ &= F_f(\alpha x) \end{aligned}$$

(4) Good news, we can now do our computation. By using the above formulae in (2) and (3), the Fourier transform of the variable in the statement is given by:

$$\begin{aligned}
 F(x) &= \left[F_f \left(\frac{x}{\sqrt{n}} \right) \right]^n \\
 &= \left[1 - \frac{M_2(f)}{2} \cdot \frac{x^2}{n} + O(n^{-2}) \right]^n \\
 &= \left[1 - \frac{tx^2}{2n} + O(n^{-2}) \right]^n \\
 &\simeq \left[1 - \frac{tx^2}{2n} \right]^n \\
 &\simeq e^{-tx^2/2}
 \end{aligned}$$

(3) We are therefore led to the conclusion in the statement, modulo the fact that we do not know yet that a density g_t having as Fourier transform $F(x) = e^{-tx^2/2}$ really exists, plus perhaps some other abstract issues, related to the continuous measures, to be discussed too. But too late to go back, both cat and sailors are happy, we will go ahead. So, theorem proved, modulo finding that law g_t , which still remains to be done. \square

Getting now to step (2) of our Plan 4.3, that is easy to work out too, via some elementary one-variable calculus, with the result here being as follows:

THEOREM 4.6. *The “normal” law g_t , having as Fourier transform*

$$F(x) = e^{-tx^2/2}$$

must have all odd moments zero, and its even moments must be the numbers

$$M_k(g_t) = t^{k/2} \times k!!$$

where $k!! = (k-1)(k-3)(k-5) \dots$, for $k \in 2\mathbb{N}$.

PROOF. Again, several things going on here, the idea being as follows:

(1) To start with, at the level of formalism and notations, in view of Question 4.1 and of Theorem 4.5, we have adopted the term “normal” for the mysterious law g_t that we are looking for, the one having $F(x) = e^{-tx^2/2}$ as Fourier transform.

(2) Getting towards the computation of the moments, as a first useful observation, according to Theorem 4.5 this normal law g_t must be centered, as shown by:

$$\begin{aligned} f_i = \text{centered} &\implies \sum_{i=1}^n f_i = \text{centered} \\ &\implies \frac{1}{\sqrt{n}} \sum_{i=1}^n f_i = \text{centered} \\ &\implies g_t = \text{centered} \end{aligned}$$

Moreover, the same argument works by replacing “centered” with “having an even function as density”, and this shows, via some standard calculus, that we will leave here as an exercise, that the odd moments of our normal law must vanish:

$$M_{2l+1}(g_t) = 0$$

Thus, first assertion proved, and we only have to care about the even moments.

(3) As a comment here, as we will see in a moment, our study below of the moments computes in fact the odd moments too, as being all equal to 0, this time without making reference to Theorem 4.5. Thus, definitely no worries with the odd moments.

(4) Getting to work now, we must reformulate the equation $F(x) = e^{-tx^2/2}$, in terms of moments. We know from the proof of Theorem 4.5 that we have:

$$F(x) = \sum_{k=0}^{\infty} \frac{i^k M_k(g_t)}{k!} x^k$$

On the other hand, we have the following formula, for the exponential:

$$e^{-tx^2/2} = \sum_{r=0}^{\infty} (-1)^r \frac{t^r x^{2r}}{2^r r!}$$

Thus, our equation $F(x) = e^{-tx^2/2}$ takes the following form:

$$\sum_{k=0}^{\infty} \frac{i^k M_k(g_t)}{k!} x^k = \sum_{r=0}^{\infty} (-1)^r \frac{t^r x^{2r}}{2^r r!}$$

(5) As a first observation, the odd moments must vanish, as said in (2) above. As for the even moments, these can be computed as follows:

$$\begin{aligned}
M_k(g_t) &= k! \times \frac{t^{k/2}}{2^{k/2}(k/2)!} \\
&= t^{k/2} \times \frac{k!}{2^{k/2}(k/2)!} \\
&= t^{k/2} \times \frac{2 \cdot 3 \cdot 4 \dots (k-1) \cdot k}{2 \cdot 4 \cdot 6 \dots (k-2) \cdot k} \\
&= t^{k/2} \times 3 \cdot 5 \dots (k-3)(k-1) \\
&= t^{k/2} \times k!!
\end{aligned}$$

Thus, we are led to the formula in the statement. \square

The moment formula that we found is quite interesting, and before going ahead with step (3) of our Plan 4.3, let us look a bit at this, and see what we can further say.

To be more precise, in analogy with what we know about the Poisson laws, and about the Bessel laws too, making reference to interesting combinatorics and partitions, when it comes to moments, we have the following result, regarding the normal laws:

THEOREM 4.7. *The moments of the normal law g_t are given by*

$$M_k(g_t) = t^{k/2} |P_2(k)|$$

for any $k \in \mathbb{N}$, with $P_2(k)$ standing for the pairings of $\{1, \dots, k\}$.

PROOF. This is a reformulation of Theorem 4.6, the idea being as follows:

(1) We know from Theorem 4.6 that the moments of the normal law $M_k = M_k(g_t)$ that we are interested in are given by the following formula, with the convention $k!! = 0$ for k odd, and $k!! = (k-1)(k-3)(k-5) \dots$ for k even, for the double factorials:

$$M_k(g_t) = t^{k/2} \times k!!$$

Now observe that, according to our above convention for the double factorials, these are subject to the following recurrence relation, with initial data $1!! = 0, 2!! = 1$:

$$k!! = (k-1)(k-2)!!$$

We conclude that the moments of the normal law $M_k = M_k(g_t)$ are subject to the following recurrence relation, with initial data $M_1 = 0, M_2 = t$:

$$M_k = t(k-1)M_{k-2}$$

(2) On the other hand, let us first count the pairings of the set $\{1, \dots, k\}$. In order to have such a pairing, we must pair 1 with one of the numbers $2, \dots, k$, and then use a

pairing of the remaining $k - 2$ numbers. Thus, we have the following recurrence formula for the number P_k of such pairings, with the initial data $P_1 = 0, P_2 = 1$:

$$P_k = (k - 1)P_{k-2}$$

Now by multiplying by $t^{k/2}$, the resulting numbers $N_k = t^{k/2}P_k$ will be subject to the following recurrence relation, with initial data $N_1 = 0, N_2 = t$:

$$N_k = t(k - 1)N_{k-2}$$

(3) Thus, the moments $M_k = M_k(g_t)$ and the numbers $N_k = t^{k/2}P_k$ are subject to the same recurrence relation, with the same initial data, so they are equal, as claimed. \square

Still in analogy with what we know about the Poisson laws, and about the Bessel laws too, we can further process what we found in Theorem 4.7, and we are led to:

THEOREM 4.8. *The moments of the normal law g_t are given by*

$$M_k(g_t) = \sum_{\pi \in P_2(k)} t^{|\pi|}$$

where $P_2(k)$ is the set of pairings of $\{1, \dots, k\}$, and $|\cdot|$ is the number of blocks.

PROOF. This is a quick reformulation of Theorem 4.7, with the number of blocks of a pairing of $\{1, \dots, k\}$ being trivially $k/2$, independently of the pairing. \square

It is possible to do some more combinatorics here, again in relation with what we know about the Poisson laws, for instance by looking at cumulants, and we have:

THEOREM 4.9. *The cumulants of the normal law g_t are the following numbers:*

$$0, t, 0, 0, \dots$$

*In particular, the normal laws satisfy $g_s * g_t = g_{s+t}$, for any $s, t > 0$.*

PROOF. We have two assertions here, the idea being as follows:

(1) For the normal law g_t we have the following computation:

$$\begin{aligned} K_\mu(\xi) &= \log F_\mu(-i\xi) \\ &= \log \exp \left[-t(-i\xi)^2/2 \right] \\ &= t\xi^2/2 \end{aligned}$$

But the plain coefficients of this series are the numbers $0, t/2, 0, 0, \dots$, and so the Taylor coefficients of this series are the numbers $0, t, 0, 0, \dots$, as claimed.

(2) As for the last assertion, regarding the semigroup property of the normal laws, this actually follows from Theorem 4.5, the log of the Fourier transform being linear in t , but is best seen by looking at the cumulants, which are obviously linear in t .

(3) However, as a technical remark here, the linearization results for the convolution that we have, be them in terms of the Fourier transform, or of the cumulants, were

formally established in chapters 2-3 only for the discrete measures. So, instead of further thinking at all this, let us pull out a third, elementary proof for $g_s * g_t = g_{s+t}$.

(4) In order to do this, consider, as in Theorem 4.5, on one hand i.i.d. centered variables f_1, f_2, f_3, \dots having variance $s > 0$, and on the other hand i.i.d. centered variables h_1, h_2, h_3, \dots having variance $t > 0$. According to Theorem 4.5, we have:

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n f_i \sim g_s \quad , \quad \frac{1}{\sqrt{n}} \sum_{i=1}^n h_i \sim g_t$$

Now let us sum these formulae. Assuming that the variables f_1, f_2, f_3, \dots that we used were independent from the variables h_1, h_2, h_3, \dots , we obtain in this way:

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n (f_i + h_i) \sim g_s * g_t$$

On the other hand, yet another application of Theorem 4.5, with the remark that by independence, the variance of $f_i + h_i$ is indeed $s + t$, gives the following formula:

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n (f_i + h_i) \sim g_{s+t}$$

Thus, we are led to the semigroup formula $g_s * g_t = g_{s+t}$, as desired. □

As a philosophical conclusion now to all this, let us formulate:

CONCLUSION 4.10. *The normal laws g_t have properties which are quite similar to those of the Poisson laws p_t , and combinatorially, the passage*

$$p_t \rightarrow g_t$$

appears by replacing the partitions with the pairings.

Which sounds quite conceptual, hope you agree with me. According to Question 4.1, compared to the Poisson Limit Theorem, the normal law g_t should be rather the “continuous version” of the Poisson law p_t , and here we are into some sort of quite advanced mathematics, telling us that the passage from discrete to continuous appears, combinatorially speaking, by replacing the partitions with the pairings.

We will explore such things, which are indeed quite advanced, later in this book. In the meantime, however, we still need to know what the density of g_t is.

4b. Normal laws

So, let us get now to step (3) of our Plan 4.3. This does not look obvious at all, but some partial integration know-how leads us to the following statement:

THEOREM 4.11. *The normal laws are given by*

$$g_t = \frac{1}{\sqrt{2t} \cdot I} e^{-x^2/2t} dx$$

with the constant on the bottom being $I = \int_{\mathbb{R}} e^{-x^2} dx$.

PROOF. This comes from partial integration, as follows:

(1) Let us first do a naive computation. Consider the following quantities:

$$M_k = \int_{\mathbb{R}} x^k e^{-x^2} dx$$

It is quite obvious that by partial integration we will get a recurrence formula for these numbers, similar to the one that we have for the moments of the normal laws. So, let us do this. By partial integration we obtain the following formula, for any $k \in \mathbb{N}$:

$$\begin{aligned} M_k &= -\frac{1}{2} \int_{\mathbb{R}} x^{k-1} \left(e^{-x^2} \right)' dx \\ &= \frac{1}{2} \int_{\mathbb{R}} (k-1) x^{k-2} e^{-x^2} dx \\ &= \frac{k-1}{2} \cdot M_{k-2} \end{aligned}$$

(2) Thus, we are on the good way, with the recurrence formula that we got being the same as that for the moments of $g_{1/2}$. Now let us fine-tune this, as to reach to the same recurrence as for the moments of g_t . Consider the following quantities:

$$N_k = \int_{\mathbb{R}} x^k e^{-x^2/2t} dx$$

By partial integration as before, we obtain the following formula:

$$\begin{aligned} N_k &= \int_{\mathbb{R}} (tx^{k-1}) \left(-e^{-x^2/2t} \right)' dx \\ &= \int_{\mathbb{R}} t(k-1) x^{k-2} e^{-x^2/2t} dx \\ &= t(k-1) \int_{\mathbb{R}} x^{k-2} e^{-x^2/2t} dx \\ &= t(k-1) N_{k-2} \end{aligned}$$

(3) Thus, almost done, and it remains to discuss normalization. We know from the above that we must have a formula as follows, with I_t being a certain constant:

$$g_t = \frac{1}{I_t} \cdot e^{-x^2/2t} dx$$

But the constant I_t must be the one making g_t of mass 1, and so:

$$\begin{aligned} I_t &= \int_{\mathbb{R}} e^{-x^2/2t} dx \\ &= \int_{\mathbb{R}} e^{-2ty^2/2t} \sqrt{2t} dy \\ &= \sqrt{2t} \int_{\mathbb{R}} e^{-y^2} dy \end{aligned}$$

Thus, we are led to the formula in the statement. \square

What we did in the above is good work, and it remains to compute the constant I appearing in Theorem 4.11, given by the following formula, and called Gauss integral:

$$I = \int_{\mathbb{R}} e^{-x^2} dx$$

Which does not look obvious at all, so time to ask again the cat. Unfortunately cat is nowhere to be found, probably gone with the sailors, for some partying, now that we are again on land, for a short break. This being said, walking through this debauchery, I can spot a small kitty, behind a pile of harbor garbage. She's really small, and can barely meow, but never knows, let's ask her about mathematics. And kitty answers:

KITTY 4.12. *The normal laws are naturally two-dimensional, and it is*

$$I^2 = \int_{\mathbb{R}} \int_{\mathbb{R}} e^{-x^2-y^2} dx dy$$

the quantity that you should compute. I bet you can do that.

Thanks kitty, this sounds wise, so again that mysterious two-dimensionality, that cat was talking about. So, I will try this, and by the way, come with me. Cat is surely very good at both catching mice and helping with trigonometry and other math, but these days he's learning more and more bad things from the sailors, I am a bit worried about this, and in any case, we need a backup. You'll like it, plenty of mice on our ship.

So, following now kitty's advice, we are led to the following result:

THEOREM 4.13. *We have the following formula,*

$$\int_{\mathbb{R}} e^{-x^2} dx = \sqrt{\pi}$$

called Gauss integral formula.

PROOF. As already mentioned, this is something which is nearly impossible to prove, with bare hands. However, this can be proved by using two dimensions, as follows:

$$\begin{aligned}
 \int_{\mathbb{R}} \int_{\mathbb{R}} e^{-x^2-y^2} dx dy &= 4 \int_0^\infty \int_0^\infty e^{-x^2-y^2} dx dy \\
 &= 4 \int_0^\infty \int_0^\infty e^{-t^2 y^2 - y^2} y dt dy \\
 &= 4 \int_0^\infty \int_0^\infty y e^{-y^2(1+t^2)} dy dt \\
 &= 2 \int_0^\infty \int_0^\infty \left(-\frac{e^{-y^2(1+t^2)}}{1+t^2} \right)' dy dt \\
 &= 2 \int_0^\infty \frac{dt}{1+t^2} \\
 &= 2 \int_0^\infty (\arctan t)' dt \\
 &= \pi
 \end{aligned}$$

Thus, we are led to the conclusion in the statement. \square

Very nice, so as a final conclusion to our study, started long ago, in the beginning of this chapter, we can now formulate the Central Limit Theorem (CLT), as follows:

THEOREM 4.14 (CLT). *Given discrete random variables f_1, f_2, f_3, \dots , which are i.i.d., centered, and with common variance $t > 0$, we have*

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n f_i \sim g_t$$

in the $n \rightarrow \infty$ limit, in moments, with the limiting measure being

$$g_t = \frac{1}{\sqrt{2\pi t}} e^{-x^2/2t} dx$$

called normal, or Gaussian law of parameter $t > 0$.

PROOF. This follows indeed from our various results above, and more specifically from Theorem 4.5, Theorem 4.6 for the terminology, Theorem 4.11 and Theorem 4.13. \square

Let us study now more in detail the laws that we found. Normally we already have everything that is needed, but it is instructive at this point to do some computations, based on the explicit formula of g_t found above, and on Theorem 4.13. We first have:

PROPOSITION 4.15. *We have the variance formula*

$$V(g_t) = t$$

valid for any $t > 0$.

PROOF. We already know this, but we can establish this as well directly, starting from our formula of g_t from Theorem 4.14. Indeed, the first moment is 0, because our normal law g_t is centered. As for the second moment, this can be computed as follows:

$$\begin{aligned}
 M_2 &= \frac{1}{\sqrt{2\pi t}} \int_{\mathbb{R}} x^2 e^{-x^2/2t} dx \\
 &= \frac{1}{\sqrt{2\pi t}} \int_{\mathbb{R}} (tx) \left(-e^{-x^2/2t} \right)' dx \\
 &= \frac{1}{\sqrt{2\pi t}} \int_{\mathbb{R}} t e^{-x^2/2t} dx \\
 &= t
 \end{aligned}$$

We conclude from this that the variance is $V = M_2 = t$, as claimed. \square

More generally, we can recover in this way the computation of all moments:

THEOREM 4.16. *The even moments of the normal law are the numbers*

$$M_k(g_t) = t^{k/2} \times k!!$$

where $k!! = (k-1)(k-3)(k-5)\dots$, and the odd moments vanish.

PROOF. Again, we already know this, but we can establish this as well directly, starting from our formula above of g_t . Indeed, we have the following computation:

$$\begin{aligned}
 M_k &= \frac{1}{\sqrt{2\pi t}} \int_{\mathbb{R}} y^k e^{-y^2/2t} dy \\
 &= \frac{1}{\sqrt{2\pi t}} \int_{\mathbb{R}} (ty^{k-1}) \left(-e^{-y^2/2t} \right)' dy \\
 &= \frac{1}{\sqrt{2\pi t}} \int_{\mathbb{R}} t(k-1)y^{k-2} e^{-y^2/2t} dy \\
 &= t(k-1) \times \frac{1}{\sqrt{2\pi t}} \int_{\mathbb{R}} y^{k-2} e^{-y^2/2t} dy \\
 &= t(k-1)M_{k-2}
 \end{aligned}$$

Thus by recurrence, we are led to the formula in the statement. \square

Here is another result, which is the key one for the study of the normal laws:

THEOREM 4.17. *We have the following formula, valid for any $t > 0$:*

$$F_{g_t}(x) = e^{-tx^2/2}$$

*In particular, the normal laws satisfy $g_s * g_t = g_{s+t}$, for any $s, t > 0$.*

PROOF. As before, we already know this, but we can establish now the Fourier transform formula as well directly, by using the explicit formula of g_t , as follows:

$$\begin{aligned}
 F_{g_t}(x) &= \frac{1}{\sqrt{2\pi t}} \int_{\mathbb{R}} e^{-y^2/2t + ixy} dy \\
 &= \frac{1}{\sqrt{2\pi t}} \int_{\mathbb{R}} e^{-(y/\sqrt{2t} - \sqrt{t/2}ix)^2 - tx^2/2} dy \\
 &= \frac{1}{\sqrt{2\pi t}} \int_{\mathbb{R}} e^{-z^2 - tx^2/2} \sqrt{2t} dz \\
 &= \frac{1}{\sqrt{\pi}} e^{-tx^2/2} \int_{\mathbb{R}} e^{-z^2} dz \\
 &= e^{-tx^2/2}
 \end{aligned}$$

As for the last assertion, this follows from the fact that $\log F_{g_t}$ is linear in t . \square

Observe that, thinking retrospectively, the above computation formally solves the question raised by Theorem 4.5, and so could have been used there, afterwards. However, and here comes the point, all this is based on Theorem 4.13, and also, crucially, on our work from Theorem 4.11, which in turn was based on moments and so on.

Thus, no really way of doing things in a simpler way. There is however one simplification, because if you know well multivariable calculus, the Gauss formula comes right away via polar coordinates, the computation here being as follows:

$$\begin{aligned}
 \int_{\mathbb{R}} \int_{\mathbb{R}} e^{-x^2 - y^2} dx dy &= \int_0^{2\pi} \int_0^\infty e^{-r^2} r dr dt \\
 &= 2\pi \int_0^\infty \left(-\frac{e^{-r^2}}{2} \right)' dr \\
 &= 2\pi \left[0 - \left(-\frac{1}{2} \right) \right] \\
 &= \pi
 \end{aligned}$$

And, based on this, it is quite natural to introduce the normal laws g_t directly, then compute Fourier, and prove the CLT, as probabilists do. More on this later, once we will know well multivariable calculus, and in particular the formula $dx dy = r dr dt$, which was crucially used in the above computation, we will certainly come back to this.

As another comment, as explained in the previous section, before even knowing the formula of the densities, the normal laws g_t have a quite interesting combinatorics, quite similar to that of the Poisson and Bessel laws. We will be back to this later, in Part IV below, when discussing on a more systematic basis probability theory, in general. In the

meantime, however, a few words about groups, which are still missing from our picture. In analogy with what we know about the Poisson and Bessel laws, we have:

FACT 4.18. *For the orthogonal group O_N , the main character*

$$\chi(U) = \sum_{i=1}^N U_{ii}$$

follows with $N \rightarrow \infty$ the normal law g_1 . More generally, the truncated character

$$\chi(U) = \sum_{i=1}^{[tN]} U_{ii}$$

follows with $N \rightarrow \infty$ the normal law g_t , for any $t \in (0, 1]$.

Which looks very good, and in tune with what we know about the Poisson and Bessel laws, with the overall conclusion being that the passage $p_t \rightarrow g_t$, from discrete to continuous, amounts in performing a passage $S_N \rightarrow O_N$ at the group theory level.

Moreover, with some group theory helping, it is possible to derive from this, directly, our previous Conclusion 4.10, in regards with partitions and pairings.

However, and here comes the point, proving such things is quite difficult, because unlike in the finite group case, where we were able to get away with elementary counts on S_N, H_N and other reflection groups, in order to deal with continuous groups like O_N , we have to know more about the uniform integration on such groups, and this is something non-trivial. So, all this will have to wait a bit, we will be back to it in chapter 14.

4c. Complex variables

Let us discuss now the complex analogues of all the above, with a notion of complex normal, or Gaussian law. To start with, we have the following definition:

DEFINITION 4.19. *A complex random variable is a variable $f : X \rightarrow \mathbb{C}$. In the discrete case, the law of such a variable is the complex probability measure*

$$\mu = \sum_i \alpha_i \delta_{z_i} \quad , \quad \alpha_i \geq 0 \quad , \quad \sum_i \alpha_i = 1 \quad , \quad z_i \in \mathbb{C}$$

given by the following formula, with P being the probability over X ,

$$\mu = \sum_{z \in \mathbb{C}} P(f = z) \delta_z$$

with the sum being finite or countable, as per our discreteness assumption.

Observe the similarity with the analogous notions introduced in chapter 2, for the real variables $f : X \rightarrow \mathbb{R}$. In fact, what we are doing here is to extend the formalism from chapter 2, from real to complex, in a straightforward way. As a basic example for this, any real variable $f : X \rightarrow \mathbb{R}$ can be regarded as a complex variable $f : X \rightarrow \mathbb{C}$.

In order to understand the precise relation with the real theory, from chapter 2, we can decompose any complex variable $f : X \rightarrow \mathbb{C}$ as a sum, as follows:

$$f = g + ih \quad , \quad g = \operatorname{Re}(f), \quad h = \operatorname{Im}(f)$$

With this done, we have the following computation, for the corresponding law:

$$\begin{aligned} \mu &= \sum_{z \in \mathbb{C}} P(f = z) \delta_z \\ &= \sum_{x, y \in \mathbb{R}} P(f = x + iy) \delta_{x+iy} \\ &= \sum_{x, y \in \mathbb{R}} P(g + ih = x + iy) \delta_{x+iy} \\ &= \sum_{x, y \in \mathbb{R}} P(g = x, h = y) \delta_{x+iy} \end{aligned}$$

In the case where the real and imaginary parts $g, h : X \rightarrow \mathbb{R}$ are independent, we can say more about this, with the above computation having the following continuation:

$$\begin{aligned} \mu &= \sum_{x, y \in \mathbb{R}} P(g = x, h = y) \delta_{x+iy} \\ &= \sum_{x, y \in \mathbb{R}} P(g = x) P(h = y) \delta_{x+iy} \\ &= \sum_{x, y \in \mathbb{R}} P(g = x) P(h = y) \delta_x * \delta_{iy} \\ &= \left(\sum_{x \in \mathbb{R}} P(g = x) \delta_x \right) * \left(\sum_{y \in \mathbb{R}} P(h = y) \delta_{iy} \right) \\ &= \mu_g * i\mu_h \end{aligned}$$

To be more precise, we have used here in the beginning the independence of the variables $h, g : X \rightarrow \mathbb{R}$, and at the end we have denoted the measure on the right, which is obtained from μ_h by putting this measure on the imaginary axis, by $i\mu_h$.

All this is quite interesting, going beyond what we know so far about basic probability, in the real case, so let us record this finding, along with a bit more, as follows:

THEOREM 4.20. *For a discrete complex random variable $f : X \rightarrow \mathbb{C}$, decomposed into real and imaginary parts as $f = g + ih$, and with g, h assumed independent, we have*

$$\mu_f = \mu_g * i\mu_h$$

with $$ being the usual convolution operation, $\delta_z * \delta_t = \delta_{z+t}$, and with $\mu \rightarrow i\mu$ denoting the rotated version, $\mathbb{R} \rightarrow i\mathbb{R}$. If g, h are not independent, this formula does not hold.*

PROOF. We already know that the first assertion holds, as explained in the above. As for the second assertion, this follows by carefully examining the above computation. Indeed, we have used only at one point the independence of g, h , so for the formula $\mu_f = \mu_g * i\mu_h$ to hold, the equality used at that point, which is as follows, must hold:

$$\sum_{x,y \in \mathbb{R}} P(g = x, h = y) \delta_{x+iy} = \sum_{x,y \in \mathbb{R}} P(g = x) P(h = y) \delta_{x+iy}$$

But this is the same as saying that the following must hold, for any x, y :

$$P(g = x, h = y) = P(g = x) P(h = y)$$

We conclude that, in order for the decomposition formula $\mu_f = \mu_g * i\mu_h$ to hold, the real and imaginary parts $g, h : X \rightarrow \mathbb{R}$ must be independent, as stated. \square

Many other things can be said, along the same lines, inspired by the basic theory of the complex numbers. Indeed, what we used in the above was the fact that any complex number decomposes as $z = x + iy$ with $x, y \in \mathbb{R}$, but at a more advanced level, we can equally use formulae of type $z = re^{it}$, or $|z|^2 = z\bar{z}$ and so on, and we are led in this way to a whole collection of results, connecting real and complex probability theory.

Going now straight to the point, probabilistic limiting theorems, let us discuss the complex analogue of the CLT. We have the following statement, to start with:

THEOREM 4.21. *Given discrete complex variables f_1, f_2, f_3, \dots whose real and imaginary parts are i.i.d., centered, and with common variance $t > 0$, we have*

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n f_i \sim C_t$$

with $n \rightarrow \infty$, in moments, where C_t is the law of a complex variable whose real and imaginary parts are independent, and each following the law g_t .

PROOF. This follows indeed from the real CLT, established in Theorem 4.14, simply by taking the real and imaginary parts of all the variables involved. \square

It is tempting at this point to call Theorem 4.21 the complex CLT, or CCLT, but before doing that, let us study a bit more all this. We would like to have a better understanding

of the limiting law C_t at the end, and for this purpose, let us look at a sum as follows, with a, b being real independent variables, both following the normal law g_t :

$$c = a + ib$$

To start with, this variable is centered, in a complex sense, because we have:

$$\begin{aligned} E(c) &= E(a + ib) \\ &= E(a) + iE(b) \\ &= 0 + i \cdot 0 \\ &= 0 \end{aligned}$$

Regarding now the variance, things are more complicated, because the usual variance formula from the real case, which is $V(c) = E(c^2)$ in the centered case, will not provide us with a positive number, in the case where our variable is not real. So, in order to have a variance which is real, and positive too, we must rather use a formula of type $V(c) = E(|c|^2)$, in the centered case. And, with this convention for the variance, we have then the following computation, for the variance of the above variable c :

$$\begin{aligned} V(c) &= E(|c|^2) \\ &= E(a^2 + b^2) \\ &= E(a^2) + E(b^2) \\ &= V(a^2) + V(b^2) \\ &= t + t \\ &= 2t \end{aligned}$$

But this suggests to divide everything by $\sqrt{2}$, as to have in the end a variable having complex variance t , in our sense, and we are led in this way into:

DEFINITION 4.22. *The complex normal, or Gaussian law of parameter $t > 0$ is*

$$G_t = \text{law} \left(\frac{1}{\sqrt{2}}(a + ib) \right)$$

where a, b are real and independent, each following the law g_t .

In short, the complex normal laws appear as natural complexifications of the real normal laws. As in the real case, these measures form convolution semigroups:

PROPOSITION 4.23. *The complex Gaussian laws have the property*

$$G_s * G_t = G_{s+t}$$

for any $s, t > 0$, and so they form a convolution semigroup.

PROOF. This follows indeed from the real result, namely $g_s * g_t = g_{s+t}$, established in Theorem 4.9, simply by taking real and imaginary parts. \square

We have as well the following complex analogue of the CLT:

THEOREM 4.24 (CCLT). *Given discrete complex variables f_1, f_2, f_3, \dots whose real and imaginary parts are i.i.d. and centered, and having variance $t > 0$, we have*

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n f_i \sim G_t$$

with $n \rightarrow \infty$, in moments.

PROOF. This follows indeed from our previous CCLT result, from Theorem 4.21, by dividing everything by $\sqrt{2}$, as explained in the above. \square

Finally, in relation with groups, we have the following analogue of Fact 4.18:

FACT 4.25. *For the unitary group U_N , the main character*

$$\chi(U) = \sum_{i=1}^N U_{ii}$$

follows with $N \rightarrow \infty$ the normal law G_1 . More generally, the truncated character

$$\chi(U) = \sum_{i=1}^{[tN]} U_{ii}$$

follows with $N \rightarrow \infty$ the normal law G_t , for any $t \in (0, 1]$.

Which looks of course very good and conceptual, but as before with Fact 4.18, when it comes to prove such things, this is no easy business. More on this later in this book.

4d. Wick formula

Regarding now the moments, the situation here is more complicated than in the real case, because in order to have good results, we have to deal with both the complex variables, and their conjugates. Let us formulate the following definition:

DEFINITION 4.26. *The moments a complex variable $f \in L^\infty(X)$ are the numbers*

$$M_k = E(f^k)$$

depending on colored integers $k = \circ \bullet \bullet \circ \dots$, with the conventions

$$f^\emptyset = 1 \quad , \quad f^\circ = f \quad , \quad f^\bullet = \bar{f}$$

and multiplicativity, in order to define the colored powers f^k .

As an illustration for this notion, which is something very intuitive, here are the formulae of the four possible order 2 moments of a complex variable f :

$$M_{\circ\circ} = E(f^2) \quad , \quad M_{\circ\bullet} = E(f\bar{f})$$

$$M_{\bullet\circ} = E(\bar{f}f) \quad , \quad M_{\bullet\bullet} = E(\bar{f}^2)$$

Observe that, since f, \bar{f} commute, we have the following identity, which shows that there is a bit of redundancy in our above definition, as formulated:

$$M_{\circ\bullet} = M_{\bullet\circ}$$

In fact, again since f, \bar{f} commute, we can permute terms, in the general context of Definition 4.26, and restrict the attention to exponents of the following type:

$$k = \dots \circ \circ \circ \bullet \bullet \bullet \dots$$

However, our results about the complex Gaussian laws, and other complex laws, later on, not to talk about laws of matrices, random matrices and other noncommuting variables, that will appear later too, will look better without doing this. So, we will use Definition 4.26 as stated. Getting to work now, we first have the following result:

THEOREM 4.27. *The moments of the complex normal law are given by*

$$M_k(G_t) = \begin{cases} t^p p! & (k \text{ uniform, of length } 2p) \\ 0 & (k \text{ not uniform}) \end{cases}$$

where $k = \circ \bullet \bullet \circ \dots$ is called *uniform* when it contains the same number of \circ and \bullet .

PROOF. We must compute the moments, with respect to colored integer exponents $k = \circ \bullet \bullet \circ \dots$ as above, of the variable from Definition 4.22, namely:

$$f = \frac{1}{\sqrt{2}}(a + ib)$$

We can assume that we are in the case $t = 1$, and the proof here goes as follows:

(1) As a first observation, in the case where our exponent $k = \circ \bullet \bullet \circ \dots$ is not uniform, a standard rotation argument shows that the corresponding moment of f vanishes. To be more precise, the variable $f' = wf$ is complex Gaussian too, for any complex number $w \in \mathbb{T}$, and from $M_k(f) = M_k(f')$ we obtain $M_k(f) = 0$, in this case.

(2) In the uniform case now, where the exponent $k = \circ \bullet \bullet \circ \dots$ consists of p copies of \circ and p copies of \bullet , the corresponding moment can be computed as follows:

$$\begin{aligned}
M_k &= \int (f \bar{f})^p \\
&= \frac{1}{2^p} \int (a^2 + b^2)^p \\
&= \frac{1}{2^p} \sum_r \binom{p}{r} \int a^{2r} \int b^{2p-2r} \\
&= \frac{1}{2^p} \sum_r \binom{p}{r} (2r)!! (2p-2r)!! \\
&= \frac{1}{2^p} \sum_r \frac{p!}{r!(p-r)!} \cdot \frac{(2r)!}{2^r r!} \cdot \frac{(2p-2r)!}{2^{p-r} (p-r)!} \\
&= \frac{p!}{4^p} \sum_r \binom{2r}{r} \binom{2p-2r}{p-r}
\end{aligned}$$

(3) In order to finish now the computation, let us recall that we have the following formula, coming from the generalized binomial formula, or from the Taylor formula:

$$\frac{1}{\sqrt{1+t}} = \sum_{q=0}^{\infty} \binom{2q}{q} \left(\frac{-t}{4}\right)^q$$

By taking the square of this series, we obtain the following formula:

$$\frac{1}{1+t} = \sum_p \left(\frac{-t}{4}\right)^p \sum_r \binom{2r}{r} \binom{2p-2r}{p-r}$$

Now by looking at the coefficient of t^p on both sides, we conclude that the sum on the right equals 4^p . Thus, we can finish the moment computation in (2), as follows:

$$M_k = \frac{p!}{4^p} \times 4^p = p!$$

We are therefore led to the conclusion in the statement. □

Before going further, let us record the following consequence, of the above:

THEOREM 4.28. *The moments of the Rayleigh law, given by*

$$R_t = \text{law}(|G_t|)$$

are given by the following formula, at the parameter value $t = 1$,

$$M_p = p!$$

and are given by the formula $M_p = t^p p!$, in general.

PROOF. This follows indeed from Theorem 4.27, or simply from the computation (2) in the above proof, at $t = 1$, and then by rescaling by t , in general. \square

Many other things can be said about the Rayleigh laws, which are quite interesting mathematical objects, notably with some further formulae, regarding the Fourier transform, and the cumulants, which can be obtained as a consequence of Theorem 4.28. We will be back to these remarkable laws, on several occasions, in what follows.

Let us record, however, the following statement, called Rayleigh Central Limiting Theorem, which is something of theoretical importance:

THEOREM 4.29 (RCLT). *Given discrete complex variables f_1, f_2, f_3, \dots whose real and imaginary parts are i.i.d. and centered, and having variance $t > 0$, we have*

$$\left| \frac{1}{\sqrt{n}} \sum_{i=1}^n f_i \right| \sim R_t$$

with $n \rightarrow \infty$, in moments.

PROOF. This follows indeed from our previous Central Limiting result, namely the CCLT from Theorem 4.24, by taking absolute values on both sides. \square

As a comment here, observe that the Rayleigh laws, while being certainly real, capture the essentials of the two-dimensional nature of the normal laws. So, probably time now for some further thinking, in relation with what cat and kitty were saying. But with this being something a bit complicated, and with our advisors being not really available, one with a hangover, and the other chasing mice, we will not further explore this subject.

Getting back now to the complex normal laws, what we have in Theorem 4.27 about them is usually what is needed in practice, when dealing with moments. But, as before with the real Gaussian laws, or even further before with the Poisson and Bessel laws, a better-looking statement regarding the moments is in terms of partitions.

Indeed, given a colored integer $k = \circ \bullet \bullet \circ \dots$, let us say that $\pi \in P_2(k)$ is matching when it pairs $\circ - \bullet$ symbols. With this convention, we have the following result:

THEOREM 4.30. *The moments of the complex normal law are the numbers*

$$M_k(G_t) = \sum_{\pi \in \mathcal{P}_2(k)} t^{|\pi|}$$

where $\mathcal{P}_2(k)$ are the matching pairings of $\{1, \dots, k\}$, and $|\cdot|$ is the number of blocks.

PROOF. This is a reformulation of Theorem 4.27. Indeed, we can assume that we are in the case $t = 1$, and here we know from Theorem 4.27 that the moments are:

$$M_k = \begin{cases} (|k|/2)! & (k \text{ uniform}) \\ 0 & (k \text{ not uniform}) \end{cases}$$

On the other hand, the numbers $|\mathcal{P}_2(k)|$ are given by exactly the same formula. Indeed, in order to have a matching pairing of k , our exponent $k = \circ \bullet \bullet \circ \dots$ must be uniform, consisting of p copies of \circ and p copies of \bullet , with $p = |k|/2$. But then the matching pairings of k correspond to the permutations of the \bullet symbols, as to be matched with \circ symbols, and so we have $p!$ such pairings. Thus, we have the same formula as for the moments of f , and we are led to the conclusion in the statement. \square

In practice, we also need to know how to compute joint moments. We have here:

THEOREM 4.31 (Wick formula). *Given independent variables f_i , each following the complex normal law G_t , with $t > 0$ being a fixed parameter, we have the formula*

$$E(f_{i_1}^{k_1} \dots f_{i_s}^{k_s}) = t^{s/2} \# \left\{ \pi \in \mathcal{P}_2(k) \mid \pi \leq \ker i \right\}$$

where $k = k_1 \dots k_s$ and $i = i_1 \dots i_s$, for the joint moments of these variables, where $\pi \leq \ker i$ means that the indices of i must fit into the blocks of π , in the obvious way.

PROOF. This is something well-known, which can be proved as follows:

(1) Let us first discuss the case where we have a single variable f , which amounts in taking $f_i = f$ for any i in the formula in the statement. What we have to compute here are the moments of f , with respect to colored integer exponents $k = \circ \bullet \bullet \circ \dots$, and the formula in the statement tells us that these moments must be:

$$E(f^k) = t^{|k|/2} |\mathcal{P}_2(k)|$$

But this is the formula in Theorem 4.30, so we are done with this case.

(2) In general now, when expanding the product $f_{i_1}^{k_1} \dots f_{i_s}^{k_s}$ and rearranging the terms, we are left with doing a number of computations as in (1), and then making the product of the expectations that we found. But this amounts in counting the partitions in the statement, with the condition $\pi \leq \ker i$ there standing for the fact that we are doing the various type (1) computations independently, and then making the product. \square

The above statement is one of the possible formulations of the Wick formula, and there are many more formulations, which are all useful. For instance, we have:

THEOREM 4.32 (Wick formula 2). *Given independent variables f_i , each following the complex normal law G_t , with $t > 0$ being a fixed parameter, we have the formula*

$$E(f_{i_1} \dots f_{i_k} f_{j_1}^* \dots f_{j_k}^*) = t^k \# \left\{ \pi \in S_k \mid i_{\pi(r)} = j_r, \forall r \right\}$$

for the non-vanishing joint moments of these variables.

PROOF. This follows from the usual Wick formula, from Theorem 4.31. With some changes in the indices and notations, the formula there reads:

$$E(f_{I_1}^{K_1} \dots f_{I_s}^{K_s}) = t^{s/2} \# \left\{ \sigma \in \mathcal{P}_2(K) \mid \sigma \leq \ker I \right\}$$

Now observe that we have $\mathcal{P}_2(K) = \emptyset$, unless the colored integer $K = K_1 \dots K_s$ is uniform, in the sense that it contains the same number of \circ and \bullet symbols. Up to permutations, the non-trivial case, where the moment is non-vanishing, is the case where the colored integer $K = K_1 \dots K_s$ is of the following special form:

$$K = \underbrace{\circ \circ \dots \circ}_k \underbrace{\bullet \bullet \dots \bullet}_k$$

So, let us focus on this case, which is the non-trivial one. Here we have $s = 2k$, and we can write the multi-index $I = I_1 \dots I_s$ in the following way:

$$I = i_1 \dots i_k j_1 \dots j_k$$

With these changes made, the above usual Wick formula reads:

$$E(f_{i_1} \dots f_{i_k} f_{j_1}^* \dots f_{j_k}^*) = t^k \# \left\{ \sigma \in \mathcal{P}_2(K) \mid \sigma \leq \ker(ij) \right\}$$

The point now is that the matching pairings $\sigma \in \mathcal{P}_2(K)$, with $K = \circ \dots \circ \bullet \dots \bullet$, of length $2k$, as above, correspond to the permutations $\pi \in S_k$, in the obvious way. With this identification made, the above modified usual Wick formula becomes:

$$E(f_{i_1} \dots f_{i_k} f_{j_1}^* \dots f_{j_k}^*) = t^k \# \left\{ \pi \in S_k \mid i_{\pi(r)} = j_r, \forall r \right\}$$

Thus, we have reached to the formula in the statement, and we are done. \square

Finally, here is one more formulation of the Wick formula, useful as well:

THEOREM 4.33 (Wick formula 3). *Given independent variables f_i , each following the complex normal law G_t , with $t > 0$ being a fixed parameter, we have the formula*

$$E(f_{i_1} f_{j_1}^* \dots f_{i_k} f_{j_k}^*) = t^k \# \left\{ \pi \in S_k \mid i_{\pi(r)} = j_r, \forall r \right\}$$

for the non-vanishing joint moments of these variables.

PROOF. This follows from our second Wick formula, from Theorem 4.32, simply by permuting the terms, as to have an alternating sequence of plain and conjugate variables. Alternatively, we can start with Theorem 4.31, and then perform the same manipulations as in the proof of Theorem 4.32, but with the exponent being this time as follows:

$$K = \underbrace{\circ \bullet \circ \bullet \dots \circ \bullet}_{2k}$$

Thus, we are led to the conclusion in the statement. \square

Many other things can be said about the normal laws, and also about Rayleigh laws. We will be back to this, notably in chapter 15 below, when doing random matrices.

4e. Exercises

Strange and exciting chapter that we had here, with our explorations involving ships, drunken sailors and mysterious cats, and as exercises on all this, we have:

EXERCISE 4.34. *Do some abstract simulations, of your choice, for the CLT.*

EXERCISE 4.35. *Do as well some computer simulations for the CLT.*

EXERCISE 4.36. *Prove that the density is even when the odd moments vanish.*

EXERCISE 4.37. *Find a genius method for dealing with the Gauss integral.*

EXERCISE 4.38. *How to best integrate g_t , over a bounded interval $[a, b]$?*

EXERCISE 4.39. *Learn more about the Rayleigh laws, and their properties.*

EXERCISE 4.40. *Work out our RCLT program, outlined in the above.*

EXERCISE 4.41. *Do some probabilistic computations for O_N, U_N .*

As bonus exercise, learn some systematic probability, say from probabilists, or from physicists. Be aware though that the books there are also based on some cat knowledge.

Part II

Measure theory

*And it's no nay never
No nay never no more
Will I play the wild rover
No never no more*

CHAPTER 5

Measure theory

5a. Metric spaces

Welcome to measure theory, take two. We have seen in Part I a lot of interesting mathematics, mostly of probabilistic nature, coming as a complement to the basic calculus that you surely know, and often pushing us into developing measure theory.

So, this is what we will do here, in this Part II, developing measure theory, in a very standard, rigorous and mathematical way. Then in Part III we will talk about function spaces, by using measure theory tools, again in a very standard way, and with this being traditionally part of measure theory too. In short, expect now 200 pages of standard mathematics, which will be something useful, improving your knowledge of calculus.

As for mysterious probability, and other strange manipulations involving ∞ , in case you got addicted to that, do not worry, we will be back to this in Part IV, that time with even more exciting considerations, mixing measure theory, math and physics.

In short, you got it, this book is conceived a bit like a sandwich, with philosophical buns and meat content, so time to get into the meat. Getting started now, before even talking measures, we need spaces. And here, we have the following question:

QUESTION 5.1. *What is a space?*

And good question this is. There are of course a myriad answers to it, but remembering that we are analysts, let us consider instead a more modest question, as follows:

QUESTION 5.2. *What are the spaces where we can do analysis?*

But, this looks again like an excessively general question. Indeed, examples of such spaces include for instance \mathbb{R}, \mathbb{C} , then the powers $\mathbb{R}^N, \mathbb{C}^N$, and then all sorts of curves, surfaces and so on $X \subset \mathbb{R}^N, X \subset \mathbb{C}^N$. And with the story being not over here, because we should normally be able to allow $N = \infty$ too in all this. Plus, there might be other spaces X where we can do analysis, which do not necessarily embed into $\mathbb{R}^N, \mathbb{C}^N$.

In order to see more clearly what is going on, let us forget about spaces, and think instead at what doing analysis means. And here, for doing analysis, we obviously need

convergence, with the usual definition of convergence being as follows:

$$x_n \rightarrow x \iff d(x, x_n) \rightarrow 0$$

Long story short, for doing analysis on a space X , we need a distance on X . Thus, as answer to Question 5.2, which is something very intuitive, we have:

ANSWER 5.3. *In order to be able to do analysis on X , we need a distance on X :*

$$d : X \times X \rightarrow \mathbb{R}_+$$

Indeed, once we have d , we can talk about convergence, then continuity and so on.

Very good all this, and the question is now, what are the exact axioms on (X, d) . But these can only be the usual properties satisfied by the distances on $\mathbb{R}^N, \mathbb{C}^N$, and a bit of thinking here leads to the following definition, fine-tuning Answer 5.3:

DEFINITION 5.4. *A metric space is a set X with a distance function $d : X \times X \rightarrow \mathbb{R}_+$, having the following properties:*

- (1) $d(x, y) > 0$ if $x \neq y$, and $d(x, x) = 0$.
- (2) $d(x, y) = d(y, x)$.
- (3) $d(x, y) \leq d(x, z) + d(y, z)$.

As a basic example, we have \mathbb{R}^N , as well as any of its subsets $X \subset \mathbb{R}^N$. Indeed, the first two axioms are clear, and for the third axiom, we must prove that:

$$\sqrt{\sum_i (a_i + b_i)^2} \leq \sqrt{\sum_i a_i^2} + \sqrt{\sum_i b_i^2}$$

Now by raising to the square, this is the same as proving that:

$$\left(\sum_i a_i b_i \right)^2 \leq \left(\sum_i a_i^2 \right) \left(\sum_i b_i^2 \right)$$

But this latter inequality is one of the many equivalent formulations of the Cauchy-Schwarz inequality, that we know well from calculus, and which follows by using the fact that $f(t) = \sum_i (a_i + t b_i)^2$ being positive, its discriminant must be negative.

As another example, we have \mathbb{C}^N , as well as any of its subsets $X \subset \mathbb{C}^N$. Indeed, this follows either from $\mathbb{C}^N \simeq \mathbb{R}^{2N}$, or directly, along the lines of the above proof for \mathbb{R}^N . To be more precise, after some algebra, we are left with proving the following inequality:

$$\left| \sum_i a_i \bar{b}_i \right|^2 \leq \left(\sum_i |a_i|^2 \right) \left(\sum_i |b_i|^2 \right)$$

But this is the complex version of the Cauchy-Schwarz inequality, that we know well from calculus, and which follows by using the fact that the function $f(t) = \sum_i |a_i + t w b_i|^2$ with $t \in \mathbb{R}$ and $|w| = 1$ being positive, its discriminant must be negative.

Here is now another example, which at first looks new and interesting, but is in fact not new, because it appears as a subspace of a suitable \mathbb{R}^N :

PROPOSITION 5.5. *Given a finite set X , the following function is a metric on it, called discrete metric:*

$$d(x, y) = \begin{cases} 1 & \text{if } x \neq y \\ 0 & \text{if } x = y \end{cases}$$

This metric space is in fact the N -simplex, which can be realized as a subspace of \mathbb{R}^{N-1} , or, more conveniently, as a subspace of \mathbb{R}^N .

PROOF. There are several things going on here, the idea being as follows:

(1) First of all, the axioms from Definition 5.4 are trivially satisfied, and with the main axiom, namely the triangle inequality, basically coming from:

$$1 \leq 1 + 1$$

(2) At the level of examples, at $|X| = 1$ we obtain a point, at $|X| = 2$ we obtain a segment, at $|X| = 3$ we obtain an equilateral triangle, at $|X| = 4$ we obtain a regular tetrahedron, and so on. Thus, what we have in general, at $|X| = N$, is the arbitrary dimensional generalization of this series of geometric objects, called N -simplex.

(3) In what regards now the geometric generalization of the N -simplex, our above examples, namely segment, triangle, tetrahedron and so on, suggest to look for an embedding $X \subset \mathbb{R}^{N-1}$. Which is something which is certainly possible, but the computations here are quite complicated, involving a lot of trigonometry, as you can check yourself by studying the problem at $N = 4$, that is, parametrizing the regular tetrahedron in \mathbb{R}^3 .

(4) However, mathematics, or perhaps physics come to the rescue, via the idea “add a dimension, for getting smarter”. Indeed, when looking for an embedding $X \subset \mathbb{R}^N$ things drastically simplify, because we can simply take X to be the standard basis of \mathbb{R}^N :

$$X = \{e_1, \dots, e_N\}$$

Indeed, we have by definition $d(e_i, e_j) = 1$ for any $i \neq j$. So, we have solved our embedding problem, just like that, without doing any computations or trigonometry. \square

Moving ahead now with some theory, and allowing us a bit of slopiness, we have:

PROPOSITION 5.6. *We can talk about limits inside metric spaces X , by saying that*

$$x_n \rightarrow x \iff d(x_n, x) \rightarrow 0$$

and we can talk as well about continuous functions $f : X \rightarrow Y$, by requiring that

$$x_n \rightarrow x \implies f(x_n) \rightarrow f(x)$$

and with these notions in hand, all the basic results from the cases $X = \mathbb{R}, \mathbb{C}$ extend.

PROOF. All this is very standard, and we will leave this as an exercise, namely carefully checking what you learned in basic calculus, in relation with limits and continuity, in the cases $X = \mathbb{R}, \mathbb{C}$, and working out the metric space extensions of this. \square

More interestingly now, we can talk about open and closed sets inside metric spaces X , again in analogy with what we did for $X = \mathbb{R}, \mathbb{C}$, but with a whole lot of interesting new phenomena appearing. So, we will do this in detail. Let us start with:

DEFINITION 5.7. *Let X be a metric space.*

- (1) *The open balls are the sets $B_x(r) = \{y \in X | d(x, y) < r\}$.*
- (2) *The closed balls are the sets $\bar{B}_x(r) = \{y \in X | d(x, y) \leq r\}$.*
- (3) *$U \subset X$ is called open if for any $x \in U$ we have a ball $B_x(r) \subset U$.*
- (4) *$F \subset X$ is called closed if its complement $F^c \subset X$ is open.*

At the level of examples, you can quickly convince yourself, by working out a few of them, that our notions above coincide with the usual ones, that we know well, in the case $X = \mathbb{R}, \mathbb{C}$. We will be back to this later, with some general results in this sense, confirming all this. But for the moment, let us work out the basics. We first have the following result, clarifying some terminology issues from Definition 5.7:

PROPOSITION 5.8. *The open balls are open, and the closed balls are closed.*

PROOF. This might sound like a joke, but it is not one, because this is the kind of thing that we have to check. Fortunately, all this is elementary, as follows:

- (1) Given an open ball $B_x(r)$ and a point $y \in B_x(r)$, by using the triangle inequality we have $B_y(r') \subset B_x(r)$, with $r' = r - d(x, y)$. Thus, $B_x(r)$ is indeed open.
- (2) Given a closed ball $\bar{B}_x(r)$ and a point $y \in B_x(r)^c$, by using the triangle inequality we have $B_y(r') \subset B_x(r)^c$, with $r' = d(x, y) - r$. Thus, $\bar{B}_x(r)$ is indeed closed. \square

Here is now something more interesting, making the link with our intuitive understanding of the notion of closedness, coming from our experience so far with analysis:

THEOREM 5.9. *For a subset $F \subset X$, the following are equivalent:*

- (1) *F is closed in our sense, meaning that F^c is open.*
- (2) *We have $x_n \rightarrow x, x_n \in F \implies x \in F$.*

PROOF. We can prove this by double implication, as follows:

(1) \implies (2) Assume by contradiction $x_n \rightarrow x, x_n \in F$ with $x \notin F$. Since we have $x \in F^c$, which is open, we can pick a ball $B_x(r) \subset F^c$. But this contradicts our convergence assumption $x_n \rightarrow x$, so we are done with this implication.

(2) \implies (1) Assume by contradiction that F is not closed in our sense, meaning that F^c is not open. Thus, we can find $x \in F^c$ such that there is no ball $B_x(r) \subset F^c$. But with $r = 1/n$ this provides us with a point $x_n \in B_x(1/n) \cap F$, and since we have $x_n \rightarrow x$, this contradicts our assumption (2). Thus, we are done with this implication too. \square

Here is another basic theorem about open and closed sets:

THEOREM 5.10. *Let X be a metric space.*

- (1) *If U_i are open, then $\cup_i U_i$ is open.*
- (2) *If F_i are closed, then $\cap_i F_i$ is closed.*
- (3) *If U_1, \dots, U_n are open, then $\cap_i U_i$ is open.*
- (4) *If F_1, \dots, F_n are closed, then $\cup_i F_i$ is closed.*

Moreover, both (3) and (4) can fail for infinite intersections and unions.

PROOF. We have several things to be proved, the idea being as follows:

(1) This is clear from definitions, because any point $x \in \cup_i U_i$ must satisfy $x \in U_i$ for some i , and so has a ball around it belonging to U_i , and so to $\cup_i U_i$.

(2) This follows from (1), by using the following well-known set theory formula:

$$\left(\bigcup_i U_i \right)^c = \bigcap_i U_i^c$$

(3) Given an arbitrary point $x \in \cap_i U_i$, we have $x \in U_i$ for any i , and so we have a ball $B_x(r_i) \subset U_i$ for any i . Now with this in hand, let us set:

$$B = B_x(r_1) \cap \dots \cap B_x(r_n)$$

As a first observation, this is a ball around x , $B = B_x(r)$, of radius given by:

$$r = \min(r_1, \dots, r_n)$$

But this ball belongs to all the U_i , and so belongs to their intersection $\cap_i U_i$. We conclude that the intersection $\cap_i U_i$ is open, as desired.

(4) This follows from (3), by using the following well-known set theory formula:

$$\left(\bigcap_i U_i \right)^c = \bigcup_i U_i^c$$

(5) Finally, in what regards the counterexamples at the end, these can be both found on \mathbb{R} , and we will leave this as an instructive exercise. \square

Still in relation with the open and closed sets, let us formulate as well:

DEFINITION 5.11. *Let X be a metric space, and $E \subset X$ be a subset.*

- (1) *The interior $E^\circ \subset E$ is the set of points $x \in E$ which admit around them open balls $B_x(r) \subset E$.*
- (2) *The closure $E \subset \bar{E}$ is the set of points $x \in X$ which appear as limits of sequences $x_n \rightarrow x$, with $x \in E$.*

These notions are quite interesting, because they make sense for any set E . That is, when E is open, that is open and end of the story, and when E is closed, that is closed and end of the story. In general, however, a set $E \subset X$ is not open or closed, and what we can best do to it, in order to study it with our tools, is to “squeeze” it, as follows:

$$E^\circ \subset E \subset \bar{E}$$

In practice now, in order to use the above notions, we need to know a number of things, including that fact that E open implies $E^\circ = E$, the fact that E closed implies $\bar{E} = E$, and many more such results. But all this can be done, and the useful statement here, summarizing all that we need to know about interiors and closures, is as follows:

THEOREM 5.12. *Let X be a metric space, and $E \subset X$ be a subset.*

- (1) *The interior $E^\circ \subset E$ is the biggest open set contained in E .*
- (2) *The closure $E \subset \bar{E}$ is the smallest closed set containing E .*

PROOF. We have several things to be proved, the idea being as follows:

(1) Let us first prove that the interior E° is open. For this purpose, pick $x \in E^\circ$. We know that we have a ball $B_x(r) \subset E$, and since this ball is open, it follows that we have $B_x(r) \subset E^\circ$. Thus, the interior E° is open, as claimed.

(2) Let us prove now that the closure \bar{E} is closed. For this purpose, we will prove that the complement \bar{E}^c is open. So, pick $x \in \bar{E}^c$. Then x cannot appear as a limit of a sequence $x_n \rightarrow x$ with $x_n \in E$, so we have a ball $B_x(r) \subset \bar{E}^c$, as desired.

(3) Finally, the maximality and minimality assertions regarding E° and \bar{E} are both routine too, coming from definitions, and we will leave them as exercises. \square

As a continuation of this, we can talk as well about density, as follows:

DEFINITION 5.13. *We say that a subset $E \subset X$ is dense when:*

$$\bar{E} = X$$

That is, any point of X must appear as a limit of points of E .

Obviously, this is something which is in tune with what we know so far from this book, and with the intuitive notion of density. As a basic example, we have $\bar{\mathbb{Q}} = \mathbb{R}$.

5b. Compactness

Again in analogy with what we know about $X = \mathbb{R}, \mathbb{C}$, we can talk about compact sets. However, things here are quite tricky, in the general metric space framework, substantially deviating from what we know, and we will do this in detail. Let us start with:

DEFINITION 5.14. A set $K \subset X$ is called compact if any cover with open sets

$$K \subset \bigcup_i U_i$$

has a finite subcover, $K \subset (U_{i_1} \cup \dots \cup U_{i_n})$.

This definition, which is probably new to you, might seem overly abstract, but our claim is that this is the correct definition, and that there is no way of doing otherwise. Let us start with some examples, with $X = \mathbb{R}$. The situation here is as follows:

(1) A point is obviously compact, and we can choose that finite subcover with $n = 1$. Similarly, 2 points are compact, and we can choose the subcover with $n = 2$. More generally, N points are compact, and we can choose the subcover with $n = N$.

(2) In contrast, the set $\mathbb{N} \subset \mathbb{R}$ is not compact, because we can cover it with of a suitable union of small open intervals around each point, and this open cover has no finite subcover. For the same reasons, the set $\{1/n | n \in \mathbb{N}\}$ is not compact either.

(3) However, and here comes an interesting point, the following set is compact:

$$K = \left\{ \frac{1}{n} \mid n \in \mathbb{N} \right\} \cup \{0\}$$

Indeed, any open cover of it $\cup_i U_i$ has to cover 0, and by selecting an open set U_i covering 0, this set U_i will cover the whole K , except for finitely many points, due to $1/n \rightarrow 0$. But these finitely points left, say N of them, can be covered by suitable sets U_{i_1}, \dots, U_{i_N} , and by adding this family to the set U_i , we have our finite subcover.

As a conclusion to this, Definition 5.14 seems to be in tune with what we know about the compact subsets $K \subset \mathbb{R}$, namely that these are the sets which are closed and bounded. However, and here comes our point, such things are wrong in general, due to:

PROPOSITION 5.15. Given an infinite set X with the discrete distance on it, namely $d(p, q) = 1 - \delta_{pq}$, which can be modeled as the basis of a suitable Hilbert space,

$$X = \{e_x\}_{x \in X} \subset l^2(X)$$

this set is closed and bounded, but not compact.

PROOF. Here the first part, regarding the modeling of X , that we will actually not need here, is something that we already know. Regarding now the second part:

(1) X being the total space, it is closed. In fact, since the points of X are open, any subset $E \subset X$ is open, and by taking complements, any set $E \subset X$ is closed as well.

(2) X is also bounded, because all distances are smaller than 1.

(3) However, our set X is not compact, because its points being open, as noted above, $X = \cup_{x \in X} \{x\}$ is an open cover, having no finite subcover. \square

Let us develop now the theory of compact sets, and see what we get. We first have:

PROPOSITION 5.16. *The following hold:*

- (1) *Compact implies closed.*
- (2) *Closed inside compact is compact.*
- (3) *Compact intersected with closed is compact.*

PROOF. These assertions are all clear from definitions, as follows:

(1) Assume that $K \subset X$ is compact, and let us prove that K is closed. For this purpose, we will prove that K^c is open. So, pick $p \in K^c$. For any $q \in K$ we set $r = d(p, q)/3$, and we consider the following balls, separating p and q :

$$U_q = B_p(r) \quad , \quad V_q = B_q(r)$$

We have then $K \subset \cup_{q \in K} V_q$, so we can pick a finite subcover, as follows:

$$K \subset (V_{q_1} \cup \dots \cup V_{q_n})$$

With this done, consider the following intersection:

$$U = U_{q_1} \cap \dots \cap U_{q_n}$$

This intersection is then a ball around p , and since this ball avoids V_{q_1}, \dots, V_{q_n} , it avoids the whole K . Thus, we have proved that K^c is open at p , as desired.

(2) Assume that $F \subset K$ is closed, with $K \subset X$ being compact. For proving our result, we can assume, by replacing X with K , that we have $X = K$. In order to prove now that F is compact, consider an open cover of it, as follows:

$$F \subset \bigcup_i U_i$$

By adding the set F^c , which is open, to this cover, we obtain a cover of K . Now since K is compact, we can extract from this a finite subcover Ω , and there are two cases:

- If $F^c \in \Omega$, by removing F^c from Ω we obtain a finite cover of F , as desired.
- If $F^c \notin \Omega$, we are done too, because in this case Ω is a finite cover of F .

(3) This follows from (1) and (2), because if $K \subset X$ is compact, and $F \subset X$ is closed, then $K \cap F \subset K$ is closed inside a compact, so it is compact. \square

As a second batch of results, which are useful as well, we have:

PROPOSITION 5.17. *The following hold:*

- (1) *If $K_i \subset X$ are compact, satisfying $K_{i_1} \cap \dots \cap K_{i_n} \neq \emptyset$, then $\cap_i K_i \neq \emptyset$.*
- (2) *If $K_1 \supset K_2 \supset K_3 \supset \dots$ are non-empty compacts, then $\cap_i K_i \neq \emptyset$.*
- (3) *If K is compact, and $E \subset K$ is infinite, then E has a limit point in K .*
- (4) *If K is compact, any sequence $\{x_n\} \subset K$ has a limit point in K .*
- (5) *If K is compact, any $\{x_n\} \subset K$ has a subsequence which converges in K .*

PROOF. Again, these are elementary results, which can be proved as follows:

(1) Assume by contradiction $\cap_i K_i = \emptyset$, and let us pick $K_1 \in \{K_i\}$. Since any $x \in K_1$ is not in $\cap_i K_i$, there is an index i such that $x \in K_i^c$, and we conclude that we have:

$$K_1 \subset \bigcup_{i \neq 1} K_i^c$$

But this can be regarded as being an open cover of K_1 , that we know to be compact, so we can extract from it a finite subcover, as follows:

$$K_1 \subset (K_{i_1}^c \cup \dots \cup K_{i_n}^c)$$

But this contradicts our non-empty intersection assumption, and we are done.

(2) This is a particular case of (1), proved above.

(3) We prove this by contradiction. So, assume that E has no limit point in K . This means that any $p \in K$ can be isolated from the rest of E by a certain open ball $V_p = B_p(r)$, and in both the cases that can appear, $p \in E$ or $p \notin E$, we have:

$$|V_p \cap E| = 0, 1$$

Now observe that these sets V_p form an open cover of K , and so of E . But due to $|V_p \cap E| = 0, 1$ and to $|E| = \infty$, this open cover of E has no finite subcover. Thus the same cover, regarded now as cover of K , has no finite subcover either, contradiction.

(4) This follows from (3) that we just proved, with $E = \{x_n\}$.

(5) This is a reformulation of (4), that we just proved. □

Getting now to more exciting theory, here is a key result about compactness:

THEOREM 5.18. *For a subset $K \subset \mathbb{R}^N$, the following are equivalent:*

- (1) *K is closed and bounded.*
- (2) *K is compact.*
- (3) *Any infinite subset $E \subset K$ has a limiting point in K .*

PROOF. This is something quite tricky, the idea being as follows:

(1) \implies (2) As a first task, let us prove that any product of closed intervals is indeed compact. We can assume by linearity that we are dealing with the unit cube:

$$C_1 = \prod_{i=1}^N [0, 1] \subset \mathbb{R}^N$$

In order to prove that C_1 is compact, we proceed by contradiction. So, assume that we have an open cover as follows, having no finite subcover:

$$C_1 \subset \bigcup_i U_i$$

Now let us cut C_1 into 2^N small cubes, in the obvious way, over the N coordinate axes. Then at least one of these small cubes, which are all covered by $\cup_i U_i$ too, has no finite subcover. So, let us call $C_2 \subset C_1$ one of these small cubes, having no finite subcover:

$$C_2 \subset \bigcup_i U_i$$

We can then cut C_2 into 2^N small cubes, and by the same reasoning, we obtain a smaller cube $C_3 \subset C_2$ having no finite subcover. And so on by recurrence, and we end up with a decreasing sequence of cubes, as follows, having no finite subcover:

$$C_1 \supset C_2 \supset C_3 \supset \dots$$

Now since these decreasing cubes have edge size $1, 1/2, 1/4, \dots$, their intersection must be a point. So, let us call p this point, defined by the following formula:

$$\{p\} = \bigcap_k C_k$$

But this point p must be covered by $\cup_i U_i$, so we can find an index i such that $p \in U_i$. Now observe that U_i must contain a whole ball around p , and so starting from a certain $K \in \mathbb{N}$, all the cubes C_k will be contained in this ball, and so in U_i :

$$C_k \subset U_i \quad , \quad \forall k \geq K$$

But this is a contradiction, because C_K , and in fact the smaller cubes C_k with $k > K$ as well, were assumed to have no finite subcover. Thus, we proved our claim. But with this claim in hand, the result is now clear. Indeed, assuming that $K \subset \mathbb{R}^N$ is closed and bounded, we can view it as a subset as a suitable big cube, of the following form:

$$K \subset \prod_{i=1}^N [-M, M] \subset \mathbb{R}^N$$

But, what we have here is a closed subset inside a compact set, which by Proposition 5.16 follows to be compact, as desired.

(2) \implies (3) This is something that we already know, not needing $K \subset \mathbb{R}^N$.

(3) \implies (1) We have to prove that K as in the statement is both closed and bounded, and we can do both these things by contradiction, as follows:

– Assume first that K is not closed. But this means that we can find a point $x \notin K$ which is a limiting point of K . Now let us pick $x_n \in K$, with $x_n \rightarrow x$, and consider the set $E = \{x_n\}$. According to our assumption, E must have a limiting point in K . But this limiting point can only be x , which is not in K , contradiction.

– Assume now that K is not bounded. But this means that we can find points $x_n \in K$ satisfying $\|x_n\| \rightarrow \infty$, and if we consider the set $E = \{x_n\}$, then again this set must have a limiting point in K , which is impossible, so we have our contradiction, as desired. \square

So long for compactness. As a last piece of general topology, in our metric space framework, we can talk as well about connectedness, as follows:

DEFINITION 5.19. *We can talk about connected sets $E \subset X$, as follows:*

- (1) *We say that E is connected if it cannot be separated as $E = E_1 \cup E_2$, with the components E_1, E_2 satisfying $E_1 \cap \bar{E}_2 = \bar{E}_1 \cap E_2 = \emptyset$.*
- (2) *We say that E is path connected if any two points $p, q \in E$ can be joined by a path, meaning a continuous $f : [0, 1] \rightarrow X$, with $f(0) = p$, $f(1) = q$.*

All this looks a bit technical, and indeed it is. To start with, (1) is something quite natural, but the separation condition there $E_1 \cap \bar{E}_2 = \bar{E}_1 \cap E_2 = \emptyset$ can be weakened into $E_1 \cap E_2 = \emptyset$, or strengthened into $\bar{E}_1 \cap \bar{E}_2 = \emptyset$, depending on purposes, and with our (1) as formulated being the good compromise, for most purposes. As for (2), this condition is obviously something stronger, and we have in fact the following implications:

$$\text{convex} \implies \text{path connected} \implies \text{connected}$$

Anyway, leaving aside the discussion here, once all these questions clarified, the idea is that any set E can be written as a disjoint union of connected components:

$$E = \bigsqcup_i E_i$$

Getting back now to more concrete things, remember that we are here in this book for studying functions, and doing calculus. And, regarding functions, we have:

THEOREM 5.20. *Assuming that $f : X \rightarrow Y$ is continuous, the following happen:*

- (1) *If O is open, then $f^{-1}(O)$ is open.*
- (2) *If C is closed, then $f^{-1}(C)$ is closed.*
- (3) *If K is compact, then $f(K)$ is compact.*
- (4) *If E is connected, then $f(E)$ is connected.*

PROOF. This is something fundamental, which can be proved as follows:

(1) This is clear from the definition of continuity, written with ε, δ . In fact, the converse holds too, in the sense that if $f^{-1}(\text{open}) = \text{open}$, then f must be continuous.

(2) This follows from (1), by taking complements. And again, the converse holds too, in the sense that if $f^{-1}(\text{closed}) = \text{closed}$, then f must be continuous.

(3) Indeed, given an open cover $f(K) \subset \cup_i U_i$, we have an open cover $K \subset \cup_i f^{-1}(U_i)$, and so by compactness of K , a finite subcover $K \subset f^{-1}(U_{i_1}) \cup \dots \cup f^{-1}(U_{i_n})$, and so finally a finite subcover $f(K) \subset U_{i_1} \cup \dots \cup U_{i_n}$, as desired.

(4) This can be proved via the same trick as for (3). Indeed, any separation of $f(E)$ into two parts can be returned via f^{-1} into a separation of E into two parts, contradiction. \square

Observe the power of (3,4) in the above result, which among others prove the mean value theorem, which is something non-trivial. Good mathematics that we have here.

5c. Measured spaces

We are now ready to attack measure theory. We first need measurable sets, and on the bottom line, we would like our open and closed sets, and their “combinations”, to be measurable. But these “combinations” can be understood and axiomatized, and are called Borel sets. So, this will be the idea in what follows, talking about abstract measurable sets, then about Borel sets, and then coming with measures, and integration theory.

Let us start with the abstract measurable sets. We have here the following definition, which is something very general, not making reference to any metric on our space X , nor making reference to any measure on X , measuring these measurable sets:

DEFINITION 5.21. *An abstract measured space is a set X , given with a set of subsets $M \subset P(X)$, called measurable sets, which form an algebra, in the sense that:*

- (1) $\emptyset, X \in M$.
- (2) $E \in M \implies E^c \in M$.
- (3) M is stable under countable unions and intersections.

Obviously, this is something quite abstract. If the last axiom, (3), is only satisfied for the finite unions and intersections, we say that $M \subset P(X)$ is a finite algebra.

As a first observation, some of the axioms above are redundant. Indeed, assuming that (2) holds, we have the following equivalence, which can help in verifying (1):

$$\emptyset \in M \iff X \in M$$

The same goes for the axiom (3), with only one of the conditions there being in need to be verified, in practice, and this due to the following formula:

$$\left(\bigcup_i E_i \right)^c = \bigcap_i E_i^c$$

However, it is most convenient to write Definition 5.21 as above, in a symmetric way with respect to the two operations involved, namely union and intersection.

At the level of examples of abstract measured spaces, there are many of them, and more on this in a moment, usually coming via the following result:

THEOREM 5.22. *Given a set of subsets $S \subset P(X)$, there is a smallest algebra*

$$M = \bar{S}$$

containing it, called algebra generated by S .

PROOF. This can be viewed in two possible ways, as follows:

(1) According to the axioms in Definition 5.21, what we have to do in order to construct $M = \bar{S}$ is to start with S , then add \emptyset, X to it, along with the complements E^c of all the

sets $E \in \mathcal{S}$, and then take countable unions and intersections of such sets. And, some elementary verifications show that what we get in this way is indeed an algebra.

(2) Alternatively, we can define $M = \bar{\mathcal{S}}$ as being the intersection of all algebras containing \mathcal{S} , and with the remark that we have at least one such algebra, namely $P(X)$ itself. It is then clear from definitions that M is an algebra, as desired. \square

Getting now to the concrete examples of abstract measured spaces, we have:

DEFINITION 5.23. *Any metric space X is automatically an abstract measured space, with the algebra of measurable sets being*

$$B = \bar{O}$$

that is, the smallest algebra containing the open sets, called Borel algebra of X .

Observe that the Borel sets include all open sets, all closed sets, as well as all countable unions of closed sets, and all countable intersections of open sets. As an example here, in the case $X = \mathbb{R}$, with its usual topology, all kinds of intervals are Borel sets:

$$(a, b), [a, b], (a, b], [a, b) \in B$$

Indeed, the first interval is open, and the second one is closed, so these are certainly Borel sets. As for the third and fourth intervals, these appear as countable unions of closed intervals, or as countable intersections of open intervals, so they are Borel too.

Getting back now to the general case, following Definition 5.21, we have:

DEFINITION 5.24. *Given a measured space (X, M) , a measure on it is a function*

$$\mu : M \rightarrow [0, \infty]$$

which is countably additive, in the sense that we have

$$\mu \left(\bigcup_{i=1}^{\infty} E_i \right) = \sum_{i=1}^{\infty} \mu(E_i)$$

for any countable family of disjoint measurable sets $E_i \in M$.

This definition, which will play a key role in what follows, is something quite tricky, and there are several comments to be made about it, as follows:

(1) Obviously, what we axiomatized above are the positive measures, and for the moment this will do, and we will omit the term “positive”. More on this later, when we will talk about differences $\mu - \eta$ of such measures, and about complex measures too.

(2) In contrast, we did not assume that our measures are finite, and this because many interesting spaces, such as $X = \mathbb{R}$ itself, are naturally of infinite measure, $\mu(X) = \infty$. In the case where the measure μ happens to be bounded, up to a rescaling we can assume $\mu(X) = 1$, and we say in this case that we have a probability measure on X .

(3) Yet another subtlety comes in relation with the countable additivity condition at the end, the point being that for many measured spaces X , the finite additivity condition is something strictly weaker, and leads to a wrong theory. More on this later.

Looking at what we have so far, Definition 5.23 and Definition 5.24, many natural questions appear, and leaving aside anything too specialized, we are led to:

QUESTION 5.25. *Given a metric space X , such as $X = \mathbb{R}$, or $X = \mathbb{R}^N$, how to construct a measure on it? Also, once we have such a measure, how to integrate the functions $f : X \rightarrow \mathbb{R}$, or $f : X \rightarrow \mathbb{C}$, with respect to this measure?*

As you can see, we have two questions here, and none is trivial:

(1) Indeed, regarding the first question, this is something that we do not know yet how to solve, even this even for very simple spaces like $X = \mathbb{R}$. Indeed, while we certainly know how to measure the real intervals, simply by setting $\mu(a, b) = b - a$, and then unions of such intervals too, by decomposing them into disjoint unions of intervals, and making sums, nothing guarantees that we can measure any Borel set $E \in B$, in this way.

(2) As for the second question, experience with the usual Riemann integral, that you know well from calculus, shows that such things can be quite tricky too. Observe also that this second question is more general than the first one, because we have $\mu(E) = \int \chi_E$ for any $E \in M$, so if we know how to integrate, we know how to measure.

In short, we are facing non-trivial questions here, and we need a plan. And, perhaps a bit surprisingly, the best plan in order to deal with Question 5.25 is as follows:

PLAN 5.26. *We will jointly develop measure and integration theory, as follows:*

- (1) *We will first keep staying abstract, and understand how the functions $f : X \rightarrow \mathbb{R}$, or $f : X \rightarrow \mathbb{C}$, can be integrated, with respect to an abstract measure.*
- (2) *With this understood, we will conclude that the integration over $X = \mathbb{R}$ can only be a straightforward extension of the usual Riemann integration.*
- (3) *But, in the end, this will enable us to both measure all the Borel sets $E \subset \mathbb{R}$, and to integrate the measurable functions $f : \mathbb{R} \rightarrow \mathbb{C}$.*
- (4) *Finally, we will discuss how to deal with $X = \mathbb{R}^N$ too, and with other product spaces $X = Y \times Z$, both measure theory and integration.*

As you can see, all this is quite tricky, the main idea behind this plan being functional analysis, that is, using functions and their integrals, instead of just fighting with abstract measure theory first, and looking at functions and their integrals afterwards.

Getting started now, let us first talk about measurable functions, in the general context of Definition 5.21, with no measure involved. We have here the following notion:

DEFINITION 5.27. *Given a measured space X , and a metric space Y , a function*

$$f : X \rightarrow Y$$

is called measurable when it satisfies the following condition:

$$U \in \mathcal{O} \implies f^{-1}(U) \in \mathcal{M}$$

When X comes with a measure, we also call such functions integrable.

Obviously, this is something simplified, because for doing abstract measurability theory, our spaces should be abstract measured spaces in the sense of Definition 5.21, and so the functions that we should normally care about should be functions $f : X \rightarrow Y$, with both X, Y being abstract measured spaces. However, in view of Definition 5.23, this is more or less that what we are doing here, by restricting the attention to the target spaces Y which are metric, with this being the case that really matters.

Many things can be said about the measurable functions. We first have:

PROPOSITION 5.28. *The measurable functions have the following properties:*

- (1) *If $f : X \rightarrow Y$ is measurable and $g : Y \rightarrow Z$ is continuous, $g \circ f$ is measurable.*
- (2) *If $f, g : X \rightarrow \mathbb{R}$ are both measurable and $h : \mathbb{R}^2 \rightarrow Y$ is continuous, then the function $x \rightarrow h(f(x), g(x))$ is measurable.*
- (3) *$f : X \rightarrow \mathbb{C}$ is measurable precisely when $\operatorname{Re}(f), \operatorname{Im}(f) : X \rightarrow \mathbb{R}$ are measurable. In this case, the function $|f| : X \rightarrow \mathbb{R}$ is measurable too.*
- (4) *If $f, g : X \rightarrow \mathbb{C}$ are measurable, then so are $f + g, fg : X \rightarrow \mathbb{C}$.*

PROOF. This is something very standard, the idea being as follows:

- (1) This is clear from definitions, because we have:

$$U \in \mathcal{O} \implies g^{-1}(U) \in \mathcal{O} \implies f^{-1}(g^{-1}(U)) \in \mathcal{M}$$

(2) By using (1), it is enough to check that the function $k(x) = (f(x), g(x))$ is measurable. But this follows by writing any open set $U \subset \mathbb{R}^2$ as a union of rectangles. Indeed, the preimage of each rectangle $R = I \times J$ is measurable, as shown by:

$$k^{-1}(R) = (f, g)^{-1}(I \times J) = f^{-1}(I) \cap g^{-1}(J) \in \mathcal{M}$$

But then, with this in hand, since any open set $U \subset \mathbb{R}^2$ can be written as a union of rectangles, it follows that $k^{-1}(U)$ is measurable, as desired.

- (3) This follows indeed by using (2), with the following functions:

$$h(z) = z, \operatorname{Re}(z), \operatorname{Im}(z), |z|$$

- (4) In the real case, $f, g : X \rightarrow \mathbb{R}$, the result follows by using (2), with:

$$h(x, y) = x + y, xy$$

Then, the result can be extended to the complex case, $f, g : X \rightarrow \mathbb{C}$, by using (3). \square

Next, we have the following useful characterization of the real measurable functions:

PROPOSITION 5.29. *A function $f : X \rightarrow [-\infty, \infty]$ is measurable precisely when*

$$f^{-1}(I) \in M$$

for any interval of type $I = (a, \infty]$, with $a \in \mathbb{R}$.

PROOF. Consider the following set, which is easily seen to be an algebra:

$$\Omega = \left\{ E \subset [-\infty, \infty] \mid f^{-1}(E) \in M \right\}$$

We want to prove that Ω contains all the open sets, and this can be done as follows:

(1) Pick $a \in \mathbb{R}$, and then pick an increasing sequence $a_n \rightarrow a$. We have then the following formula, which shows that we have $[-\infty, a) \in \Omega$:

$$[-\infty, a) = \bigcup_n [-\infty, a_n) = \bigcup_n (a_n, \infty]^c \in \Omega$$

(2) But with this in hand, we obtain, for any $a < b$, that we have:

$$(a, b) = [-\infty, b) \cap (a, \infty] \in \Omega$$

(3) Now since any open set $U \subset [-\infty, \infty]$ can be written as a union of open intervals, we conclude that we have $U \in \Omega$, as desired. \square

Getting now to limits of measurable functions, we have the following result:

PROPOSITION 5.30. *Given an abstract measured space X , the measurable functions $f : X \rightarrow [-\infty, \infty]$ have the following properties:*

- (1) *If f_n are measurable, so are $g = \sup_n f_n$, and $h = \limsup_n f_n$.*
- (2) *If f_n are measurable, so are $k = \inf_n f_n$, and $l = \liminf_n f_n$.*
- (3) *If $f_n \rightarrow f$ and f_n are measurable, then f is measurable.*
- (4) *If f, g are measurable, so are $h = \min(f, g)$ and $k = \max(f, g)$.*
- (5) *If f is measurable, so are $f^+ = \max(f, 0)$ and $f^- = -\min(f, 0)$.*

PROOF. This is again something very standard, the idea being as follows:

(1) For the function $g = \sup_n f_n$ we can use the measurability criterion from Proposition 5.29, along with the following fact, valid for any $a \in \mathbb{R}$:

$$g = \sup_n f_n \implies g^{-1}(a, \infty] = \bigcup_n f_n^{-1}(a, \infty]$$

By symmetry we obtain that the function $k = \inf_n f_n$ is measurable as well. But with these results in hand, the last assertion follows too, by using the following formula:

$$\limsup_n f_n = \inf_k \left(\sup_{n \geq k} f_n \right)$$

(2) Here the fact that $k = \inf_n f_n$ is measurable was already proved in the above, and for $l = \liminf_n f_n$ we can use the same argument, symmetry, or the following formula:

$$\liminf_n f_n = \sup_k \left(\inf_{n \geq k} f_n \right)$$

(3) This follows indeed from (1), or from (2), and from the following fact:

$$f_n \rightarrow f \implies f = \inf_n f_n = \sup_n f_n$$

(4) This is a trivial application of (1) and (2).

(5) This follows from (4), and from the fact that if f is measurable, so is $-f$. \square

As a main result now regarding the measurable functions, which will be of key importance in what follows, we have the following statement, with the convention that a step function means a function which takes finitely many values:

THEOREM 5.31. *Given a measurable function $f : X \rightarrow [0, \infty]$, we can write*

$$f(x) = \lim_{n \rightarrow \infty} \varphi_n(x)$$

with $0 \leq \varphi_1 \leq \varphi_2 \leq \dots \leq f$, and with each φ_i being a measurable step function.

PROOF. As a first observation, the converse holds too, thanks to the results in Proposition 5.30. Regarding now the proof, this goes as follows:

(1) First, it is clear by drawing a picture that we can approximate the identity of $[0, \infty]$ with step functions as in the statement. That is, we can obviously find an increasing sequence of measurable step functions $0 \leq \psi_1 \leq \psi_2 \leq \dots \leq id$, satisfying:

$$\lim_{n \rightarrow \infty} \psi_n(x) = x$$

(2) Now let us set $\varphi_n = \psi_n \circ f$. According to the limiting formula above, we have:

$$\lim_{n \rightarrow \infty} \varphi_n(x) = f(x)$$

On the other hand, by using the measurability criterion from Proposition 5.29, it follows that our truncation functions $\varphi_n = \psi_n \circ f$ are measurable, as desired. \square

5d. Integration theory

Good news, with the above general theory understood, we can now integrate functions, by following the good old method of Riemann, that you know well from one-variable calculus. To be more precise, we have the following result, which is of course stated a bit informally, with some of the details being left to you, as an instructive exercise:

THEOREM 5.32. *We can integrate the measurable functions $f : X \rightarrow \mathbb{R}_+$ by setting*

$$\int_X f(x) d\mu(x) = \sup_{0 \leq \varphi \leq f} \int_X \varphi(x) d\mu(x)$$

with sup over measurable step functions, then extend this by linearity.

PROOF. This is something very standard, and we will leave the clarification of all this, both precise statement, and proof, as an instructive exercise. To be more precise:

(1) We can certainly integrate the step functions $\varphi : X \rightarrow \mathbb{R}_+$, by writing each such function as a linear combination of characteristic functions, as follows:

$$\varphi = \sum_i \lambda_i \chi_{E_i}$$

Indeed, with this formula in hand, we can integrate our function φ , as follows:

$$\int_X \varphi(x) d\mu(x) = \sum_i \lambda_i \mu(E_i)$$

The integral of step functions constructed in this way has then all the linearity and positivity properties that you might expect, and behaves well with respect to limits.

(2) Next, consider an arbitrary measurable function $f : X \rightarrow [0, \infty]$. We know from Theorem 5.31 that we can write this function as an increasing limit, as follows, with $0 \leq \varphi_1 \leq \varphi_2 \leq \dots \leq f$, and with each φ_i being a measurable step function:

$$f(x) = \lim_{n \rightarrow \infty} \varphi_n(x)$$

But this suggests to define the integral of f by the formula in the statement, namely:

$$\int_X f(x) d\mu(x) = \sup_{0 \leq \varphi \leq f} \int_X \varphi(x) d\mu(x)$$

Indeed, we can see that the integral constructed in this way has all the linearity and positivity properties that you might expect, and behaves well with respect to limits. \square

More in detail now, here are some basic properties of the integrals, which are very similar to those of the Riemann integral, that you know from calculus:

PROPOSITION 5.33. *The integrals of measurable functions $f : X \rightarrow \mathbb{R}_+$ have the following properties:*

- (1) $f \leq g$ implies $\int f \leq \int g$.
- (2) $E \subset F$ implies $\int_E f \leq \int_F f$.
- (3) $\int f + g = \int f + \int g$.
- (4) $\int \lambda f = \lambda \int f$.

PROOF. All this is indeed very standard, all routine verifications. \square

Summarizing, we know how to integrate the real positive functions, in our abstract measure theory setting. In general, we can use the following formula:

$$\int_X (f - g)(x) d\mu(x) = \int_X f(x) d\mu(x) - \int_X g(x) d\mu(x)$$

We can integrate as well the complex functions, by setting:

$$\int_X (f + ig)(x) d\mu(x) = \int_X f(x) d\mu(x) + i \int_X g(x) d\mu(x)$$

All this is, indeed, very standard, exactly as for the Riemann integral. Let us record these findings as an upgrade of Theorem 5.32, as follows, once again with the statement being a bit informal, and with some of the details being left as instructive exercises:

THEOREM 5.34. *We can integrate the measurable functions $f : X \rightarrow \mathbb{C}$ by setting*

$$\int_X f(x) d\mu(x) = \sup_{0 \leq \varphi \leq f} \int_X \varphi(x) d\mu(x)$$

with sup over measurable step functions, for $f : X \rightarrow \mathbb{R}_+$ then extend this by linearity.

PROOF. This follows indeed from what we already know from Theorem 5.32, and from the above discussion, exactly as for the Riemann integral. \square

Many other things can be said here, following Lebesgue, Fatou and others. We first have the following result, regarding the monotone convergence, due to Lebesgue:

THEOREM 5.35 (Lebesgue). *Given an increasing sequence of measurable functions*

$$0 \leq f_1 \leq f_2 \leq \dots \leq \infty$$

which converges pointwise, $f_n \rightarrow f$, their limit is measurable, and we have

$$\int_X f_n(x) d\mu(x) \rightarrow \int_X f(x) d\mu(x)$$

for any positive measure on X .

PROOF. This is indeed something very standard, the idea being as follows:

(1) We first have the following obvious implication, showing that the sequence of integrals on the right converges, to a certain number in $[0, \infty]$:

$$f_1 \leq f_2 \leq \dots \implies \int_X f_1(x) d\mu(x) \leq \int_X f_2(x) d\mu(x) \leq \dots$$

Moreover, since we have $f_n \rightarrow f$, it follows that we have the following inequality:

$$\lim_{n \rightarrow \infty} \int_X f_n(x) d\mu(x) \leq \int_X f(x) d\mu(x)$$

(2) In order to prove now the reverse inequality, pick a measurable simple function $0 \leq \varphi \leq f$, pick also a number $c \in (0, 1)$, and consider the following sets:

$$E_n = \left\{ x \in X \mid f_n(x) \geq c\varphi(x) \right\}$$

These sets are then measurable, we have $E_1 \subset E_2 \subset \dots$, and our claim is that:

$$X = \bigcup_n E_n$$

Indeed, given $x \in X$, if $f(x) = 0$ then $x \in E_1$ and things fine. Otherwise, we have $f(x) > 0$, and so $f(x) > c\varphi(x)$, since $c < 1$, and so $x \in E_n$ for some n , as desired.

(3) Now observe that, with $0 \leq \varphi \leq f$ and $c \in (0, 1)$ as above, we have:

$$\int_X f_n(x) d\mu(x) \geq \int_{E_n} f_n(x) d\mu(x) \geq c \int_{E_n} \varphi(x) d\mu(x)$$

By taking the limit of this estimate, with $n \rightarrow \infty$, we obtain:

$$\lim_{n \rightarrow \infty} \int f_n(x) d\mu(x) \geq c \int_X \varphi(x) d\mu(x)$$

Now with $c \rightarrow 1$, we obtain from this the following estimate:

$$\lim_{n \rightarrow \infty} \int f_n(x) d\mu(x) \geq \int_X \varphi(x) d\mu(x)$$

But this being true for any simple function $0 \leq \varphi \leq f$, we conclude that we have:

$$\lim_{n \rightarrow \infty} \int f_n(x) d\mu(x) \geq \int_X f(x) d\mu(x)$$

Thus we have the reverse of the estimate found in (1), which finishes the proof. \square

Regarding now the series of functions, again following Lebesgue, we have:

THEOREM 5.36. *Given measurable functions $f_n : X \rightarrow [0, \infty]$, their sum*

$$f(x) = \sum_{n=1}^{\infty} f_n(x)$$

is measurable, and we have the formula

$$\int_X f(x) d\mu(x) = \sum_{n=1}^{\infty} \int_X f_n(x) d\mu(x)$$

for any positive measure on X .

PROOF. This follows indeed from Theorem 5.35, applied to the partial sums. \square

Following now Fatou, we have as well the following key result:

THEOREM 5.37 (Fatou). *Given measurable functions $f_n : X \rightarrow [0, \infty]$, their limit*

$$f(x) = \liminf_n f_n(x)$$

is measurable, and we have the formula

$$\int_X f(x) d\mu(x) \leq \liminf_n \int_X f_n(x) d\mu(x)$$

for any positive measure on X .

PROOF. The first assertion is something that we know, from Proposition 5.30. As for the second assertion, this can be proved by using the same trick as there, namely:

$$\liminf_n f_n = \sup_k \left(\inf_{n \geq k} f_n \right)$$

Indeed, we can apply Theorem 5.35 to the following sequence of functions:

$$g_k(x) = \inf_{n \geq k} f_n(x)$$

Thus, we are led to the conclusion in the statement. \square

Regarding the above result of Fatou, let us mention that there are examples where the inequality can be strict, with the standard example here being as follows:

$$f_n = \begin{cases} \chi_E & (n \text{ odd}) \\ 1 - \chi_E & (n \text{ even}) \end{cases}$$

Finally, as a last basic result, due to Lebesgue again, we have:

THEOREM 5.38 (Lebesgue). *Assuming $f_n \rightarrow f$ pointwise, and assuming too*

$$|f_n| \leq g$$

with $\int g < \infty$, the following happen:

- (1) $\int |f| < \infty$.
- (2) $\int |f_n - f| \rightarrow 0$.
- (3) $\int f_n \rightarrow \int f$.

PROOF. This is something very standard, using Fatou, the idea being as follows:

- (1) This follows from $|f_n| \leq g$, which in the limit gives $|f| < g$, so $\int |f| < \infty$.
- (2) Since $|f_n - f| \leq 2g$, we can apply Theorem 5.37 to the following functions:

$$h_n = 2g - |f_n - f|$$

We obtain in this way the following estimate:

$$\begin{aligned} \int_X 2g &\leq \liminf_n \int_X 2g - |f_n - f| \\ &= \int_X 2g + \liminf_n \left(- \int_X |f_n - f| \right) \\ &= \int_X 2g - \limsup_n \int_X |f_n - f| \end{aligned}$$

Now by subtracting $\int_X 2g$, this estimate gives the following formula:

$$\limsup_n \int_X |f_n - f| \leq 0$$

We conclude that this limit must be 0, as claimed in (2).

(3) This follows indeed from (2). □

Many other things can be said, along these lines, which are more specialized, and such knowledge can be very useful when doing applied probability, and statistics. For our purposes here, which will be mostly pure mathematical, with a touch of theoretical physics, the above general theory, and results of Lebesgue and Fatou, will do.

5e. Exercises

This was a very standard chapter, all fundamentals of advanced analysis, and as exercises on all this, numerous, and all mandatory, we have:

EXERCISE 5.39. *Master Cauchy-Schwarz, and all its generalizations.*

EXERCISE 5.40. *Find the coordinates of the regular N -simplex, inside \mathbb{R}^{N-1} .*

EXERCISE 5.41. *Learn more about connectedness, and connected components.*

EXERCISE 5.42. *Learn about topological spaces, generalizing the metric spaces.*

EXERCISE 5.43. *Learn also about measure theory, on such topological spaces.*

EXERCISE 5.44. *Have some fun with Borel and non-Borel sets, on \mathbb{R} .*

EXERCISE 5.45. *Check all the details for our Riemann integration method.*

EXERCISE 5.46. *Look up and solve exercises, using Lebesgue and Fatou.*

As bonus exercise, as already suggested, have a look into what probabilists and statisticians say too, about Lebesgue, Fatou and related topics. All this is good learning.

CHAPTER 6

Main theorems

6a. Preliminaries

Welcome to measure theory, again. What we did in the previous chapter, while certainly useful, was not very connected to life and reality, because even for simple measured spaces like $X = \mathbb{R}$, we still don't know what the measure μ exactly is.

So, time to discuss such questions, following Riesz, Lebesgue and others. All this will be quite tricky and technical, and as a guiding principle, we will have:

PRINCIPLE 6.1. *The following are the same thing:*

- (1) *Measure theory.*
- (2) *Integration theory.*
- (3) *Probability theory.*

Now back to our question, constructing measures μ on spaces like $X = \mathbb{R}$, or more generally $X = \mathbb{R}^N$, and perhaps on some other metric spaces X too, this does not look obvious to do, with bare hands. However, and here comes the point, in view of Principle 6.1, it is normally enough to know how to integrate, or equivalently, to know how to do probability, on our spaces $X = \mathbb{R}$, or more generally $X = \mathbb{R}^N$, and so on.

And here, we certainly know from calculus how to integrate over $X = \mathbb{R}$, and that is excellent knowledge, that we should normally be able to covert into systematic measure theory on \mathbb{R} , in particular with μ fully constructed. As for $X = \mathbb{R}^N$, and more complicated spaces, such as various curves, surfaces and other manifolds $X \subset \mathbb{R}^N$, here we certainly have some integration knowledge too, from multivariable calculus, so with a bit of luck, we should be able to improve that knowledge, and convert it into measure theory too.

But probably too much talking, let us make a concrete plan, based on this:

PLAN 6.2. *In order to fully develop measure theory:*

- (1) *Following Riesz, we will first need to fully understand, abstractly, the precise relation between measure and integration.*
- (2) *Then, following Lebesgue, we will develop measure theory on $X = \mathbb{R}$, and more generally on $X = \mathbb{R}^N$, based on our integration knowledge there.*

And afterwards, we can have a look at more complicated spaces $X \subset \mathbb{R}^N$ too.

Getting to work now, what can be the “Riesz theorem”, answering (1) above? Ask Riesz, of course, or intercept a cat, and ask the cat, or simply meditate on all this, so many techniques are available to us. In practice, however, simplest is to have a look at a good measure theory book, such as Rudin [76], which reveals the following answer:

FACT 6.3 (Riesz theorem). *Any positive functional $I : C_c(X) \rightarrow \mathbb{R}$ comes by integrating with respect to a certain measure on X ,*

$$I(f) = \int_X f(x) d\mu(x)$$

which is unique, modulo null sets. In addition, the following happen:

- (1) $\mu(K) < \infty$, for any $K \subset X$ compact.
- (2) $\mu(E) = \inf\{\mu(U) | E \subset U \text{ open}\}$, for any E measurable.
- (3) $\mu(E) = \sup\{\mu(K) | K \subset E \text{ compact}\}$, for any E open, or of finite measure.

Which looks a bit scary, but quite reasonable. Indeed, the main assertion, with $C_c(X)$ standing as usual for the continuous, compactly supported functions on X , is more or less the equivalence between (1) and (2) in our guiding Principle 6.1. As for the various properties (1,2,3), which are part of the whole story too, because they sort of tell us how to construct the measure μ , from the knowledge of the integration I , and more on this in a moment, these are certainly very reasonable properties, that we can expect to hold.

So, very good all this, we have a concrete theorem now, waiting to be proved, and the rest will naturally come afterwards, according to Plan 6.2. Now by looking more in detail at what our theorem exactly says, we are led to the following battle plan:

STRATEGY 6.4. *In order to prove the Riesz theorem, we must:*

- (1) *Have a quick discussion about null sets.*
- (2) *Have some squeezing technology for the measurable sets, $K \subset E \subset U$.*
- (3) *And then get to work, and prove the theorem.*

In short, and bad news here, not yet time to get to work and do (3), we have to deal with (1) first, and then especially with (2), which suspiciously reminds what mathematicians call “Lemmas”, and with these having the reputation of being the worst of mathematics, unless of course you are a pure mathematician yourself.

Nevermind. As I always say to my students, one day you will have to work, and work can be quite varied, sometimes you have to move big furniture around the office, install or debug computers, fix electricity or even plumbing problems, wait for someone at the airport, calm down a colleague who gets crazy, encourage another one who starts crying, and so on. So, doing what is to be done, and we will not be scared about some Lemmas, if they are on the menu for today, we will prove Lemmas, no problem with that.

Getting started now, according to our Strategy 6.4, we first need to talk about null sets. So, let us first review the material from the previous chapter, by paying particular attention to the null sets, meaning sets of zero measure. We first have here:

THEOREM 6.5. *Any measured space can be completed,*

$$(X, M, \mu) \rightarrow (X, \bar{M}, \mu)$$

by saturating everything with subsets of null sets.

PROOF. This is something quite self-explanatory, the idea being as follows:

(1) To start with, once we have a null set, $\mu(A) = 0$, we can declare any subset $B \subset A$ to be measurable too, and of course of mass zero, $\mu(B) = 0$.

(2) But then, we can add the sets B found above to all the measurable sets $E \in M$, and we obtain in this way a bigger algebra, $M \subset \bar{M}$.

(3) Finally, our original measure μ extends from the original algebra M to the bigger algebra \bar{M} in the obvious way, according to the formula $\mu(B) = 0$ from (1).

(4) Thus, result basically proved, modulo a number of technical details, which look quite straightforward, and whose clarification we will leave, as an instructive exercise. \square

Along the same lines, but at the level of functions now, we have:

THEOREM 6.6. *At the level of functions on a measured space, we can talk about integration by avoiding null sets, based on the equivalence*

$$\int_X f(x) d\mu(x) = 0 \iff f = 0 \text{ a.e.}$$

for any positive function $f \geq 0$, and then on $f = g - h$ with $g, h \geq 0$, in general.

PROOF. This is indeed something very standard, coming as a functional analytic complement to Theorem 6.5, the idea with all this being as follows:

(1) To start with, we have indeed the equivalence in the statement, for any $f \geq 0$, with this being something standard, coming from the general theory from chapter 5.

(2) Now based on this, we can develop some theory, by neglecting the null sets, which in practice means reformulating Theorem 6.5, in functional analysis terms.

(3) To be more precise, given (X, M, μ) , we certainly know from chapter 5 how to integrate the measurable functions $f : X \rightarrow \mathbb{C}$. And the point is that, by using the equivalence in the statement, this tells us in fact how to integrate the measurable functions $f : X \rightarrow \mathbb{C}$, with respect to the saturated space (X, \bar{M}, μ) from Theorem 6.5.

(4) So, this was for the idea, and we will leave clarifying the details here, both precise statement, and few things that are needed for the proof, as an instructive exercise. \square

As before with Fatou and Lebesgue, many other things can be said, along these lines, which are more specialized, and such knowledge can be very useful when doing applied probability, and statistics. For our purposes here, the above general theory will do.

We will be actually back to this in Part III of this book, when doing more systematically functional analysis, which is the good framework for such things. The idea there will be indeed that of defining various spaces of functions by neglecting the null sets.

Moving now towards what we want to do, namely measuring and integrating over \mathbb{R} , in full generality, the idea here will be that of squeezing the arbitrary measurable sets $E \subset X$ between compact sets $K \subset X$ and open sets $U \subset X$, as follows:

$$K \subset E \subset U$$

In order to do this, which is something quite tricky, we will need a number of technical preliminaries, as a continuation of the material from chapter 5. First, we have:

LEMMA 6.7. *Given $K \subset U$, compact inside open, we can find inclusions*

$$K \subset V \subset \bar{V} \subset U$$

with the set V being open, with compact closure \bar{V} .

PROOF. This is something elementary, the idea being as follows:

(1) For any $x \in K$ pick a neighborhood W_x having compact closure. Since K is compact, we can find finitely many such neighborhoods, covering it:

$$K \subset W_{x_1} \cup \dots \cup W_{x_n}$$

With this done, consider now the set on the right, namely:

$$W = W_{x_1} \cup \dots \cup W_{x_n}$$

This is then an open set containing our compact set K , having compact closure, that we will use in what follows, in our constructions.

(2) In order to prove now our result, observe first that if we are in the case $U = X$ we are done, because here we can simply take $V = W$. So, assume $U \neq X$.

(3) Given a point $x \in U^c$, since this point is compact, and our set K not containing it is compact as well, we can find an open set V_x having the following properties:

$$K \subset V_x \quad , \quad x \notin \bar{V}_x$$

Now consider the following family of sets, with W being the open set in (1):

$$\left\{ K \cap \bar{W} \cap \bar{V}_x \mid x \in U^c \right\}$$

This is then a family of compact sets having empty intersection, so we can find a finite subfamily having empty intersection too. That is, we can find points $x_i \in U^c$ such that:

$$K \cap \bar{W} \cap \bar{V}_{x_1} \cap \dots \cap \bar{V}_{x_m} = \emptyset$$

(4) With this done, consider now the following open set:

$$V = W \cap V_{x_1} \cap \dots \cap V_{x_m}$$

Then V has compact closure, and we have the following inclusion:

$$\bar{V} \subset \bar{W} \cap \bar{V}_{x_1} \cap \dots \cap \bar{V}_{x_m}$$

Thus, V is the open set we are looking for, and we are done. \square

The above result, separating $K \subset U$, compact set inside open set, via an open set V with compact closure \bar{V} , is something quite useful, in practice. However, for our measure theory purposes here, we will need more, along the same lines.

To be more precise, under the same assumptions, $K \subset U$, compact inside open, we would like to come up with a continuous, compactly supported function f such that:

$$\chi_K \leq f \leq \chi_U$$

Which is something quite non-trivial, when thinking a bit, and so again, and with our apologies here, we are in need of some more technical preliminaries.

To be more precise, we will need the following standard notions:

DEFINITION 6.8. *A function $f : X \rightarrow \mathbb{R}$, with X topological space, is called:*

- (1) *Upper semicontinuous, if $\{x \in X | f(x) < a\}$ is open, for any $a \in \mathbb{R}$.*
- (2) *Lower semicontinuous, if $\{x \in X | f(x) > a\}$ is open, for any $a \in \mathbb{R}$.*

These notions are actually quite interesting on their own, having many uses in many contexts in analysis, and as basic illustrations for them, we have:

(1) A characteristic function χ_E is upper semicontinuous when $E \subset X$ is closed. This follows indeed from the above definition of the upper semicontinuity.

(2) Also, a characteristic function χ_E is lower semicontinuous when $E \subset X$ is open. Again, this follows from definitions, or simply from (1), by using the set E^c .

(3) Observe also that a continuous function is trivially both upper and lower semicontinuous. The converse of this holds too, and more on this in a moment.

Many things can be said about the upper and lower semicontinuous functions, which some knowledge here being something quite useful, when doing in analysis in general. In what follows we will only need a handful of basic results on the subject, as follows:

PROPOSITION 6.9. *The upper and lower semicontinuous functions $f : X \rightarrow \mathbb{R}$ have the following properties:*

- (1) *If $F \subset X$ is closed, χ_F is upper semicontinuous.*
- (2) *If $U \subset X$ is open, χ_U is lower semicontinuous.*
- (3) *Infimum of upper semicontinuous functions is upper semicontinuous.*
- (4) *Supremum of lower semicontinuous functions is lower semicontinuous.*
- (5) *f is continuous precisely when it is upper and lower semicontinuous.*

PROOF. All this is elementary, the idea being as follows:

- (1) This assertion, already mentioned in the above, follows from definitions.
- (2) This assertion, also mentioned in the above, also follows from definitions.
- (3) This assertion is also trivial, also coming from definitions.
- (4) And this assertion is trivial too, also coming from definitions.
- (5) In one sense, this is clear, as already mentioned in the above. In the other sense, assuming that $f : X \rightarrow \mathbb{R}$ is both upper and lower semicontinuous, we can see that the preimage of any open interval $(a, b) \subset \mathbb{R}$ is open, due to the following formula:

$$f^{-1}(a, b) = f^{-1}(a, \infty] \cap f^{-1}[-\infty, b)$$

Now since any open set $U \subset \mathbb{R}$ can be written a union of open intervals $(a, b) \subset \mathbb{R}$, it follows that $f^{-1}(U)$ is open, and so that f is continuous, as desired. \square

Many other things can be said, about the upper and lower semicontinuous functions, notably with many examples. However, for our purposes here, the above will do.

Good news, we can now formulate the main technical result that we will need, for our measure theory purposes here. With the convention that the support of a function f is the closure of the set $\{x | f(x) \neq 0\}$, this key statement is as follows:

LEMMA 6.10 (Urysohn). *Given $K \subset U$, compact inside open, we can find*

$$\chi_K \leq f \leq \chi_U$$

with f being continuous, and compactly supported.

PROOF. This is something very standard, the idea being as follows:

- (1) Given sets $K \subset U$ as in the statement, by using Lemma 6.7 we can find an open set V_0 , having compact closure \bar{V}_0 , such that:

$$K \subset V_0 \subset \bar{V}_0 \subset U$$

But then, by using again Lemma 6.7, applied this time to the inclusion $K \subset V_0$, we can find a second open set V_1 , having compact closure \bar{V}_1 , such that:

$$K \subset V_1 \subset \bar{V}_1 \subset V_0 \subset \bar{V}_0 \subset U$$

And so on, would be the idea. In practice now, by using the fact that the rational numbers are countable, we can construct in this way a whole family of open sets V_r , having compact closures \bar{V}_r , one for each rational number $r \in [0, 1]$, such that:

$$r < s \implies \bar{V}_s \subset V_r$$

(2) Time now to construct our function f . Let us set, for any $r \in [0, 1]$ rational:

$$f_r(x) = \begin{cases} r & \text{if } x \in V_r \\ 0 & \text{otherwise} \end{cases}$$

Then, we can construct our function f in the following way:

$$f = \sup_r f_r$$

It is clear then that f is lower semicontinuous, that we have $\chi_K \leq f \leq \chi_U$, and also that f is compactly supported, with support included in the compact set \bar{V}_0 . Thus, we are almost there, and it remains to prove that f is upper semicontinuous as well.

(3) For this purpose, let us set as well, for any $r \in [0, 1]$ rational:

$$g_s(x) = \begin{cases} 1 & \text{if } x \in \bar{V}_s \\ 0 & \text{otherwise} \end{cases}$$

Then, we can construct another function g , in the following way:

$$g = \inf_s g_s$$

It is then clear, exactly as in (2), that g is lower semicontinuous, and also that we have $\chi_K \leq g \leq \chi_U$, and that g is compactly supported. Of interest for us is the lower semicontinuity of g , because in order to finish the proof, it is enough to show that:

$$f = g$$

(4) So, let us prove this, $f = g$. By definition of the functions f_r and g_s , we have:

$$f_r < g_s$$

Thus $f \leq g$. Now assume that we have somewhere a strict inequality, $f(x) < g(x)$. Then, we can find two rational numbers r, s in between, as follows:

$$f(x) < r < s < g(x)$$

But the first inequality tells us that $x \notin V_r$, and the last inequality tells us that $x \in \bar{V}_s$, and this contradicts the condition in (1) on the sets V_r , as desired. \square

Very nice all this, but in fact, contrary to what was advertised before, we will actually need in what follows, besides the Urysohn lemma, another technical result, which is something useful too, and actually of independent interest too, as follows:

THEOREM 6.11. *Given a compact set inside a union of open sets*

$$K \subset U_1 \cup \dots \cup U_n$$

we can find an associated partition of unity, that is, a decomposition of type

$$f_1(x) + \dots + f_n(x) = 1 \quad , \quad \forall x \in K$$

with each f_i being continuous, supported on U_i .

PROOF. This follows by using the Urysohn lemma, as follows:

(1) For any $x \in K$, let us pick a neighborhood V_x such that $\bar{V}_x \subset U_i$, for some i . Since K is compact, we can find finitely many points $x_1, \dots, x_m \in K$ such that:

$$K \subset V_{x_1} \cup \dots \cup V_{x_m}$$

Now for any index $i \in \{1, \dots, n\}$ let us consider the following union:

$$K_i = \bigcup_{\bar{V}_{x_i} \subset U_i} \bar{V}_{x_i}$$

We can apply then the Urysohn lemma to the inclusion $K_i \subset U_i$, and we obtain in this way a continuous, compactly supported function g_i , such that:

$$\chi_{K_i} \leq g_i \leq \chi_{U_i}$$

(2) With this done, consider the following sequence of functions:

$$f_1 = g_1$$

$$f_2 = (1 - g_1)g_2$$

$$\vdots$$

$$f_n = (1 - g_1) \dots (1 - g_{n-1})g_n$$

Then each f_i is continuous, supported on U_i , and we have:

$$f_1 + \dots + f_n = 1 - (1 - g_1) \dots (1 - g_n)$$

(3) On the other hand, recall from the construction of K_i above that we have:

$$K \subset K_1 \cup \dots \cup K_n$$

We conclude from this that we have the following formula:

$$f_1 + \dots + f_n = 1$$

Thus, we have our partition of the unity, as desired. □

6b. Riesz theorem

Good news, eventually, the above is all that we need, as set theoretic preliminaries. We can now formulate the key result in abstract measure theory, as follows:

THEOREM 6.12 (Riesz). *Any positive functional $I : C_c(X) \rightarrow \mathbb{R}$ comes by integrating with respect to a certain measure on X ,*

$$I(f) = \int_X f(x) d\mu(x)$$

which has the following additional properties,

- (1) $\mu(K) < \infty$, for any $K \subset X$ compact.
- (2) $\mu(E) = \inf\{\mu(U) | E \subset U \text{ open}\}$, for any E measurable.
- (3) $\mu(E) = \sup\{\mu(K) | K \subset E \text{ compact}\}$, for any E open, or of finite measure.

and which is unique with these properties, modulo the null sets.

PROOF. This is something quite long and tricky, the idea being as follows:

(1) Let us first prove the uniqueness. Assuming that our measure μ produces I , and has the various properties in the statement, it is clear that μ is uniquely determined by its values on the compact sets $K \subset X$. Thus, we must show that given two measures μ_1, μ_2 as in the statement, and a compact set $K \subset X$, we have:

$$\mu_1(K) = \mu_2(K)$$

For this purpose, let us pick $\varepsilon > 0$. By using the properties (1,3) in the statement, for the measure μ_2 , we can find $K \subset U$ open such that:

$$\mu_2(U) < \mu_2(K) + \varepsilon$$

Now by using the Urysohn lemma for the inclusion $K \subset U$, we obtain a certain continuous, compactly supported function f , such that:

$$\chi_K \leq f \leq \chi_U$$

But, with this choice of f , we have the following computation:

$$\begin{aligned} \mu_1(K) &\leq \int_X f(x) d\mu_1(x) \\ &= \int_X f(x) d\mu_2(x) \\ &\leq \mu_2(U) \\ &< \mu_2(K) + \varepsilon \end{aligned}$$

Thus $\mu_1(K) \leq \mu_2(K)$, and by interchanging μ_1, μ_2 we have the reverse inequality as well, so we obtain, as desired, $\mu_1(K) = \mu_2(K)$, proving the uniqueness statement.

(2) With this done, let us get now to the main part, and discuss the construction of the algebra M , that we will choose to be saturated in the sense of Theorem 6.5, and of the measure μ . What we have as data is a positive functional, as follows:

$$I : C_c(X) \rightarrow \mathbb{R}$$

We must first measure the Borel sets $E \subset X$. And here, to start with, in order to measure the open sets $U \subset X$, the formula is straightforward, namely:

$$\mu(U) = \sup \left\{ I(f) \mid f \leq \chi_U \right\}$$

Now observe that with this definition in hand, for the open sets, we have:

$$U_1 \subset U_2 \implies \mu(U_1) \leq \mu(U_2)$$

We deduce that the following happens, for the open sets:

$$\mu(E) = \inf \left\{ \mu(U) \mid E \subset U \text{ open} \right\}$$

But this can serve as the definition of μ , for all the subsets $E \subset X$.

(3) Regarding now the measurable sets, let us first consider the class N of subsets $E \subset X$ which are of finite measure, $\mu(E) < \infty$, and satisfy the following condition:

$$\mu(E) = \sup \left\{ \mu(K) \mid K \subset E \text{ compact} \right\}$$

Then, we can define the class of measurable sets $M \subset P(X)$ as follows:

$$M = \left\{ E \subset X \mid E \cap K \in N, \forall K \subset X \text{ compact} \right\}$$

Summarizing, done with definitions, and it remains now to prove that M is an algebra, and that μ is a measure on it, along with the other things claimed in the statement.

(4) In order to prove that μ is indeed a measure, our first claim is that we have the following inequality, for any two open sets $U_1, U_2 \subset X$:

$$\mu(U_1 \cup U_2) \leq \mu(U_1) + \mu(U_2)$$

In order to prove this claim, choose an arbitrary function $g : X \rightarrow [0, 1]$, supported on $U_1 \cup U_2$. By using Theorem 6.11 we obtain a certain partition of unity, $f_1 + f_2 = 1$ on the support of g , and it follows that we have the following equality:

$$g = f_1 g + f_2 g$$

Now by applying our integration functional I , we obtain from this:

$$\begin{aligned} I(g) &= I(f_1 g + f_2 g) \\ &= I(f_1 g) + I(f_2 g) \\ &\leq \mu(U_1) + \mu(U_2) \end{aligned}$$

Now since this holds for any $g : X \rightarrow [0, 1]$ as above, this proves our claim.

(5) Our second claim is that we have in fact, more generally, the following inequality, valid for any family of subsets E_1, E_2, E_3, \dots of our space X :

$$\mu\left(\bigcup_{i=1}^{\infty} E_i\right) \leq \sum_{i=1}^{\infty} \mu(E_i)$$

Indeed, this inequality trivially holds if $\mu(E_i) = \infty$ for some i , so we can assume $\mu(E_i) < \infty$ for any i . Now pick $\varepsilon > 0$, and choose open sets $U_i \supset E_i$ such that:

$$\mu(U_i) < \mu(E_i) + \frac{\varepsilon}{2^i}$$

Consider also the union of these open sets U_i , chosen as above:

$$U = \bigcup_{i=1}^{\infty} U_i$$

Now pick an arbitrary function $f : X \rightarrow [0, 1]$, supported on this open set U . Since f has compact support, we can find a certain integer $n \in \mathbb{N}$ such that:

$$\text{supp}(f) \subset U_1 \cup \dots \cup U_n$$

Now by using what we found in (4), recursively, we obtain:

$$\begin{aligned} I(f) &\leq \mu(U_1 \cup \dots \cup U_n) \\ &\leq \mu(U_1) + \dots + \mu(U_n) \\ &\leq \sum_{i=1}^n \mu(E_i) + \varepsilon \end{aligned}$$

Since this holds for any $f : X \rightarrow [0, 1]$ as above, we deduce that we have:

$$\mu\left(\bigcup_{i=1}^{\infty} E_i\right) \leq \mu(U) \leq \sum_{i=1}^{\infty} \mu(E_i) + \varepsilon$$

Now since the number $\varepsilon > 0$ was arbitrary, this proves our claim.

(6) Our next claim, which proves in the assertion (1) in the theorem, is that any compact set $K \subset X$ is measurable, and that we have:

$$\mu(K) < \infty$$

In order to prove this claim, consider an arbitrary continuous function $f : X \rightarrow [0, 1]$ satisfying $f(x) = 1$ on K , then pick $\alpha < 1$, and consider the following open set:

$$U_\alpha = \left\{x \in X \mid f(x) > \alpha\right\}$$

We have then $K \subset U_\alpha$, and so $\alpha g \leq f$, whenever a continuous function $g : X \rightarrow [0, 1]$ is supported by U_α . By using this, we obtain the following estimate:

$$\begin{aligned} \mu(K) &\leq \mu(U_\alpha) \\ &= \sup \left\{ I(g) \mid g : X \rightarrow [0, 1], \text{supp}(g) \subset U_\alpha \right\} \\ &\leq \alpha^{-1} I(f) \end{aligned}$$

Now with $\alpha \rightarrow 1$, we obtain from this the following estimate:

$$\mu(K) \leq I(f)$$

Thus $\mu(K) < \infty$, as claimed, and the fact that K is measurable is clear too.

(7) Our next claim, which improves what we found in (6), is that for any compact set $K \subset X$ we have in fact the following estimate:

$$\mu(K) = \inf \left\{ I(f) \mid f : X \rightarrow [0, 1], f(x) = 1 \text{ on } K \right\}$$

In order to prove this, let us go back to the proof of (6). We know from there that for any continuous function $f : X \rightarrow [0, 1]$ satisfying $f(x) = 1$ on K , as above, we have:

$$\mu(K) \leq I(f)$$

Now pick $\varepsilon > 0$, and choose an open set $U \supset K$ satisfying:

$$\mu(U) < \mu(K) + \varepsilon$$

By using Lemma 6.10, we can find a compactly supported function f such that:

$$\chi_K \leq f \leq \chi_U$$

But this gives the following estimate, using the above inequality $\mu(K) \leq I(f)$:

$$\begin{aligned} I(f) &\leq \mu(U) \\ &< \mu(K) + \varepsilon \\ &\leq I(f) + \varepsilon \end{aligned}$$

Now since the number $\varepsilon > 0$ was arbitrary, this proves our claim.

(8) Our claim now is that any open set $U \subset X$ satisfies the following condition, that we used in (3) in order to define the measurable sets:

$$\mu(U) = \sup \left\{ \mu(K) \mid K \subset U \text{ compact} \right\}$$

In order to prove this, pick an arbitrary number $\alpha < \mu(U)$. We can then find a continuous function $f : X \rightarrow [0, 1]$ supported on U , such that:

$$I(f) > \alpha$$

Consider now the compact set $K = \text{supp}(f)$. Then for any open set V containing K we have $I(f) \leq \mu(V)$, and we deduce from this that we have:

$$I(f) \leq \mu(K)$$

In other words, given $\alpha < \mu(U)$, we have found a compact set K such that:

$$\mu(K) \geq \alpha$$

Now since our number $\alpha < \mu(U)$ was arbitrary, this proves our claim.

(9) Getting now towards the proof of the additivity of our measure μ , let us first prove that for any two disjoint compact sets $K_1, K_2 \subset X$, we have:

$$\mu(K_1 \cup K_2) = \mu(K_1) + \mu(K_2)$$

For this purpose, pick an arbitrary $\varepsilon > 0$. By using Lemma 6.10, we can find a compactly supported continuous function $f : X \rightarrow [0, 1]$ satisfying:

$$f(x) = \begin{cases} 1 & \text{on } K_1 \\ 0 & \text{on } K_2 \end{cases}$$

On the other hand, by using (7) we can find a compactly supported continuous function $g : X \rightarrow [0, 1]$ satisfying $g(x) = 1$ on $K_1 \cup K_2$, such that:

$$I(g) \leq \mu(K_1 \cup K_2) + \varepsilon$$

By using this, and the linearity of I , we have the following estimate:

$$\begin{aligned} \mu(K_1) + \mu(K_2) &\leq I(fg) + I(g - fg) \\ &= I(g) \\ &< \mu(K_1 \cup K_2) + \varepsilon \end{aligned}$$

Now since our number $\varepsilon > 0$ was arbitrary, this proves our claim, via (5).

(10) We are now in position of making a key verification, that of the additivity property of our measure μ . To be more precise, our claim is that we have the following equality, valid for any family of pairwise disjoint subsets E_1, E_2, E_3, \dots of our space X :

$$\mu\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} \mu(E_i)$$

In order to prove this, consider the union on the left, namely:

$$E = \bigcup_{i=1}^{\infty} E_i$$

In the case $\mu(E) = \infty$ our claim is proved by (5). So, assume $\mu(E) < \infty$, and pick an arbitrary number $\varepsilon > 0$. We can then find compact sets $K_i \subset E_i$ such that:

$$\mu(K_i) > \mu(E_i) - \frac{\varepsilon}{2^i}$$

Now consider the following unions, which are compact sets too:

$$H_n = K_1 \cup \dots \cup K_n$$

By using the additivity formula that we found in (9), recursively, we obtain:

$$\mu(E) \geq \mu(H_n) = \sum_{i=1}^n \mu(K_i) > \sum_{i=1}^n \mu(E_i) - \varepsilon$$

Now since our number $\varepsilon > 0$ was arbitrary, this proves our claim.

(11) Thus, additivity property proved, and with this in hand, the other assertions are quite standard. In order to prove these, our first claim is that for any measurable set E and any $\varepsilon > 0$, there is a compact set K , and an open set U , satisfying:

$$K \subset E \subset U \quad , \quad \mu(U - K) < \varepsilon$$

Indeed, in order to prove this claim, the first remark is that we can find indeed a compact set K and an open set U , satisfying the following conditions:

$$K \subset E \subset U \quad , \quad \mu(U) - \frac{\varepsilon}{2} < \mu(E) < \mu(K) + \frac{\varepsilon}{2}$$

Since the set $U - K$ is open, by using (8) and then (10) we obtain:

$$\mu(U) = \mu(K) + \mu(U - K)$$

But with this, we can establish our inequality, as follows:

$$\mu(U - K) = \mu(U) - \mu(K) < \varepsilon$$

(12) Our next claim is that if A, B are measurable, then so are the following sets:

$$A - B \quad , \quad A \cup B \quad , \quad A \cap B$$

In order to prove this, pick an arbitrary number $\varepsilon > 0$. By using (11) we can find certain compact sets K, H , and certain open sets U, V , satisfying:

$$K \subset A \subset U \quad , \quad \mu(U - K) < \varepsilon$$

$$H \subset B \subset V \quad , \quad \mu(V - H) < \varepsilon$$

Now observe that we have the following inclusion:

$$A - B \subset U - H \subset (U - K) \cup (K - V) \cup (V - H)$$

By using now (5), we obtain from this the following estimate:

$$\begin{aligned}\mu(A - B) &\leq \mu(U - K) + \mu(K - V) + \mu(V - H) \\ &< \varepsilon + \mu(K - V) + \varepsilon \\ &= \mu(K - V) + 2\varepsilon\end{aligned}$$

Now since $K - V$ is a compact subset of $A - B$, we conclude that $A - B$ is measurable, as desired. Regarding now $A \cup B$, we can use here the following formula:

$$A \cup B = (A - B) \cup B$$

Indeed, by using (10) we obtain from this that $A \cup B$ is measurable as well. Finally, regarding $A \cap B$, we can use here the following formula:

$$A \cap B = A - (A - B)$$

Thus $A \cap B$ is measurable as well, and this finishes the proof of our claim.

(13) Our claim now is that the set M constructed in (3) is indeed an algebra, which contains all Borel sets. In order to prove this, consider an arbitrary compact set $K \subset X$. Given $A \in M$, we can write $A^c \cap K$ as a difference of two sets in M , as follows:

$$A^c \cap K = K - (A \cap K)$$

Thus $A^c \cap K \in M$, and by using this, we conclude that we have:

$$A \in M \implies A^c \in M$$

Next, let us look at unions. Assume $A_i \in M$, and consider their union:

$$A = \bigcup_{i=1}^{\infty} A_i$$

With $K \subset X$ being an arbitrary compact set, as above, set $B_1 = A_1 \cap K$, and then define recursively the following sets:

$$B_n = (A_n \cap K) - (B_1 \cup \dots \cup B_{n-1})$$

In terms of these sets B_n , we have the following formula:

$$A \cap K = \bigcup_{i=1}^{\infty} B_n$$

On the other hand, according to our definition of the sets B_n , and by using (12), we have $B_n \in M$. Thus, by using (10), we obtain $A \in M$. Thus, we have proved that:

$$A_i \in M \implies \bigcup_{i=1}^{\infty} A_i \in M$$

Finally, if a subset $C \subset X$ is closed, then $C \cap K$ is compact, and so $C \in M$. Thus, as claimed above, M is indeed an algebra, which contains all the Borel sets.

(14) Our next claim, which together with what we found in (10) will prove that μ is indeed a measure on M , is that, in the context of our constructions in (3), the family N constructed there consists precisely of the sets $E \in M$ satisfying:

$$\mu(E) < \infty$$

In order to prove this, let $E \in N$. By using (7) and (12) we deduce that we have $E \cap K \in N$ for any compact set $K \subset X$, and we conclude from this that we have:

$$E \in N \implies E \in M$$

Conversely, assume that $E \in M$ has finite measure, $\mu(E) < \infty$. In order to prove $E \in N$, we pick a number $\varepsilon > 0$. We can then find an open set $U \subset X$ such that:

$$E \subset U \quad , \quad \mu(U) < \infty$$

Also, by (8) and (11) we can find a compact set $K \subset X$ such that:

$$K \subset U \quad , \quad \mu(U - K) < \varepsilon$$

Now since $E \cap K \in N$, we can find a compact set $H \subset X$ such that:

$$H \subset E \cap K \quad , \quad \mu(E \cap K) < \mu(H) + \varepsilon$$

Observe now that we have the following inclusion:

$$E \subset (E \cap K) \cup (U - K)$$

By using this, we obtain the following inequality:

$$\begin{aligned} \mu(E) &\leq \mu((E \cap K) \cup (U - K)) \\ &\leq \mu(E \cap K) + \mu(U - K) \\ &< \mu(H) + 2\varepsilon \end{aligned}$$

Thus $E \in N$, which finishes the proof of our claim, and proves that μ is a measure.

(15) Time now to get into the final verification, that of the following formula:

$$I(f) = \int_X f(x) d\mu(x)$$

As a first observation, by taking real and imaginary parts, it is enough to prove this for the real functions f . Moreover, and assuming now that f is real, by using $f \rightarrow -f$, in order to prove the above equality, it is enough to prove the following inequality:

$$I(f) \leq \int_X f(x) d\mu(x)$$

(16) So, let us prove now this latter inequality. For this purpose, let $K \subset X$ be the support of our function f , choose an interval containing the range, $Im(f) \subset [a, b]$, pick an arbitrary $\varepsilon > 0$, and then numbers y_0, \dots, y_n such that $y_i - y_{i-1} < \varepsilon$, and:

$$y_0 < a < y_1 < \dots < y_n = b$$

With this done, consider now the following sets:

$$E_i = \left\{ x \in X \mid y_{i-1} < f(x) < y_i \right\} \cap K$$

Since f is continuous, these are disjoint Borel sets, whose union is K . Thus, we can find open sets $U_i \subset X$, subject to the following conditions:

$$E_i \subset U_i \quad , \quad \mu(U_i) < \mu(E_i) + \frac{\varepsilon}{n} \quad , \quad f(x) < y_i + \varepsilon, \forall x \in U_i$$

By using now Theorem 6.11, we can find certain continuous functions $h_i : X \rightarrow [0, 1]$, with h_i supported on U_i , such that the following equality holds, on K :

$$h_1(x) + \dots + h_n(x) = 1$$

By multiplying by f , we conclude that the following equality holds, on K :

$$f(x) = h_1(x)f(x) + \dots + h_n(x)f(x)$$

On the other hand, by using (7), we have the following estimate:

$$\mu(K) \leq I \left(\sum_{i=1}^n h_i \right) = \sum_{i=1}^n I(h_i)$$

Good news, with all these ingredients in hand, we can now finish. Since we have $h_i f \leq (y_i + \varepsilon)h_i$, and since $y_i - \varepsilon < f(x)$ on E , we have the following estimate:

$$\begin{aligned} I(f) &= \sum_{i=1}^n I(h_i f) \\ &\leq \sum_{i=1}^n (y_i + \varepsilon) I(h_i) \\ &= \sum_{i=1}^n (|a| + y_i + \varepsilon) I(h_i) - |a| \sum_{i=1}^n I(h_i) \\ &\leq \sum_{i=1}^n (|a| + y_i + \varepsilon) \left(\mu(E_i) + \frac{\varepsilon}{n} \right) - |a| \mu(K) \\ &= \sum_{i=1}^n (y_i - \varepsilon) \mu(E_i) + 2\varepsilon \mu(K) + \frac{\varepsilon}{n} \sum_{i=1}^n (|a| + y_i + \varepsilon) \\ &\leq \int_X f(x) d\mu(x) + \varepsilon (2\mu(K) + |a| + b + \varepsilon) \end{aligned}$$

Now since $\varepsilon > 0$ was arbitrary, this finishes the proof of our latest claim, that is, of the inequality in (15), and so finishes the proof of the present theorem. \square

6c. Regular measures

We proved the Riesz theorem, and that was good learning. In practice now, in order to measure the various measurable sets $E \subset X$, the tools that we have come from (2) and (3) in that theorem. So, let us have a closer look at that two properties.

Let us start with a definition, inspired from what we have:

DEFINITION 6.13. *Let μ be a Borel measure on a space X .*

- (1) *$E \subset X$ is called inner regular if $\mu(E) = \sup\{\mu(K) | K \subset E \text{ compact}\}$.*
- (2) *$E \subset X$ is called outer regular if $\mu(E) = \inf\{\mu(U) | E \subset U \text{ open}\}$.*

If all the Borel sets $E \subset X$ are inner and outer regular, we say that μ is regular.

We know from Theorem 6.12 that, under the assumptions there, all sets $E \subset X$ are outer regular. However, in what regards inner regularity, we only have that for the sets $E \subset X$ which are open, or of finite measure. So, there is certainly room for improvement here, and in relation with this, we first have the following result:

THEOREM 6.14. *Assume, in the context of the Riesz theorem, that X is locally compact, and is σ -compact too, in the sense that it is a countable union of compact sets.*

- (1) *Given a measurable set $E \subset X$ and a number $\varepsilon > 0$, we can find $F \subset E \subset U$ with F closed, U open, and $\mu(U - F) < \varepsilon$.*
- (2) *Also given $E \subset X$ measurable, we can find $A \subset E \subset B$, with A countable union of closed sets, B countable intersection of open sets, and $\mu(B - A) = 0$.*
- (3) *The measure μ is regular, in the above sense.*

PROOF. This follows indeed by further building on Theorem 6.12 and its proof, under the assumptions in the statement, the idea being as follows:

(1) Let us write $X = \cup_i K_i$, with each K_i being compact. Given a measurable set $E \subset X$ and a number $\varepsilon > 0$, we can find then open sets $U_i \supset (K_i \cap E)$ such that:

$$\mu(U_i - (K_i \cap E)) < \frac{\varepsilon}{2^i}$$

Now if we set $U = \cup_i U_i$, we have $U - E \subset \cup_i (U_i - (K_i \cap E))$, which gives:

$$\mu(U - E) < \frac{\varepsilon}{2}$$

We can apply this to $E^c = X - E$ too, and we get an open set $V \supset E^c$ such that:

$$\mu(V - E^c) < \frac{\varepsilon}{2}$$

But this gives the result, with U being as above, and with $F = V^c$.

(2) This follows indeed from (1), in the obvious way, applied with $\varepsilon_i = 1/i$.

(3) This follows again from (1), which proves the needed inner regularity of E . □

Along the same lines, we have as well the following result:

THEOREM 6.15. *Assume that X is locally compact, with all its open sets $U \subset X$ being σ -compact. If λ is a Borel measure on X having the property that*

$$\lambda(K) < \infty$$

for any compact subset $K \subset X$, then λ is regular.

PROOF. Consider indeed the integration with respect to the measure λ :

$$I(f) = \int_X f(x) d\lambda(x)$$

By using our assumption, $\lambda(K) < \infty$ for any $K \subset X$ compact, we conclude that the Riesz theorem applies to this functional. Thus, we obtain a certain measure μ , that we know in addition to be regular, as shown in Theorem 6.14, such that:

$$I(f) = \int_X f(x) d\mu(x)$$

Thus, it remains to prove that we have $\lambda = \mu$, and this can be done as follows:

(1) Let us first prove that we have $\lambda(U) = \mu(U)$, for any open set $U \subset X$. For this purpose, we can use the σ -compactness assumption in the statement, which allows us to write our set U as follows, with the sets $K_i \subset X$ being compact:

$$U = \bigcup_i K_i$$

By using the Urysohn lemma, we can find continuous functions $f_i : X \rightarrow [0, 1]$ which are supported on U , and are equal to 1 on K_i . With this done, let us set:

$$g_n = \max(f_1, \dots, f_n)$$

The monotone convergence theorem applies then, and gives, as desired:

$$\begin{aligned} \lambda(U) &= \lim_{n \rightarrow \infty} \int_X g_n(x) d\lambda(x) \\ &= \lim_{n \rightarrow \infty} \int_X g_n(x) d\mu(x) \\ &= \mu(U) \end{aligned}$$

(2) Now let us prove that we have $\lambda(E) = \mu(E)$, for any Borel set $E \subset X$. For this purpose, pick $\varepsilon > 0$. By using Theorem 6.14 (1), applied to the above measure μ , coming from Riesz, we can find $F \subset E \subset U$, with F closed and U open, such that:

$$\mu(U - F) < \varepsilon$$

Now by using (1), applied to the open set $U - F$, we obtain from this $|\lambda(E) - \mu(E)| < \varepsilon$. But this being true for any $\varepsilon > 0$, we obtain $\lambda(E) = \mu(E)$, as desired. \square

6d. Lebesgue measure

Good work that we did, in the above, and time now to enjoy what we learned. As a first consequence of the Riesz theorem, we can now formulate, following Lebesgue:

THEOREM 6.16 (Lebesgue). *We can measure all the Borel sets*

$$E \subset \mathbb{R}$$

and the corresponding integration theory extends the usual Riemann integral.

PROOF. This follows indeed from Theorem 6.12, by using the Riemann integral, that we know well from basic calculus, as the positive functional needed there:

$$f \rightarrow \int_{\mathbb{R}} f(x) dx$$

Indeed, it follows from basic calculus that the Riemann integral satisfies indeed all the needed properties in Theorem 6.12, and this gives the result. We will leave the various verifications here, which are all elementary, as an instructive exercise. \square

Going ahead with more theory, let us discuss now the case of \mathbb{R}^N . This is something quite tricky, and you would probably say here, why not simply generalizing Theorem 6.16 and its proof, by using this time the multivariable Riemann integral, that we know as well from basic calculus, as the positive functional needed there:

$$f \rightarrow \int_{\mathbb{R}^N} f(z) dz$$

However, let me ask you here, for instance at $N = 2$, do we really master well the above integral? For instance, what is its precise definition? And here, I bet that your answer would be that we can define the multivariable integral by iterating, as follows:

$$\int_{\mathbb{R}^2} f(z) dz = \int_{\mathbb{R}} \int_{\mathbb{R}} f(x, y) dx dy$$

Which looks fine, at a first glance, but there is in fact a bug, with this. Indeed, assuming so, we would have by symmetry the following formula too:

$$\int_{\mathbb{R}^2} f(z) dz = \int_{\mathbb{R}} \int_{\mathbb{R}} f(x, y) dy dx$$

Thus, and forgetting now about what we wanted to do, we can see that our method is based on the Fubini formula, stating that we must have:

$$\int_{\mathbb{R}} \int_{\mathbb{R}} f(x, y) dx dy = \int_{\mathbb{R}} \int_{\mathbb{R}} f(x, y) dy dx$$

But, you might already know from calculus that Fubini does not always work, with the counterexamples being not very difficult to construct, as follows:

PROPOSITION 6.17. *The Fubini formula, namely*

$$\int_{\mathbb{R}} \int_{\mathbb{R}} f(x, y) dx dy = \int_{\mathbb{R}} \int_{\mathbb{R}} f(x, y) dy dx$$

can fail, for certain suitably chosen functions.

PROOF. We have indeed the following computation:

$$\begin{aligned} \int_0^1 \int_0^1 \frac{y^2 - x^2}{(x^2 + y^2)^2} dx dy &= \int_0^1 \left[\frac{x}{x^2 + y^2} \right]_0^1 dy \\ &= \int_0^1 \frac{1}{1 + y^2} dy \\ &= \frac{\pi}{4} \end{aligned}$$

On the other hand, by using this, and symmetry, we have as well:

$$\begin{aligned} \int_0^1 \int_0^1 \frac{y^2 - x^2}{(x^2 + y^2)^2} dy dx &= \int_0^1 \int_0^1 \frac{x^2 - y^2}{(x^2 + y^2)^2} dx dy \\ &= - \int_0^1 \int_0^1 \frac{y^2 - x^2}{(x^2 + y^2)^2} dx dy \\ &= -\frac{\pi}{4} \end{aligned}$$

Thus Fubini fails. Please pass this knowledge to your children, and grandchildren. \square

Summarizing, our method does not work. Which can be regarded as good news for us mathematicians, time now to get into some action, show our skills, and fix all this. Let us first formulate a general objective, inspired from the above, as follows:

OBJECTIVE 6.18. *We would like to measure all the Borel sets*

$$E \subset \mathbb{R}^N$$

with the corresponding integration theory extending and clarifying the Riemann integral, in several variables. Also, we would like to clarify the validity of the Fubini formula

$$\int_{\mathbb{R}} \int_{\mathbb{R}} f(x, y) dx dy = \int_{\mathbb{R}} \int_{\mathbb{R}} f(x, y) dy dx$$

in this formulation, in 2 variables, and in general $N \geq 2$ variables too.

Getting to work now, it is pretty much obvious that, in order to deal with this problem, rigorously, it is better to forget everything that we know from multivariable calculus, regarding the integration in several variables, and recreate everything from scratch, with our tools being Theorem 6.12, complemented by Theorem 6.14 and Theorem 6.15.

So, here is our first result, using our abstract measure theory techniques:

PROPOSITION 6.19. *Given a compactly supported function $f : \mathbb{R}^N \rightarrow \mathbb{C}$, if we set*

$$I_k(f) = \frac{1}{2^{Nk}} \sum_{x \in \mathbb{Z}^N / 2^k} f(x)$$

with $\mathbb{Z}^N / 2^k \subset \mathbb{R}^N$ being the points having as coordinates integer multiples of $1/2^k$, then

$$I(f) = \lim_{k \rightarrow \infty} I_k(f)$$

converges, and $f \rightarrow I(f)$ satisfies the assumptions of the Riesz theorem.

PROOF. In order to prove the first assertion, we can assume that our function is real, $f : \mathbb{R}^N \rightarrow \mathbb{R}$. But then, we can use the fact that f is uniformly continuous, and this gives the convergence in the statement. As for the second assertion, the fact that $f \rightarrow I(f)$ satisfies indeed the assumptions of the Riesz theorem is clear from definitions. \square

In view of the above, we can indeed apply the Riesz theorem. We obtain:

THEOREM 6.20 (Lebesgue). *There is a unique regular, translation-invariant measure on \mathbb{R}^N , normalized as for the volume of the unit cube to be 1. This measure is called Lebesgue measure, and appears alternatively as the measure on the Borel sets*

$$E \subset \mathbb{R}^N$$

whose corresponding integration theory extends the usual Riemann integral, as constructed in Proposition 6.19, or equivalently, as in multivariable calculus.

PROOF. Obviously, many things going on here, and as a first remark, this fulfills indeed our requirements from Objective 6.18, save for Fubini, that we will discuss later, in the next chapter. In what regards the rest, the idea is as follows:

(1) We want to prove that the Lebesgue measure exists, and is unique, with the various properties from the statement. But here, a bit of thinking shows that this measure can only be the one coming from Proposition 6.19, via the Riesz theorem.

(2) So, let us go back to Proposition 6.19, which allows us to apply the Riesz theorem, and apply indeed this Riesz theorem. We obtain in this way a certain regular Borel measure on \mathbb{R}^N , that we have to prove to have the properties in the statement.

(3) But here, all the assertions follow from what we have in the Riesz theorem, complemented where needed by Theorem 6.14 and Theorem 6.15, that we can use too. We will leave the various verifications here as an instructive exercise. \square

Summarizing, we know now how to integrate on \mathbb{R}^N . As a continuation of this, as a first task, let us go back to the Fubini formula problematics. We have:

THEOREM 6.21 (Fubini, Tonelli). *Given a function $f : \mathbb{R}^N \rightarrow \mathbb{R}$ which is measurable and integrable, in the sense that the following integral is finite,*

$$\int_{\mathbb{R}^N} |f(z)| dz < \infty$$

we have the following equalities, for any decomposition $N = N_1 + N_2$:

$$\int_{\mathbb{R}^{N_1}} \int_{\mathbb{R}^{N_2}} f(x, y) dy dx = \int_{\mathbb{R}^{N_2}} \int_{\mathbb{R}^{N_1}} f(x, y) dx dy = \int_{\mathbb{R}^N} f(z) dz$$

Moreover, the same holds when $f : \mathbb{R}^N \rightarrow \mathbb{R}$ is assumed positive, and measurable.

PROOF. In this statement, which is something quite compact, the first assertion is called Fubini theorem, and the second assertion is called Tonelli theorem. These two theorems are related, and can be also subject to various minor improvements. In what follows we will mostly focus on proving the Fubini theorem in its basic form, as stated.

(1) In order to prove the Fubini theorem, consider the vector space L consisting of the functions $f : \mathbb{R}^N \rightarrow \mathbb{R}$ which are measurable, and integrable:

$$L = \left\{ f : \mathbb{R}^N \rightarrow \mathbb{R} \mid \int_{\mathbb{R}^N} |f(z)| dz < \infty \right\}$$

Now fix a decomposition $N = N_1 + N_2$, and consider the family $F \subset L$ of functions $f : \mathbb{R}^N \rightarrow \mathbb{R}$ which satisfy the formula in the statement, namely:

$$\int_{\mathbb{R}^{N_1}} \int_{\mathbb{R}^{N_2}} f(x, y) dy dx = \int_{\mathbb{R}^{N_2}} \int_{\mathbb{R}^{N_1}} f(x, y) dx dy = \int_{\mathbb{R}^N} f(z) dz$$

With this conventions, we want to prove that the inclusion $F \subset L$ is an equality.

(2) But this comes from the following sequence of observations, which all come from definitions, and from the Lebesgue dominated convergence theorem from chapter 5:

- F is stable under taking finite linear combinations.
- F is stable under taking pointwise limits of increasing sequences.
- $1_E \in F$, for any set of finite measure of type $E = E_1 \times E_2 \subset \mathbb{R}^N$.
- $1_E \in F$, for any open set of finite measure $E \subset \mathbb{R}^N$.
- $1_E \in F$, for any countable intersection of open sets, of finite measure $E \subset \mathbb{R}^N$.
- $1_E \in F$, for any set of null measure $E \subset \mathbb{R}^N$.
- $1_E \in F$, for any set of finite measure $E \subset \mathbb{R}^N$.

Indeed, the proof of all these assertions, in the above precise order, is something quite straightforward, and once we have the last assertion, the result follows, because we can approximate any given function $f \in L$ by step functions, and we obtain $f \in F$.

(3) Summarizing, Fubini proved, modulo some straightforward verifications left to you, and the same goes with Tonelli, whose proof is quite routine, based on Fubini. \square

As a conclusion, we know how to rigorously integrate on \mathbb{R}^N , and as a bonus, the theory that we developed seems to have many follow-ups. More on this later.

6e. Exercises

Tough chapter that we had here, with many things explained quite quickly, and as exercises on this, all in relation with better understanding what we did, we have:

EXERCISE 6.22. *Clarify what has been said above about saturating with null sets.*

EXERCISE 6.23. *Clarify as well what we said about integration, modulo null sets.*

EXERCISE 6.24. *Learn more about almost everywhere equality, from probabilists.*

EXERCISE 6.25. *Meditate about Fubini, its victories, and its failures.*

EXERCISE 6.26. *Fill in the details for the properties of the Lebesgue measure.*

EXERCISE 6.27. *Learn as well other methods for constructing the Lebesgue measure.*

EXERCISE 6.28. *Learn about the Lusin theorem, and its applications.*

EXERCISE 6.29. *Learn also about the Vitali-Carathéodory theorem.*

As bonus exercise, read some systematic measure theory, say from Rudin [76], with what we did here being an introduction to that, and from an alternative book too.

CHAPTER 7

Integration theory

7a. Multiple integrals

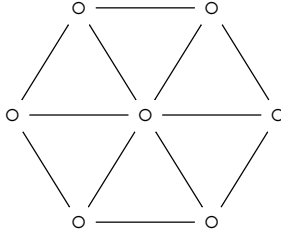
In this chapter and in the next one we discuss more advanced aspects of integration, on various explicit spaces $X \subset \mathbb{R}^N$, notably in relation with derivatives. To start with, we will be interested in the explicit computation of the integrals in \mathbb{R}^N .

Let us start with some basics. In what regards the area of the circle, sure you know that, but never too late to learn the truth about it, which is as follows:

THEOREM 7.1. *The following two definitions of π are equivalent:*

- (1) *The length of the unit circle is $L = 2\pi$.*
- (2) *The area of the unit disk is $A = \pi$.*

PROOF. In order to prove this theorem let us cut the unit disk as a pizza, into N slices, and forgetting about gastronomy, leave aside the rounded parts:



The area to be eaten can be then computed as follows, where H is the height of the slices, S is the length of their sides, and $P = NS$ is the total length of the sides:

$$A = N \times \frac{HS}{2} = \frac{HP}{2} \simeq \frac{1 \times L}{2}$$

Thus, with $N \rightarrow \infty$ we obtain that we have $A = L/2$, as desired. □

As a second computation, the area of an ellipse can be computed as follows:

THEOREM 7.2. *The area of an ellipse, given by the equation*

$$\left(\frac{x}{a}\right)^2 + \left(\frac{y}{b}\right)^2 = 1$$

with $a, b > 0$ being half the size of a box containing the ellipse, is $A = \pi ab$.

PROOF. The idea is that of cutting the ellipse into vertical slices. First observe that, according to our equation $(x/a)^2 + (y/b)^2 = 1$, the x coordinate can range as follows:

$$x \in [-a, a]$$

For any such x , the other coordinate y , satisfying $(x/a)^2 + (y/b)^2 = 1$, is given by:

$$y = \pm b \sqrt{1 - \frac{x^2}{a^2}}$$

We conclude that the area of the ellipse is given by the following formula:

$$\begin{aligned} A &= 2b \int_{-a}^a \sqrt{1 - \frac{x^2}{a^2}} dx \\ &= 4ab \int_0^1 \sqrt{1 - y^2} dy \\ &= \pi ab \end{aligned}$$

Finally, as a verification, for $a = b = R$ we get $A = \pi R^2$, as we should. \square

In what regards now the length of the ellipse, the “pizza” argument from the proof of Theorem 7.1 does not work, and things here get fairly complicated, as follows:

FACT 7.3. *The length of an ellipse, given by $(x/a)^2 + (y/b)^2 = 1$, is*

$$L = 4 \int_0^{\pi/2} \sqrt{a^2 \sin^2 t + b^2 \cos^2 t} dt$$

and with this integral being generically not computable.

To be more precise, the above formula can be deduced in the following way, and more on such things in chapter 8 below, when systematically discussing the curves:

$$\begin{aligned} L &= 4 \int_0^{\pi/2} \sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2} dt \\ &= 4 \int_0^{\pi/2} \sqrt{\left(\frac{da \cos t}{dt}\right)^2 + \left(\frac{db \sin t}{dt}\right)^2} dt \\ &= 4 \int_0^{\pi/2} \sqrt{(-a \sin t)^2 + (b \cos t)^2} dt \\ &= 4 \int_0^{\pi/2} \sqrt{a^2 \sin^2 t + b^2 \cos^2 t} dt \end{aligned}$$

As for the last assertion, when $a = b = R$ we get of course $L = 2\pi R$, as we should, but in general, when $a \neq b$, there is no trick for computing the above integral.

Moving now to 3D, as an obvious challenge here, we can try to compute the area and volume of the sphere, and more generally of the ellipsoids. We have here:

THEOREM 7.4. *The volume of the unit sphere in \mathbb{R}^3 is given by:*

$$V = \frac{4\pi}{3}$$

More generally, the volume of an ellipsoid, given by $(x/a)^2 + (y/b)^2 + (z/c)^2 = 1$, is:

$$V = \frac{4\pi abc}{3}$$

The area of the sphere is $A = 4\pi$. For ellipsoids, the area is generically not computable.

PROOF. There are several things going on here, as follows:

(1) Let us first compute the volume of the ellipsoid, which at $a = b = c = 1$ will give the volume of the unit sphere. The range of the first coordinate x is as follows:

$$x \in [-a, a]$$

Now when the first coordinate x is fixed, the other coordinates y, z vary on an ellipse, given by the equation $(y/b)^2 + (z/c)^2 = 1 - (x/a)^2$, which can be written as follows:

$$\left(\frac{y}{\beta}\right)^2 + \left(\frac{z}{\gamma}\right)^2 = 1 \quad : \quad \beta = b\sqrt{1 - \left(\frac{x}{a}\right)^2}, \quad \gamma = c\sqrt{1 - \left(\frac{x}{a}\right)^2}$$

Thus, the vertical slice of our ellipsoid at x has area as follows:

$$A(x) = \pi\beta\gamma = \pi bc \left[1 - \left(\frac{x}{a}\right)^2\right]$$

We conclude that the volume of the ellipsoid is given, as claimed, by:

$$\begin{aligned} V &= \pi bc \int_{-a}^a \left[1 - \left(\frac{x}{a}\right)^2\right] dx \\ &= \pi bc \left[x - \frac{x^3}{3a^2} \right]_{-a}^a \\ &= \pi bc \left(\frac{2a}{3} + \frac{2a}{3} \right) \\ &= \frac{4\pi abc}{3} \end{aligned}$$

(2) At $a = b = c = 1$ we get $V = 4\pi/3$, and this gives the area of the unit sphere too, because the “pizza” method from the proof of Theorem 7.1 applies, and gives:

$$A = 3 \times V = 3 \times \frac{4\pi}{3} = 4\pi$$

(3) Finally, the last assertion is something quite informal, inspired from Fact 7.3. \square

In order to deal now with N dimensions, we will need the Wallis formula:

THEOREM 7.5 (Wallis). *We have the following formulae,*

$$\int_0^{\pi/2} \cos^n t \, dt = \int_0^{\pi/2} \sin^n t \, dt = \left(\frac{\pi}{2}\right)^{\varepsilon(n)} \frac{n!!}{(n+1)!!}$$

where $\varepsilon(n) = 1$ if n is even, and $\varepsilon(n) = 0$ if n is odd, and where

$$m!! = (m-1)(m-3)(m-5)\dots$$

with the product ending at 2 if m is odd, and ending at 1 if m is even.

PROOF. Let us first compute the integral on the left in the statement:

$$I_n = \int_0^{\pi/2} \cos^n t \, dt$$

We do this by partial integration. We have the following formula:

$$\begin{aligned} (\cos^n t \sin t)' &= n \cos^{n-1} t (-\sin t) \sin t + \cos^n t \cos t \\ &= n \cos^{n+1} t - n \cos^{n-1} t + \cos^{n+1} t \\ &= (n+1) \cos^{n+1} t - n \cos^{n-1} t \end{aligned}$$

By integrating between 0 and $\pi/2$, we obtain the following formula:

$$(n+1)I_{n+1} = nI_{n-1}$$

Thus we can compute I_n by recurrence, and we obtain:

$$\begin{aligned} I_n &= \frac{n-1}{n} I_{n-2} \\ &= \frac{n-1}{n} \cdot \frac{n-3}{n-2} I_{n-4} \\ &= \frac{n-1}{n} \cdot \frac{n-3}{n-2} \cdot \frac{n-5}{n-4} I_{n-6} \\ &\vdots \\ &= \frac{n!!}{(n+1)!!} I_{1-\varepsilon(n)} \end{aligned}$$

But $I_0 = \frac{\pi}{2}$ and $I_1 = 1$, so we get the result. As for the second formula, this follows from the first one, with $t = \frac{\pi}{2} - s$. Thus, we have proved both formulae in the statement. \square

We can now compute the volumes of the N dimensional spheres, as follows:

THEOREM 7.6. *The volume of the unit sphere in \mathbb{R}^N is given by*

$$V = \left(\frac{\pi}{2}\right)^{[N/2]} \frac{2^N}{(N+1)!!}$$

with our usual convention $N!! = (N-1)(N-3)(N-5)\dots$

PROOF. If we denote by V_N the volume of the unit sphere in \mathbb{R}^N , we have:

$$\begin{aligned}
 V_N &= \int_{-1}^1 (1-x^2)^{(N-1)/2} dx \cdot V_{N-1} \\
 &= 2V_{N-1} \int_0^1 (1-x^2)^{(N-1)/2} dx \\
 &= 2V_{N-1} \int_0^{\pi/2} (1-\sin^2 t)^{(N-1)/2} \cos t dt \\
 &= 2V_{N-1} \int_0^{\pi/2} \cos^{N-1} t \cos t dt \\
 &= 2V_{N-1} \int_0^{\pi/2} \cos^N t dt
 \end{aligned}$$

Now by recurrence, and by using the formula in Theorem 7.5, we obtain:

$$\begin{aligned}
 V_N &= 2^N \int_0^{\pi/2} \cos^N t dt \int_0^{\pi/2} \cos^{N-1} t dt \dots \int_0^{\pi/2} \cos t dt \\
 &= 2^N \left(\frac{\pi}{2}\right)^{\varepsilon(N)+\varepsilon(N-1)+\dots+\varepsilon(1)} \frac{N!!}{(N+1)!!} \cdot \frac{(N-1)!!}{N!!} \dots \frac{1!!}{2!!} \\
 &= \left(\frac{\pi}{2}\right)^{\varepsilon(N)+\varepsilon(N-1)+\dots+\varepsilon(1)} \frac{2^N}{(N+1)!!} \\
 &= \left(\frac{\pi}{2}\right)^{[N/2]} \frac{2^N}{(N+1)!!}
 \end{aligned}$$

Thus, we are led to the formula in the statement. \square

As main particular cases of the above formula, we have:

PROPOSITION 7.7. *The volumes of the low-dimensional spheres are as follows:*

- (1) At $N = 1$, the length of the unit interval is $V = 2$.
- (2) At $N = 2$, the area of the unit disk is $V = \pi$.
- (3) At $N = 3$, the volume of the unit sphere is $V = \frac{4\pi}{3}$.
- (4) At $N = 4$, the volume of the corresponding unit sphere is $V = \frac{\pi^2}{2}$.

PROOF. Some of these results are well-known, but we can obtain all of them as particular cases of the general formula in Theorem 7.6, as follows:

- (1) At $N = 1$ we obtain $V = 1 \cdot \frac{2}{1} = 2$.
- (2) At $N = 2$ we obtain $V = \frac{\pi}{2} \cdot \frac{4}{2} = \pi$.
- (3) At $N = 3$ we obtain $V = \frac{\pi}{2} \cdot \frac{8}{3} = \frac{4\pi}{3}$.
- (4) At $N = 4$ we obtain $V = \frac{\pi^2}{4} \cdot \frac{16}{8} = \frac{\pi^2}{2}$. \square

Next, we can compute as well the volumes of general ellipsoids, as follows:

THEOREM 7.8. *The volume of an arbitrary ellipsoid in \mathbb{R}^N , given by*

$$\left(\frac{x_1}{a_1}\right)^2 + \dots + \left(\frac{x_N}{a_N}\right)^2 = 1$$

is $V = a_1 \dots a_N V_0$, with V_0 being the volume of the unit sphere in \mathbb{R}^N .

PROOF. We already know this at $N = 2, 3$, from Theorem 7.2 and Theorem 7.4, and the proof in general is similar, by suitably adapting the proof for the sphere. We will leave the computations here as an exercise, and we will come back to this in a moment, with a more conceptual argument, based on Theorem 7.6, and on a change of variables. \square

Finally, let us record as well the formula of the area of the sphere, as follows:

THEOREM 7.9. *The area of the unit sphere in \mathbb{R}^N is given by:*

$$A = \left(\frac{\pi}{2}\right)^{[N/2]} \frac{2^N}{(N-1)!!}$$

In particular, at $N = 2, 3, 4$ we obtain respectively $A = 2\pi, 4\pi, 2\pi^2$.

PROOF. As shown by the pizza argument from the proof of Theorem 7.1, which extends to arbitrary dimensions, the area and volume of the sphere in \mathbb{R}^N are related by:

$$A = N \cdot V$$

Together with the formula in Theorem 7.6 for V , this gives the result. \square

Needless to say, in what regards the area of the general ellipsoids in \mathbb{R}^N , this is in general not computable, as we already know well at $N = 2, 3$, from the above.

7b. Change of variables

We know from one-variable calculus that the change of variables for integrals is related to differentiation, and we can of course expect the same to happen in several variables. So, let us first recall the basics of differentiation, in several variables. We first have:

THEOREM 7.10. *The functions $f : \mathbb{R}^N \rightarrow \mathbb{R}^M$ can be locally approximated as*

$$f(x+t) \simeq f(x) + f'(x)t$$

with $f'(x)$ being by definition the matrix of partial derivatives at x ,

$$f'(x) = \left(\frac{df_i}{dx_j}(x) \right)_{ij} \in M_{M \times N}(\mathbb{R})$$

acting on the vectors $t \in \mathbb{R}^N$ by usual multiplication.

PROOF. This is obviously something a bit informal, because the precise assumptions on f , for everything to work fine, are not mentioned in the statement. However, you surely know this, and the discussion below is just a matter of recalling the basics:

(1) First of all, at $N = M = 1$ what we have is a usual 1-variable function $f : \mathbb{R} \rightarrow \mathbb{R}$, and the formula in the statement is something that we know well, namely:

$$f(x + t) \simeq f(x) + f'(x)t$$

(2) Let us discuss now the case $N = 2, M = 1$. Here what we have is a function $f : \mathbb{R}^2 \rightarrow \mathbb{R}$, and by using twice the basic approximation result from (1), we obtain:

$$\begin{aligned} f\begin{pmatrix} x_1 + t_1 \\ x_2 + t_2 \end{pmatrix} &\simeq f\begin{pmatrix} x_1 + t_1 \\ x_2 \end{pmatrix} + \frac{df}{dx_2}(x)t_2 \\ &\simeq f\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \frac{df}{dx_1}(x)t_1 + \frac{df}{dx_2}(x)t_2 \\ &= f\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} \frac{df}{dx_1}(x) & \frac{df}{dx_2}(x) \end{pmatrix} \begin{pmatrix} t_1 \\ t_2 \end{pmatrix} \end{aligned}$$

(3) More generally, we can deal in this way with the general case $M = 1$, with the formula here, obtained via a straightforward recurrence, being as follows:

$$\begin{aligned} f\begin{pmatrix} x_1 + t_1 \\ \vdots \\ x_N + t_N \end{pmatrix} &\simeq f\begin{pmatrix} x_1 \\ \vdots \\ x_N \end{pmatrix} + \frac{df}{dx_1}(x)t_1 + \dots + \frac{df}{dx_N}(x)t_N \\ &= f\begin{pmatrix} x_1 \\ \vdots \\ x_N \end{pmatrix} + \begin{pmatrix} \frac{df}{dx_1}(x) & \dots & \frac{df}{dx_N}(x) \end{pmatrix} \begin{pmatrix} t_1 \\ \vdots \\ t_N \end{pmatrix} \end{aligned}$$

(4) But this gives the result in the case where both $N, M \in \mathbb{N}$ are arbitrary too. Indeed, we can apply (3) to each of the components $f_i : \mathbb{R}^N \rightarrow \mathbb{R}$, and we get:

$$f_i\begin{pmatrix} x_1 + t_1 \\ \vdots \\ x_N + t_N \end{pmatrix} \simeq f_i\begin{pmatrix} x_1 \\ \vdots \\ x_N \end{pmatrix} + \begin{pmatrix} \frac{df_i}{dx_1}(x) & \dots & \frac{df_i}{dx_N}(x) \end{pmatrix} \begin{pmatrix} t_1 \\ \vdots \\ t_N \end{pmatrix}$$

But this collection of M formulae tells us precisely that the following happens, as an equality, or rather approximation, of vectors in \mathbb{R}^M :

$$f\begin{pmatrix} x_1 + t_1 \\ \vdots \\ x_N + t_N \end{pmatrix} \simeq f\begin{pmatrix} x_1 \\ \vdots \\ x_N \end{pmatrix} + \begin{pmatrix} \frac{df_1}{dx_1}(x) & \dots & \frac{df_1}{dx_N}(x) \\ \vdots & & \vdots \\ \frac{df_M}{dx_1}(x) & \dots & \frac{df_M}{dx_N}(x) \end{pmatrix} \begin{pmatrix} t_1 \\ \vdots \\ t_N \end{pmatrix}$$

Thus, we are led to the conclusion in the statement. \square

Generally speaking, Theorem 7.10 is all you need to know for upgrading from calculus to multivariable calculus. As a standard result here, we have:

THEOREM 7.11. *We have the chain derivative formula*

$$(f \circ g)'(x) = f'(g(x)) \cdot g'(x)$$

as an equality of matrices.

PROOF. Consider indeed a composition of functions, as follows:

$$f : \mathbb{R}^N \rightarrow \mathbb{R}^M, \quad g : \mathbb{R}^K \rightarrow \mathbb{R}^N, \quad f \circ g : \mathbb{R}^K \rightarrow \mathbb{R}^M$$

According to Theorem 7.10, the derivatives of these functions are certain linear maps, corresponding to certain rectangular matrices, as follows:

$$f'(g(x)) \in M_{M \times N}(\mathbb{R}), \quad g'(x) \in M_{N \times K}(\mathbb{R}), \quad (f \circ g)'(x) \in M_{M \times K}(\mathbb{R})$$

Thus, our formula makes sense indeed. As for proof, this comes from:

$$\begin{aligned} (f \circ g)(x+t) &= f(g(x+t)) \\ &\simeq f(g(x) + g'(x)t) \\ &\simeq f(g(x)) + f'(g(x))g'(x)t \end{aligned}$$

Thus, we are led to the conclusion in the statement. □

Getting now to integration matters, let us first recall that we have:

PROPOSITION 7.12. *We have the change of variable formula*

$$\int_a^b f(x)dx = \int_c^d f(\varphi(t))\varphi'(t)dt$$

where $c = \varphi^{-1}(a)$ and $d = \varphi^{-1}(b)$.

PROOF. This follows with $f = F'$, from the following differentiation rule:

$$(F\varphi)'(t) = F'(\varphi(t))\varphi'(t)$$

Indeed, by integrating between c and d , we obtain the result. □

In several variables now, things are quite similar, the result being as follows:

THEOREM 7.13. *Given a transformation $\varphi = (\varphi_1, \dots, \varphi_N)$, we have*

$$\int_E f(x)dx = \int_{\varphi^{-1}(E)} f(\varphi(t))|J_\varphi(t)|dt$$

with the J_φ quantity, called *Jacobian*, being given by

$$J_\varphi(t) = \det \left[\left(\frac{d\varphi_i}{dx_j}(x) \right)_{ij} \right]$$

and with this generalizing the usual formula from one variable calculus.

PROOF. This is something quite tricky, the idea being as follows:

(1) Observe first that this generalizes indeed the change of variable formula in 1 dimension, from Proposition 7.12, the point here being that the absolute value on the derivative appears as to compensate for the lack of explicit bounds for the integral.

(2) As a second observation, we can assume if we want, by linearity, that we are dealing with the constant function $f = 1$. For this function, our formula reads:

$$\text{vol}(E) = \int_{\varphi^{-1}(E)} |J_\varphi(t)| dt$$

In terms of $D = \varphi^{-1}(E)$, this amounts in proving that we have:

$$\text{vol}(\varphi(D)) = \int_D |J_\varphi(t)| dt$$

Now since this latter formula is additive with respect to D , it is enough to prove it for small cubes D . And here, as a first remark, our formula is clear for the linear maps φ , by using the definition of the determinant of real matrices, as a signed volume.

(3) However, the extension of this to the case of non-linear maps φ is something which looks non-trivial, so we will not follow this path, in what follows. So, while the above $f = 1$ discussion is certainly something nice, our theorem is still in need of a proof.

(4) In order to prove the theorem, as stated, let us rather focus on the transformations used φ , instead of the functions to be integrated f . Our first claim is that the validity of the theorem is stable under taking compositions of such transformations φ .

(5) In order to prove this claim, consider a composition, as follows:

$$\varphi : E \rightarrow F \quad , \quad \psi : D \rightarrow E \quad , \quad \varphi \circ \psi : D \rightarrow F$$

Assuming that the theorem holds for φ, ψ , we have the following computation:

$$\begin{aligned} \int_F f(x) dx &= \int_E f(\varphi(s)) |J_\varphi(s)| ds \\ &= \int_D f(\varphi \circ \psi(t)) |J_\varphi(\psi(t))| \cdot |J_\psi(t)| dt \\ &= \int_D f(\varphi \circ \psi(t)) |J_{\varphi \circ \psi}(t)| dt \end{aligned}$$

Thus, our theorem holds as well for $\varphi \circ \psi$, and we have proved our claim.

(6) Next, as a key ingredient, let us examine the case where we are in $N = 2$ dimensions, and our transformation φ has one of the following special forms:

$$\varphi(x, y) = (\psi(x, y), y) \quad , \quad \varphi(x, y) = (x, \psi(x, y))$$

By symmetry, it is enough to deal with the first case. Here the Jacobian is $d\psi/dx$, and by replacing if needed $\psi \rightarrow -\psi$, we can assume that this Jacobian is positive, $d\psi/dx > 0$. Now by assuming as before that $D = \varphi^{-1}(E)$ is a rectangle, $D = [a, b] \times [c, d]$, we can prove our formula by using the change of variables in 1 dimension, as follows:

$$\begin{aligned} \int_E f(s) ds &= \int_{\varphi(D)} f(x, y) dx dy \\ &= \int_c^d \int_{\psi(a, y)}^{\psi(b, y)} f(x, y) dx dy \\ &= \int_c^d \int_a^b f(\psi(x, y), y) \frac{d\psi}{dx} dx dy \\ &= \int_D f(\varphi(t)) J_\varphi(t) dt \end{aligned}$$

(7) But with this, we can now prove the theorem, in $N = 2$ dimensions. Indeed, given a transformation $\varphi = (\varphi_1, \varphi_2)$, consider the following two transformations:

$$\phi(x, y) = (\varphi_1(x, y), y) \quad , \quad \psi(x, y) = (x, \varphi_2 \circ \phi^{-1}(x, y))$$

We have then $\varphi = \psi \circ \phi$, and by using (6) for ψ, ϕ , which are of the special form there, and then (3) for composing, we conclude that the theorem holds for φ , as desired.

(8) Thus, theorem proved in $N = 2$ dimensions, and the extension of the above proof to arbitrary N dimensions is straightforward, that we will leave this as an exercise. \square

As a basic application of this technology, we can recover Theorem 7.8, as follows:

THEOREM 7.14. *The volume of an arbitrary ellipsoid in \mathbb{R}^N , given by*

$$\left(\frac{x_1}{a_1}\right)^2 + \dots + \left(\frac{x_N}{a_N}\right)^2 = 1$$

is $V = a_1 \dots a_N V_0$, with V_0 being the volume of the unit sphere in \mathbb{R}^N .

PROOF. This is indeed something clear, with the change of variables $x_i = a_i y_i$, whose Jacobian is constant, $J = a_1 \dots a_N$. Thus, we are led to the formula in the statement. \square

In order to discuss now some other applications, in 2 dimensions, let us start with:

THEOREM 7.15. *We have polar coordinates in 2 dimensions,*

$$\begin{cases} x = r \cos t \\ y = r \sin t \end{cases}$$

the corresponding Jacobian being $J = r$.

PROOF. This is something elementary, with the Jacobian being as follows:

$$\begin{aligned}
 J &= \begin{vmatrix} \frac{d(r \cos t)}{dr} & \frac{d(r \cos t)}{dt} \\ \frac{d(r \sin t)}{dr} & \frac{d(r \sin t)}{dt} \end{vmatrix} \\
 &= \begin{vmatrix} \cos t & -r \sin t \\ \sin t & r \cos t \end{vmatrix} \\
 &= r \cos^2 t + r \sin^2 t \\
 &= r
 \end{aligned}$$

Thus, we have indeed the formula in the statement. \square

We can now compute the Gauss integral, which is the best calculus formula ever:

THEOREM 7.16. *We have the following formula,*

$$\int_{\mathbb{R}} e^{-x^2} dx = \sqrt{\pi}$$

called Gauss integral formula.

PROOF. Let I be the above integral. By using polar coordinates, we obtain:

$$\begin{aligned}
 I^2 &= \int_{\mathbb{R}} \int_{\mathbb{R}} e^{-x^2-y^2} dx dy \\
 &= \int_0^{2\pi} \int_0^\infty e^{-r^2} r dr dt \\
 &= 2\pi \int_0^\infty \left(-\frac{e^{-r^2}}{2} \right)' dr \\
 &= 2\pi \left[0 - \left(-\frac{1}{2} \right) \right] \\
 &= \pi
 \end{aligned}$$

Thus, we are led to the formula in the statement. \square

Moving now to 3 dimensions, we have here the following result:

THEOREM 7.17. *We have spherical coordinates in 3 dimensions,*

$$\begin{cases} x = r \cos s \\ y = r \sin s \cos t \\ z = r \sin s \sin t \end{cases}$$

the corresponding Jacobian being $J(r, s, t) = r^2 \sin s$.

PROOF. The fact that we have indeed spherical coordinates is clear. Regarding now the Jacobian, this is given by the following formula:

$$\begin{aligned}
& J(r, s, t) \\
&= \begin{vmatrix} \cos s & -r \sin s & 0 \\ \sin s \cos t & r \cos s \cos t & -r \sin s \sin t \\ \sin s \sin t & r \cos s \sin t & r \sin s \cos t \end{vmatrix} \\
&= r^2 \sin s \sin t \begin{vmatrix} \cos s & -r \sin s \\ \sin s \sin t & r \cos s \sin t \end{vmatrix} + r \sin s \cos t \begin{vmatrix} \cos s & -r \sin s \\ \sin s \cos t & r \cos s \cos t \end{vmatrix} \\
&= r \sin s \sin^2 t \begin{vmatrix} \cos s & -r \sin s \\ \sin s & r \cos s \end{vmatrix} + r \sin s \cos^2 t \begin{vmatrix} \cos s & -r \sin s \\ \sin s & r \cos s \end{vmatrix} \\
&= r \sin s (\sin^2 t + \cos^2 t) \begin{vmatrix} \cos s & -r \sin s \\ \sin s & r \cos s \end{vmatrix} \\
&= r \sin s \times 1 \times r \\
&= r^2 \sin s
\end{aligned}$$

Thus, we have indeed the formula in the statement. \square

Let us work out now the general spherical coordinate formula, in arbitrary N dimensions. The formula here, which generalizes those at $N = 2, 3$, is as follows:

THEOREM 7.18. *We have spherical coordinates in N dimensions,*

$$\begin{cases} x_1 &= r \cos t_1 \\ x_2 &= r \sin t_1 \cos t_2 \\ \vdots & \\ x_{N-1} &= r \sin t_1 \sin t_2 \dots \sin t_{N-2} \cos t_{N-1} \\ x_N &= r \sin t_1 \sin t_2 \dots \sin t_{N-2} \sin t_{N-1} \end{cases}$$

the corresponding Jacobian being given by the following formula,

$$J(r, t) = r^{N-1} \sin^{N-2} t_1 \sin^{N-3} t_2 \dots \sin^2 t_{N-3} \sin t_{N-2}$$

and with this generalizing the known formulae at $N = 2, 3$.

PROOF. As before, the fact that we have spherical coordinates is clear. Regarding now the Jacobian, also as before, by developing over the last column, we have:

$$\begin{aligned}
J_N &= r \sin t_1 \dots \sin t_{N-2} \sin t_{N-1} \times \sin t_{N-1} J_{N-1} \\
&+ r \sin t_1 \dots \sin t_{N-2} \cos t_{N-1} \times \cos t_{N-1} J_{N-1} \\
&= r \sin t_1 \dots \sin t_{N-2} (\sin^2 t_{N-1} + \cos^2 t_{N-1}) J_{N-1} \\
&= r \sin t_1 \dots \sin t_{N-2} J_{N-1}
\end{aligned}$$

Thus, we obtain the formula in the statement, by recurrence. \square

Let us discuss now the computation of the arbitrary integrals over the sphere. We will need a technical result extending Theorem 7.5, as follows:

THEOREM 7.19 (Wallis). *We have the following formula,*

$$\int_0^{\pi/2} \cos^p t \sin^q t \, dt = \left(\frac{\pi}{2}\right)^{\varepsilon(p)\varepsilon(q)} \frac{p!!q!!}{(p+q+1)!!}$$

where $\varepsilon(p) = 1$ if p is even, and $\varepsilon(p) = 0$ if p is odd, and where

$$m!! = (m-1)(m-3)(m-5)\dots$$

with the product ending at 2 if m is odd, and ending at 1 if m is even.

PROOF. We use the same idea as in Theorem 7.5. Let I_{pq} be the integral in the statement. In order to do the partial integration, observe that we have:

$$\begin{aligned} (\cos^p t \sin^q t)' &= p \cos^{p-1} t (-\sin t) \sin^q t \\ &\quad + \cos^p t \cdot q \sin^{q-1} t \cos t \\ &= -p \cos^{p-1} t \sin^{q+1} t + q \cos^{p+1} t \sin^{q-1} t \end{aligned}$$

By integrating between 0 and $\pi/2$, we obtain, for $p, q > 0$:

$$pI_{p-1, q+1} = qI_{p+1, q-1}$$

Thus, we can compute I_{pq} by recurrence. When q is even we have:

$$\begin{aligned} I_{pq} &= \frac{q-1}{p+1} I_{p+2, q-2} \\ &= \frac{q-1}{p+1} \cdot \frac{q-3}{p+3} I_{p+4, q-4} \\ &= \vdots \\ &= \frac{p!!q!!}{(p+q)!!} I_{p+q} \end{aligned}$$

But the last term comes from Theorem 7.5, and we obtain the result:

$$\begin{aligned} I_{pq} &= \frac{p!!q!!}{(p+q)!!} I_{p+q} \\ &= \frac{p!!q!!}{(p+q)!!} \left(\frac{\pi}{2}\right)^{\varepsilon(p+q)} \frac{(p+q)!!}{(p+q+1)!!} \\ &= \left(\frac{\pi}{2}\right)^{\varepsilon(p)\varepsilon(q)} \frac{p!!q!!}{(p+q+1)!!} \end{aligned}$$

Observe that this gives the result for p even as well, by using $I_{pq} = I_{qp}$, coming via $t = \frac{\pi}{2} - s$. In the remaining case now, where both p, q are odd, we can use once again the

formula $pI_{p-1,q+1} = qI_{p+1,q-1}$ established above, and the recurrence goes as follows:

$$\begin{aligned}
 I_{pq} &= \frac{q-1}{p+1} I_{p+2,q-2} \\
 &= \frac{q-1}{p+1} \cdot \frac{q-3}{p+3} I_{p+4,q-4} \\
 &= \vdots \\
 &= \frac{p!!q!!}{(p+q-1)!!} I_{p+q-1,1}
 \end{aligned}$$

In order to compute the last term, observe that we have:

$$\begin{aligned}
 I_{p1} &= \int_0^{\pi/2} \cos^p t \sin t \, dt \\
 &= -\frac{1}{p+1} \int_0^{\pi/2} (\cos^{p+1} t)' \, dt \\
 &= \frac{1}{p+1}
 \end{aligned}$$

Thus, we can finish our computation in the case p, q odd, as follows:

$$\begin{aligned}
 I_{pq} &= \frac{p!!q!!}{(p+q-1)!!} I_{p+q-1,1} \\
 &= \frac{p!!q!!}{(p+q-1)!!} \cdot \frac{1}{p+q} \\
 &= \frac{p!!q!!}{(p+q+1)!!}
 \end{aligned}$$

Thus, we obtain the formula in the statement, the exponent of $\pi/2$ appearing there being $\varepsilon(p)\varepsilon(q) = 0 \cdot 0 = 0$ in the present case, and this finishes the proof. \square

We can now integrate over the spheres, as follows:

THEOREM 7.20. *The polynomial integrals over the unit sphere $S_{\mathbb{R}}^{N-1} \subset \mathbb{R}^N$, with respect to the normalized, mass 1 measure, are given by the following formula,*

$$\int_{S_{\mathbb{R}}^{N-1}} x_1^{k_1} \dots x_N^{k_N} \, dx = \frac{(N-1)!!k_1!! \dots k_N!!}{(N + \sum k_i - 1)!!}$$

valid when all exponents k_i are even. If an exponent is odd, the integral vanishes.

PROOF. If one exponent k_i is odd, with $x_i \rightarrow -x_i$ we can see that the integral in the statement vanishes. So, assume that all k_i are even. As a first observation, the result

holds indeed at $N = 2$, due to the formula from Theorem 7.19, which reads:

$$\int_0^{\pi/2} \cos^p t \sin^q t dt = \left(\frac{\pi}{2}\right)^{\varepsilon(p)\varepsilon(q)} \frac{p!!q!!}{(p+q+1)!!} = \frac{p!!q!!}{(p+q+1)!!}$$

In the general case now, where the dimension $N \in \mathbb{N}$ is arbitrary, the integral in the statement can be written in spherical coordinates, as follows:

$$I = \frac{2^N}{A} \int_0^{\pi/2} \dots \int_0^{\pi/2} x_1^{k_1} \dots x_N^{k_N} J dt_1 \dots dt_{N-1}$$

Here A is the area of the sphere, J is the Jacobian, and the 2^N factor comes from the restriction to the $1/2^N$ part of the sphere where all the coordinates are positive. According to Theorem 7.9, the normalization constant in front of the integral is:

$$\frac{2^N}{A} = \left(\frac{2}{\pi}\right)^{[N/2]} (N-1)!!$$

As for the unnormalized integral, this is given by:

$$\begin{aligned} I' = \int_0^{\pi/2} \dots \int_0^{\pi/2} & (\cos t_1)^{k_1} (\sin t_1 \cos t_2)^{k_2} \\ & \vdots \\ & (\sin t_1 \sin t_2 \dots \sin t_{N-2} \cos t_{N-1})^{k_{N-1}} \\ & (\sin t_1 \sin t_2 \dots \sin t_{N-2} \sin t_{N-1})^{k_N} \\ & \sin^{N-2} t_1 \sin^{N-3} t_2 \dots \sin^2 t_{N-3} \sin t_{N-2} \\ & dt_1 \dots dt_{N-1} \end{aligned}$$

By rearranging the terms, we obtain:

$$\begin{aligned} I' = & \int_0^{\pi/2} \cos^{k_1} t_1 \sin^{k_2+\dots+k_N+N-2} t_1 dt_1 \\ & \int_0^{\pi/2} \cos^{k_2} t_2 \sin^{k_3+\dots+k_N+N-3} t_2 dt_2 \\ & \vdots \\ & \int_0^{\pi/2} \cos^{k_{N-2}} t_{N-2} \sin^{k_{N-1}+k_N+1} t_{N-2} dt_{N-2} \\ & \int_0^{\pi/2} \cos^{k_{N-1}} t_{N-1} \sin^{k_N} t_{N-1} dt_{N-1} \end{aligned}$$

Now by using the above-mentioned formula at $N = 2$, this gives:

$$\begin{aligned}
 I' &= \frac{k_1!!(k_2 + \dots + k_N + N - 2)!!}{(k_1 + \dots + k_N + N - 1)!!} \left(\frac{\pi}{2}\right)^{\varepsilon(N-2)} \\
 &\quad \frac{k_2!!(k_3 + \dots + k_N + N - 3)!!}{(k_2 + \dots + k_N + N - 2)!!} \left(\frac{\pi}{2}\right)^{\varepsilon(N-3)} \\
 &\quad \vdots \\
 &\quad \frac{k_{N-2}!!(k_{N-1} + k_N + 1)!!}{(k_{N-2} + k_{N-1} + l_N + 2)!!} \left(\frac{\pi}{2}\right)^{\varepsilon(1)} \\
 &\quad \frac{k_{N-1}!!k_N!!}{(k_{N-1} + k_N + 1)!!} \left(\frac{\pi}{2}\right)^{\varepsilon(0)}
 \end{aligned}$$

Now let F be the part involving the double factorials, and P be the part involving the powers of $\pi/2$, so that $I' = F \cdot P$. Regarding F , by cancelling terms we have:

$$F = \frac{k_1!! \dots k_N!!}{(\sum k_i + N - 1)!!}$$

As in what regards P , by summing the exponents, we obtain $P = \left(\frac{\pi}{2}\right)^{[N/2]}$. We can now put everything together, and we obtain the formula in the statement. \square

In the case of the complex spheres, we have the following result:

THEOREM 7.21. *We have the following integration formula over the complex sphere $S_{\mathbb{C}}^{N-1} \subset \mathbb{C}^N$, with respect to the normalized uniform measure,*

$$\int_{S_{\mathbb{C}}^{N-1}} |z_1|^{2k_1} \dots |z_N|^{2k_N} dz = \frac{(N-1)!k_1! \dots k_n!}{(N + \sum k_i - 1)!}$$

valid for any exponents $k_i \in \mathbb{N}$. As for the other polynomial integrals in z_1, \dots, z_N and their conjugates $\bar{z}_1, \dots, \bar{z}_N$, these all vanish.

PROOF. The last assertion being clear by using $z \rightarrow \lambda z$, assume that we are in the non-vanishing case. By using the standard identification $S_{\mathbb{C}}^{N-1} \simeq S_{\mathbb{R}}^{2N-1}$, we obtain:

$$\begin{aligned}
 I &= \int_{S_{\mathbb{C}}^{N-1}} |z_1|^{2k_1} \dots |z_N|^{2k_N} dz \\
 &= \int_{S_{\mathbb{R}}^{2N-1}} (x_1^2 + y_1^2)^{k_1} \dots (x_N^2 + y_N^2)^{k_N} d(x, y) \\
 &= \sum_{r_1 \dots r_N} \binom{k_1}{r_1} \dots \binom{k_N}{r_N} \int_{S_{\mathbb{R}}^{2N-1}} x_1^{2k_1-2r_1} y_1^{2r_1} \dots x_N^{2k_N-2r_N} y_N^{2r_N} d(x, y)
 \end{aligned}$$

By using now our integration formula for the real spheres, we obtain:

$$\begin{aligned}
& I \\
&= \sum_{r_1 \dots r_N} \binom{k_1}{r_1} \cdots \binom{k_N}{r_N} \frac{(2N-1)!!(2r_1)!! \cdots (2r_N)!!(2k_1-2r_1)!! \cdots (2k_N-2r_N)!!}{(2N+2\sum k_i-1)!!} \\
&= \sum_{r_1} \binom{2r_1}{r_1} \binom{2k_1-2r_1}{k_1-r_1} \cdots \sum_{r_N} \binom{2r_N}{r_N} \binom{2k_N-2r_N}{k_N-r_N} \frac{(N-1)!k_1! \cdots k_N!}{4^{\sum k_i}(N+\sum k_i-1)!} \\
&= 4^{k_1} \times \cdots \times 4^{k_N} \times \frac{(N-1)!k_1! \cdots k_N!}{4^{\sum k_i}(N+\sum k_i-1)!} \\
&= \frac{(N-1)!k_1! \cdots k_N!}{(N+\sum k_i-1)!}
\end{aligned}$$

Thus, we are led to the formula in the statement. \square

7c. Probability, revised

Good news, with our accumulated measure theory knowledge, we can now lay down some solid foundations for probability theory, both discrete and continuous. With the idea in mind of doing things a bit abstractly, our starting point will be:

DEFINITION 7.22. *Let X be a probability space, that is, a space with a probability measure, and with the corresponding integration denoted E , and called expectation.*

- (1) *The random variables are the real functions $f \in L^\infty(X)$.*
- (2) *The moments of such a variable are the numbers $M_k(f) = E(f^k)$.*
- (3) *The law of such a variable is the measure given by $M_k(f) = \int_{\mathbb{R}} x^k d\mu_f(x)$.*

Here the fact that μ_f exists indeed is well-known. By linearity, we would like to have a real probability measure making hold the following formula, for any $P \in \mathbb{R}[X]$:

$$E(P(f)) = \int_{\mathbb{R}} P(x) d\mu_f(x)$$

By using a standard continuity argument, it is enough to have this formula for the characteristic functions χ_I of the measurable sets of real numbers $I \subset \mathbb{R}$:

$$E(\chi_I(f)) = \int_{\mathbb{R}} \chi_I(x) d\mu_f(x)$$

But this latter formula, which reads $P(f \in I) = \mu_f(I)$, can serve as a definition for μ_f , and we are done. Next in line, we need to talk about independence:

DEFINITION 7.23. *Two variables $f, g \in L^\infty(X)$ are called independent when*

$$E(f^k g^l) = E(f^k) E(g^l)$$

happens, for any $k, l \in \mathbb{N}$.

Again, this definition hides some non-trivial things. Indeed, by linearity, we would like to have a formula as follows, valid for any polynomials $P, Q \in \mathbb{R}[X]$:

$$E[P(f)Q(g)] = E[P(f)] E[Q(g)]$$

By using a continuity argument, it is enough to have this formula for characteristic functions χ_I, χ_J of the measurable sets of real numbers $I, J \subset \mathbb{R}$:

$$E[\chi_I(f)\chi_J(g)] = E[\chi_I(f)] E[\chi_J(g)]$$

Thus, we are led to the usual definition of independence, namely:

$$P(f \in I, g \in J) = P(f \in I) P(g \in J)$$

All this might seem a bit abstract, but in practice, the idea is of course that f, g must be independent, in an intuitive, real-life sense. As a first result now, we have:

PROPOSITION 7.24. *Assuming that $f, g \in L^\infty(X)$ are independent, we have*

$$\mu_{f+g} = \mu_f * \mu_g$$

where $*$ is the convolution of real probability measures.

PROOF. We have the following formula, coming from the independence of f, g :

$$M_k(f+g) = \sum_r \binom{k}{r} M_r(f) M_{k-r}(g)$$

On the other hand, by using the Fubini theorem, we have as well:

$$\begin{aligned} \int_{\mathbb{R}} x^k d(\mu_f * \mu_g)(x) &= \int_{\mathbb{R} \times \mathbb{R}} (x+y)^k d\mu_f(x) d\mu_g(y) \\ &= \sum_r \binom{k}{r} \int_{\mathbb{R}} x^r d\mu_f(x) \int_{\mathbb{R}} y^{k-r} d\mu_g(y) \\ &= \sum_r \binom{k}{r} M_r(f) M_{k-r}(g) \end{aligned}$$

Thus μ_{f+g} and $\mu_f * \mu_g$ have the same moments, so they coincide, as claimed. \square

Here is now a second result on independence, which is something more advanced:

THEOREM 7.25. *Assuming that $f, g \in L^\infty(X)$ are independent, we have*

$$F_{f+g} = F_f F_g$$

where $F_f(x) = E(e^{ixf})$ is the Fourier transform.

PROOF. We have the following computation, using Proposition 7.24 and Fubini:

$$\begin{aligned}
 F_{f+g}(x) &= \int_{\mathbb{R}} e^{ixz} d\mu_{f+g}(z) \\
 &= \int_{\mathbb{R}} e^{ixz} d(\mu_f * \mu_g)(z) \\
 &= \int_{\mathbb{R} \times \mathbb{R}} e^{ix(z+t)} d\mu_f(z) d\mu_g(t) \\
 &= \int_{\mathbb{R}} e^{ixz} d\mu_f(z) \int_{\mathbb{R}} e^{ixt} d\mu_g(t) \\
 &= F_f(x) F_g(x)
 \end{aligned}$$

Thus, we are led to the conclusion in the statement. \square

But with this in hand, and with the Gauss formula helping too, we can now have a new, final look at the normal variables, discussed in chapter 4. Let us start with:

DEFINITION 7.26. *The normal law of parameter 1 is the following measure:*

$$g_1 = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$$

More generally, the normal law of parameter $t > 0$ is the following measure:

$$g_t = \frac{1}{\sqrt{2\pi t}} e^{-x^2/2t} dx$$

These are also called Gaussian distributions, with “g” standing for Gauss.

Observe that these laws have indeed mass 1, as they should, due to:

$$\int_{\mathbb{R}} e^{-x^2/2t} dx = \int_{\mathbb{R}} e^{-y^2} \sqrt{2t} dy = \sqrt{2\pi t}$$

Generally speaking, the normal laws appear as bit everywhere, in real life. The reasons for this come from the Central Limit Theorem (CLT), that we will explain in a moment, after developing some more general theory. As a first result, we have:

THEOREM 7.27. *We have the following formula, valid for any $t > 0$:*

$$F_{g_t}(x) = e^{-tx^2/2}$$

*In particular, the normal laws satisfy $g_s * g_t = g_{s+t}$, for any $s, t > 0$.*

PROOF. The Fourier transform formula can be established as follows:

$$\begin{aligned}
 F_{g_t}(x) &= \frac{1}{\sqrt{2\pi t}} \int_{\mathbb{R}} e^{-z^2/2t+ixz} dz \\
 &= \frac{1}{\sqrt{2\pi t}} \int_{\mathbb{R}} e^{-(z/\sqrt{2t}-\sqrt{t/2}iz)^2-tx^2/2} dz \\
 &= \frac{1}{\sqrt{2\pi t}} \int_{\mathbb{R}} e^{-y^2-tx^2/2} \sqrt{2t} dy \\
 &= \frac{1}{\sqrt{\pi}} e^{-tx^2/2} \int_{\mathbb{R}} e^{-y^2} dy \\
 &= e^{-tx^2/2}
 \end{aligned}$$

As for $g_s * g_t = g_{s+t}$, this follows via Theorem 7.25, $\log F_{g_t}$ being linear in t . \square

We are now ready to state and prove the CLT, as follows:

THEOREM 7.28 (CLT). *Given real variables $f_1, f_2, f_3, \dots \in L^\infty(X)$ which are i.i.d., centered, and with common variance $t > 0$, we have*

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n f_i \sim g_t$$

with $n \rightarrow \infty$, in moments.

PROOF. The Fourier transform of the variable in the statement is:

$$\begin{aligned}
 F(x) &= \left[F_f \left(\frac{x}{\sqrt{n}} \right) \right]^n \\
 &= \left[1 - \frac{tx^2}{2n} + O(n^{-2}) \right]^n \\
 &\simeq e^{-tx^2/2}
 \end{aligned}$$

But this function being the Fourier transform of g_t , we obtain the result. \square

Finally, we will need the following result, regarding the moments:

THEOREM 7.29. *The even moments of the normal law are the numbers*

$$M_k(g_t) = t^{k/2} \times k!!$$

where $k!! = (k-1)(k-3)(k-5)\dots$, and the odd moments vanish.

PROOF. We have the following computation, valid for any integer $k \in \mathbb{N}$:

$$\begin{aligned}
 M_k &= \frac{1}{\sqrt{2\pi t}} \int_{\mathbb{R}} y^k e^{-y^2/2t} dy \\
 &= \frac{1}{\sqrt{2\pi t}} \int_{\mathbb{R}} (ty^{k-1}) \left(-e^{-y^2/2t}\right)' dy \\
 &= \frac{1}{\sqrt{2\pi t}} \int_{\mathbb{R}} t(k-1)y^{k-2} e^{-y^2/2t} dy \\
 &= t(k-1) \times \frac{1}{\sqrt{2\pi t}} \int_{\mathbb{R}} y^{k-2} e^{-y^2/2t} dy \\
 &= t(k-1)M_{k-2}
 \end{aligned}$$

Thus, by recurrence, we are led to the formula in the statement. \square

In relation now with geometry, we first have the following result:

THEOREM 7.30. *The moments of the hyperspherical variables are*

$$\int_{S_{\mathbb{R}}^{N-1}} x_i^p dx = \frac{(N-1)!!p!!}{(N+p-1)!!}$$

and the rescaled variables $y_i = \sqrt{N}x_i$ become normal and independent with $N \rightarrow \infty$.

PROOF. The moment formula in the statement follows from our general integration formula over the real spheres, which was as follows:

$$\int_{S_{\mathbb{R}}^{N-1}} x_1^{k_1} \dots x_N^{k_N} dx = \frac{(N-1)!!k_1!! \dots k_N!!}{(N + \sum k_i - 1)!!}$$

Indeed, as a consequence, with $N \rightarrow \infty$ we have the following estimate:

$$\int_{S_{\mathbb{R}}^{N-1}} x_i^p dx \simeq N^{-p/2} \times p!! = N^{-p/2} M_p(g_1)$$

Thus, the variables $\sqrt{N}x_i$ become normal with $N \rightarrow \infty$, as claimed. As for the proof of the asymptotic independence, this is standard too, once again by using our general integration formula above. Indeed, the joint moments of x_1, \dots, x_N are given by:

$$\begin{aligned}
 \int_{S_{\mathbb{R}}^{N-1}} x_1^{k_1} \dots x_N^{k_N} dx &= \frac{(N-1)!!k_1!! \dots k_N!!}{(N + \sum k_i - 1)!!} \\
 &\simeq N^{-\sum k_i} \times k_1!! \dots k_N!!
 \end{aligned}$$

By rescaling, the joint moments of the variables $y_i = \sqrt{N}x_i$ are given by:

$$\int_{S_{\mathbb{R}}^{N-1}} y_1^{k_1} \dots y_N^{k_N} dx \simeq k_1!! \dots k_N!!$$

Thus, we have multiplicativity, and so independence with $N \rightarrow \infty$, as claimed. \square

In the case of the complex spheres, we have a similar result, as follows:

THEOREM 7.31. *The rescalings $\sqrt{N}z_i$ of the unit complex sphere coordinates*

$$z_i : S_{\mathbb{C}}^{N-1} \rightarrow \mathbb{C}$$

become complex Gaussian and independent with $N \rightarrow \infty$.

PROOF. As before, we have the following estimate, for a single coordinate, which shows that the rescaled variables $\sqrt{N}z_i$ become normal with $N \rightarrow \infty$:

$$\int_{S_{\mathbb{C}}^{N-1}} |z_i|^{2k} dz = \frac{(N-1)!k!}{(N+k-1)!} \simeq N^{-k} \times k!$$

As for the asymptotic independence, this is standard too, again by using our general integration formulae. Indeed, the joint moments of z_1, \dots, z_N are given by:

$$\begin{aligned} \int_{S_{\mathbb{C}}^{N-1}} |z_1|^{2k_1} \dots |z_N|^{2k_N} dx &= \frac{(N-1)!k_1! \dots k_N!}{(N + \sum k_i - 1)!} \\ &\simeq N^{-\sum k_i} \times k_1! \dots k_N! \end{aligned}$$

Thus, we have multiplicativity, and so independence with $N \rightarrow \infty$, as claimed. \square

7d. Radon-Nikodym

How many probability measures are there? In order to discuss this question, we need to compare abstract measures, and the main theorem here is a very useful decomposition result, due to Radon-Nikodym, that we would like to explain now.

Let us start with some abstract notions, in relation with this, as follows:

DEFINITION 7.32. *Let λ, μ be two measures on a measurable space X .*

- (1) *We say that λ is absolutely continuous with respect to μ , and write $\lambda < \mu$, when $\mu(E) = 0 \implies \lambda(E) = 0$, for any measurable set E .*
- (2) *We say that λ, μ are orthogonal, and write $\lambda \perp \mu$, when we can find disjoint sets A, B such that λ is concentrated on A , and μ is concentrated on B .*

As illustrations for this, consider the case where $X = \mathbb{R}^N$, and μ is the Lebesgue measure. In this case it is easy to construct measures $\lambda < \mu$, because any integrable function $f : \mathbb{R}^N \rightarrow \mathbb{R}$ produces such a measure, according to the following formula:

$$\lambda(E) = \int_E f(x) d\mu(x)$$

As for the measures $\lambda \perp \mu$, there are many of them too, because any discrete measure λ , appearing by definition as follows, with $\{x_i\} \subset \mathbb{R}^N$ being a sequence of points, and

$\{c_i\} \subset [0, \infty)$ being a sequence of positive numbers, has this orthogonality property:

$$\lambda = \sum_i c_i \delta_{x_i}$$

Getting back now to the general case, where the measurable space X is arbitrary, we have the following remarkable decomposition result, due to Radon-Nikodym:

THEOREM 7.33. *Consider a measured space (X, μ) , with μ being σ -finite, and let λ be another measure on X . We can then find a decomposition as follows,*

$$\lambda = \nu + \eta \quad , \quad \nu < \mu \quad , \quad \eta \perp \mu$$

and in what regards the absolutely continuous part ν , this appears as

$$\nu(E) = \int_E f(x) d\mu(x)$$

for a certain function f , integrable with respect to μ , called density of ν .

PROOF. This is something quite standard, based on the material that we know, but instead of explaining this, which is quite long, here is a quick proof due to von Neumann, using some functional analysis material that we will learn later in this book:

- (1) Pick a function $w : X \rightarrow (0, 1)$ which is integrable with respect to μ , and set:

$$\rho = \lambda + w\mu$$

The integration with respect to λ being a bounded linear functional on the space of square-integrable functions with respect to ρ , we must have a formula as follows for this functional, for a certain function g , which is square-integrable with respect to ρ :

$$\int_X f(x) d\lambda(x) = \int_X f(x) g(x) d\rho(x)$$

We have then $g : X \rightarrow [0, 1]$, and the above formula can be rewritten as follows:

$$\int_X (1 - g(x)) f(x) d\lambda(x) = \int_X f(x) g(x) w(x) d\mu(x)$$

As already mentioned, we have used here some functional analysis material, that we will learn later in this book. So, don't forget to come back later to review this proof.

- (2) But this gives both the assertions. Indeed, consider first the following sets:

$$A = \left\{ x \in X \mid g(x) < 1 \right\} \quad , \quad B = \left\{ x \in X \mid g(x) = 1 \right\}$$

We can then construct the needed decomposition $\lambda = \nu + \eta$ by setting:

$$\nu(E) = \lambda(E \cap A) \quad , \quad \eta(E) = \lambda(E \cap B)$$

Indeed, we have $\lambda = \nu + \eta$, and the conditions $\nu < \mu$ and $\eta \perp \mu$ are both clear.

(3) Moreover, with $f = 1 + g + \dots + g^{n-1}$ in the formula from (1), we obtain:

$$\int_X (1 - g^n(x)) d\lambda(x) = \int_X (g(x) + g^2(x) + \dots + g^n(x)) w(x) d\mu(x)$$

But with this, we can set the density function f to be the limit of the integrands on the right, and the density function property is satisfied indeed, as desired. \square

Getting back now to \mathbb{R}^N , the Radon-Nikodym theorem has the following consequence:

THEOREM 7.34. *Any probability measure on \mathbb{R}^N is of the following form,*

$$d\lambda(x) = f(x)dx + d\eta(x)$$

with $f : \mathbb{R}^N \rightarrow \mathbb{R}_+$ being a certain density, and η being supported on a measure 0 set.

PROOF. This follows indeed from Theorem 7.33, when applied to the case of the probability measures, that is, of the positive measures of mass 1. Observe that, as basic examples, we have the measures of the following type, with $f \geq 0$ and $c_i \geq 0$:

$$\lambda = f(x)dx + \sum_i c_i \delta_{x_i}$$

To be more precise, here $f : \mathbb{R}^N \rightarrow \mathbb{R}$ is a certain density function, $x_i \in \mathbb{R}^N$ are certain points, and $c_i \in \mathbb{R}$ are certain numbers, subject to the following conditions:

$$f \geq 0 \quad , \quad c_i \geq 0 \quad , \quad \int_{\mathbb{R}^N} f(x)dx + \sum_i c_i = 1$$

However, these are just the basic examples, and this because exotic measures $\eta \perp \mu$ do exist. For more on this, we refer to any specialized measure theory book. \square

7e. Exercises

We had a quite exciting chapter here, and as exercises on all this, we have:

EXERCISE 7.35. *Meditate some more on the various definitions of π .*

EXERCISE 7.36. *Do some computations for the lengths of the ellipses.*

EXERCISE 7.37. *Do some computations for the areas of ellipsoids, too.*

EXERCISE 7.38. *Learn if needed the theory of partial derivatives, with full details.*

EXERCISE 7.39. *Fill in the details, in the proof of the change of the variable formula.*

EXERCISE 7.40. *Learn about the stereographic projection, and cartography in general.*

EXERCISE 7.41. *Compute the laws of hyperspherical variables at small values of N .*

EXERCISE 7.42. *Learn, with full details, the Radon-Nikodym theorem.*

As bonus exercise, find and solve 100 difficult exercises on multivariable calculus.

CHAPTER 8

Contour integrals

8a. Curves, surfaces

Our goal in this chapter is to integrate over curves, surfaces, and other manifolds. But, as a question coming before that, what are the manifolds themselves?

In answer, things can be quite complicated, but everything comes from conics. We have indeed the following result, which is at the origin of modern mathematics:

THEOREM 8.1. *The conics, which are the algebraic curves of degree 2 in the plane,*

$$C = \left\{ (x, y) \in \mathbb{R}^2 \mid P(x, y) = 0 \right\}$$

with $\deg P \leq 2$, appear modulo degeneration by cutting a 2-sided cone with a plane, and can be classified into ellipses, parabolas and hyperbolas.

PROOF. This is something very standard, the idea being as follows:

(1) Let us first classify the conics up to non-degenerate linear transformations of the plane, which are by definition transformations as follows, with $\det A \neq 0$:

$$\begin{pmatrix} x \\ y \end{pmatrix} \rightarrow A \begin{pmatrix} x \\ y \end{pmatrix}$$

Our claim is that as solutions we have the circles, parabolas, hyperbolas, along with some degenerate solutions, namely \emptyset , points, lines, pairs of lines, \mathbb{R}^2 .

(2) As a first remark, it looks like we forgot precisely the ellipses, but via linear transformations these become circles, so things fine. As a second remark, all our claimed solutions can appear. Indeed, the circles, parabolas, hyperbolas can appear as follows:

$$x^2 + y^2 = 1 \quad , \quad x^2 = y \quad , \quad xy = 1$$

As for \emptyset , points, lines, pairs of lines, \mathbb{R}^2 , these can appear too, as follows, and with our polynomial P chosen, whenever possible, to be of degree exactly 2:

$$x^2 = -1 \quad , \quad x^2 + y^2 = 0 \quad , \quad x^2 = 0 \quad , \quad xy = 0 \quad , \quad 0 = 0$$

Observe here that, when dealing with these degenerate cases, assuming $\deg P = 2$ instead of $\deg P \leq 2$ would only rule out \mathbb{R}^2 itself, which is not worth it.

(3) Getting now to the proof of our claim in (1), classification up to linear transformations, consider an arbitrary conic, written as follows, with $a, b, c, d, e, f \in \mathbb{R}$:

$$ax^2 + by^2 + cxy + dx + ey + f = 0$$

Assume first $a \neq 0$. By making a square out of ax^2 , up to a linear transformation in (x, y) , we can get rid of the term cxy , and we are left with:

$$ax^2 + by^2 + dx + ey + f = 0$$

In the case $b \neq 0$ we can make two obvious squares, and again up to a linear transformation in (x, y) , we are left with an equation as follows:

$$x^2 \pm y^2 = k$$

In the case of positive sign, $x^2 + y^2 = k$, the solutions are the circle, when $k \geq 0$, the point, when $k = 0$, and \emptyset , when $k < 0$. As for the case of negative sign, $x^2 - y^2 = k$, which reads $(x - y)(x + y) = k$, here once again by linearity our equation becomes $xy = l$, which is a hyperbola when $l \neq 0$, and two lines when $l = 0$.

(4) In the case $b \neq 0$ the study is similar, with the same solutions, so we are left with the case $a = b = 0$. Here our conic is as follows, with $c, d, e, f \in \mathbb{R}$:

$$cxy + dx + ey + f = 0$$

If $c \neq 0$, by linearity our equation becomes $xy = l$, which produces a hyperbola or two lines, as explained before. As for the remaining case, $c = 0$, here our equation is:

$$dx + ey + f = 0$$

But this is generically the equation of a line, unless we are in the case $d = e = 0$, where our equation is $f = 0$, having as solutions \emptyset when $f \neq 0$, and \mathbb{R}^2 when $f = 0$.

(5) Thus, done with the classification, up to linear transformations as in (1). But this classification leads to the classification in general too, by applying now linear transformations to the solutions that we found. So, done with this, and very good.

(6) It remains to discuss the cone cutting. But here, what we have to do is to see how the cone equation $x^2 + y^2 = kz^2$ changes, under a linear change of coordinates, and then set $z = 0$, as to get the (x, y) equation of the intersection. But this leads, via some thinking or computations, to the conclusion that the cone equation $x^2 + y^2 = kz^2$ becomes in this way a degree 2 equation in (x, y) , which can be arbitrary, as desired. \square

Ready for some physics? And here, good news, not only what we did in the above is relevant, but is actually at the origin of the whole modern physics, thanks to:

THEOREM 8.2. *Planets and other celestial bodies move around the Sun on conics,*

$$C = \left\{ (x, y) \in \mathbb{R}^2 \mid P(x, y) = 0 \right\}$$

with $P \in \mathbb{R}[x, y]$ being of degree 2, which can be ellipses, parabolas or hyperbolas.

PROOF. This is something quite long, due to Kepler and Newton, as follows:

(1) The force of attraction between two bodies of masses M, m is given by:

$$||F|| = G \cdot \frac{Mm}{d^2}$$

Here d is the distance between the two bodies, and $G \simeq 6.674 \times 10^{-11}$ is a constant. Now assuming that M is fixed at the origin $0 \in \mathbb{R}^2$, the force exerted on m positioned at $z = (x, y) \in \mathbb{R}^2$, regarded as a vector $F \in \mathbb{R}^2$, is given by the following formula:

$$F = -||F|| \cdot \frac{z}{||z||} = -\frac{GMm}{||z||^2} \cdot \frac{z}{||z||} = -\frac{GMmz}{||z||^3}$$

But $F = ma = m\ddot{z}$, with $a = \ddot{z}$ being the acceleration, so the equation of motion of m , assuming that M is fixed at 0, and with the notation $K = GM$, is:

$$\ddot{z} = -\frac{Kz}{||z||^3}$$

(2) Let us write now our vector $z = (x, y)$ in polar coordinates, as follows:

$$x = r \cos \theta \quad , \quad y = r \sin \theta$$

We have then $||z|| = r$, and our equation of motion becomes:

$$\ddot{z} = -\frac{Kz}{r^3}$$

Let us differentiate now x, y . By using the standard calculus rules, we have:

$$\ddot{x} = \ddot{r} \cos \theta - 2\dot{r} \sin \theta \cdot \dot{\theta} - r \cos \theta \cdot \dot{\theta}^2 - r \sin \theta \cdot \ddot{\theta}$$

$$\ddot{y} = \ddot{r} \sin \theta + 2\dot{r} \cos \theta \cdot \dot{\theta} - r \sin \theta \cdot \dot{\theta}^2 + r \cos \theta \cdot \ddot{\theta}$$

Consider now the following two quantities, appearing as coefficients in the above:

$$a = \ddot{r} - r\dot{\theta}^2 \quad , \quad b = 2\dot{r}\dot{\theta} + r\ddot{\theta}$$

In terms of these quantities, our second derivative formulae read:

$$\ddot{x} = a \cos \theta - b \sin \theta \quad , \quad \ddot{y} = a \sin \theta + b \cos \theta$$

(3) We can now solve the equation of motion. Indeed, with the formulae that we found above for \ddot{x}, \ddot{y} , our equation of motion takes the following form:

$$a \cos \theta - b \sin \theta = -\frac{K}{r^2} \cos \theta \quad , \quad a \sin \theta + b \cos \theta = -\frac{K}{r^2} \sin \theta$$

But these two formulae can be written in the following way:

$$\left(a + \frac{K}{r^2}\right) \cos \theta = b \sin \theta \quad , \quad \left(a + \frac{K}{r^2}\right) \sin \theta = -b \cos \theta$$

By making now the product, and assuming that we are in a non-degenerate case, where the angle θ varies indeed, we obtain by positivity that we must have:

$$a + \frac{K}{r^2} = b = 0$$

(4) Let us first examine the second equation, $b = 0$. This can be solved as follows:

$$b = 0 \iff 2\dot{r}\dot{\theta} + r\ddot{\theta} = 0 \iff \dot{\theta} = \frac{\lambda}{r^2}$$

As for the first equation, namely $a + K/r^2 = 0$, with $K = \lambda^2/c$ this becomes:

$$\ddot{r} = \frac{\lambda^2}{r^2} \left(\frac{1}{r} - \frac{1}{c} \right)$$

Now set $f = 1/r$. With the convention that dots mean as usual derivatives with respect to t , and that primes denote derivatives with respect to $\theta = \theta(t)$, we have:

$$\dot{r} = -\frac{f'\dot{\theta}}{f^2} = -\frac{f'}{f^2} \cdot \frac{\lambda}{r^2} = -\lambda f'$$

By differentiating one more time with respect to t , we obtain:

$$\ddot{r} = -\lambda f''\dot{\theta} = -\lambda f'' \cdot \frac{\lambda}{r^2} = -\frac{\lambda^2}{r^2} f''$$

Thus, in terms of $f = 1/r$ as above, our equation for \ddot{r} simply reads:

$$f'' + f = \frac{1}{c}$$

But this latter equation is easy to solve, and by inverting, we obtain:

$$r = \frac{1}{f} = \frac{c}{1 + \varepsilon \cos \theta + \delta \sin \theta}$$

(5) But this leads to the conclusion that the trajectory is a conic. Indeed, in terms of the parameter θ , the formulae of the coordinates are:

$$x = \frac{c \cos \theta}{1 + \varepsilon \cos \theta + \delta \sin \theta} \quad , \quad y = \frac{c \sin \theta}{1 + \varepsilon \cos \theta + \delta \sin \theta}$$

We conclude from this that our coordinates x, y satisfy the following equation:

$$x^2 + y^2 = (\varepsilon x + \delta y - c)^2$$

But what we have here is an equation of a conic, and we are done. \square

All the above was very nice, all good old things, but the continuation of the story is more complicated, because beyond conics, things ramify. A first idea, in order to generalize the conics, is to look at the algebraic manifolds, in the following sense:

DEFINITION 8.3. *An algebraic manifold $X \subset \mathbb{R}^N$ is a space of the form*

$$X = \left\{ x \in \mathbb{R}^N \mid P_i(x) = 0 \right\}$$

with $P_i \in \mathbb{R}[x_1, \dots, x_N]$ being certain polynomials.

Manhy things can be said about such manifolds, both of mathematical and physical nature, and as a fundamental result here, generalizing Theorem 8.1, let us record:

THEOREM 8.4. *The degree 2 hypersurfaces $S \subset \mathbb{R}^N$, called quadrics, are up to degeneracy and to linear transformations the hypersurfaces of the following form,*

$$\pm x_1^2 \pm \dots \pm x_N^2 = 0, 1$$

and with the sphere being the only compact one.

PROOF. We have two statements here, the idea being as follows:

(1) The equations for a quadric $S \subset \mathbb{R}^N$ are best written as follows, with $A \in M_N(\mathbb{R})$ being a matrix, $B \in M_{1 \times N}(\mathbb{R})$ being a row vector, and $C \in \mathbb{R}$ being a constant:

$$\langle Ax, x \rangle + Bx + C = 0$$

(2) By doing the linear algebra, or by invoking the theorem of Sylvester on quadratic forms, we are left, modulo linear transformations, with signed sums of squares:

$$\pm x_1^2 \pm \dots \pm x_N^2 = 0, 1$$

(3) To be more precise, with linear algebra, by evenly distributing the terms $x_i x_j$ above and below the diagonal, we can assume that our matrix $A \in M_N(\mathbb{R})$ is symmetric. Thus A must be diagonalizable, and by changing the basis of \mathbb{R}^N , as to have it diagonal, our equation becomes as follows, with $D \in M_N(\mathbb{R})$ being now diagonal:

$$\langle Dx, x \rangle + Ex + F = 0$$

(4) But now, by making squares in the obvious way, which amounts in applying yet another linear transformation to our quadric, the equation takes the following form, with $G \in M_N(-1, 0, 1)$ being diagonal, and with $H \in \{0, 1\}$ being a constant:

$$\langle Gx, x \rangle = H$$

(5) Now barring the degenerate cases, we can further assume $G \in M_N(-1, 1)$, and we are led in this way to the equation claimed in (2) above, namely:

$$\pm x_1^2 \pm \dots \pm x_N^2 = 0, 1$$

(6) In particular we see that, up to some degenerate cases, namely emptyset and point, the only compact quadric, up to linear transformations, is the one given by:

$$x_1^2 + \dots + x_N^2 = 1$$

(7) But this is the unit sphere, so are led to the conclusions in the statement. □

Summarizing, we have some beginning of algebraic geometry theory going on. However, for our purposes here, which are integration, along with applications to physics, although many interesting algebraic manifolds appear at the advanced level, in what concerns the basics, here we are mostly in need of a different definition, namely:

DEFINITION 8.5. *A smooth manifold is a space X which is locally isomorphic to \mathbb{R}^N . To be more precise, this space X must be covered by charts, bijectively mapping open pieces of it to open pieces of \mathbb{R}^N , with the changes of charts being C^∞ functions.*

As a basic example, we have \mathbb{R}^N itself, or any open subset $X \subset \mathbb{R}^N$. Another example is the circle, or curves like ellipses and so on, for obvious reasons. However, this might seem a bit vague, and here is a more precise statement in this sense:

THEOREM 8.6. *The following are smooth manifolds, in the plane:*

- (1) *The circles.*
- (2) *The ellipses.*
- (3) *The non-degenerate conics.*
- (4) *Smooth deformations of these.*

PROOF. All this is quite intuitive, the idea being as follows:

(1) Consider the unit circle, $x^2 + y^2 = 1$. We can write then $x = \cos t$, $y = \sin t$, with $t \in [0, 2\pi)$, and we seem to have here the solution to our problem, just using 1 chart. But this is of course wrong, because $[0, 2\pi)$ is not open, and we have a problem at 0. In practice we need to use 2 such charts, say with the first one being with $t \in (0, 3\pi/2)$, and the second one being with $t \in (\pi, 5\pi/2)$. As for the fact that the change of charts is indeed smooth, this comes by writing down the formulae, or just thinking a bit, and arguing that this change of chart being actually a translation, it is automatically linear.

(2) This follows from (1), by pulling the circle in both the Ox and Oy directions, and the formulae here, based on our various formulae above, are left to you reader.

(3) We already have the ellipses, and the case of the parabolas and hyperbolas is elementary as well, and in fact simpler than the case of the ellipses. Indeed, a parabolola is clearly homeomorphic to \mathbb{R} , and a hyperbola, to two copies of \mathbb{R} .

(4) This is something which is clear too, depending of course on what exactly we mean by “smooth deformation”, and by using a bit of multivariable calculus if needed. \square

In higher dimensions now, as basic examples here, we have the unit sphere in \mathbb{R}^N , and smooth deformations of it, once again, somehow by obvious reasons, and with the formal proof of this fact coming for instance by using the spherical coordinate formulae from chapter 7. Alternatively, we can use the stereographic projection, which provides us with a proof that the sphere $S_{\mathbb{R}}^{N-1}$ is indeed a smooth manifold, with just 2 charts.

8b. Contour integrals

We would like to discuss now the integration over curves, surfaces, and other smooth manifolds. Let us start with something very basic, regarding the curves, namely:

THEOREM 8.7. *The length of a curve $\gamma : [a, b] \rightarrow \mathbb{R}^N$ is given by*

$$L(\gamma) = \int_a^b \|\gamma'(t)\| dt$$

with $\|\cdot\|$ being the usual norm of \mathbb{R}^N .

PROOF. This is something quite intuitive, that can even stand as a definition for the length of the curves, and that we will not prove in detail here, the idea being as follows:

(1) To start with, what is the length of a curve? Good question, and in answer, a physicist would say that this is the quantity obtained by integrating the magnitude of the velocity vector over the curve, with respect to time. But this velocity vector is $\gamma'(t)$, having magnitude $\|\gamma'(t)\|$, so we are led to the formula in the statement.

(2) Regarding now mathematicians, these would say that the length of a curve is the following quantity, with $(t_1 = a, t_2, \dots, t_{n-1}, t_n = b)$ being a uniform division of (a, b) :

$$L(\gamma) = \lim_{n \rightarrow \infty} \sum_{i=1}^n \|f(t_i) - f(t_{i-1})\|$$

But, by using the fundamental theorem of calculus, we can write this as follows:

$$L(\gamma) = \lim_{n \rightarrow \infty} \sum_{i=1}^n \left\| \int_{t_{i-1}}^{t_i} f'(t) dt \right\|$$

And the point now is that, by doing some standard analysis, that we will leave here as an instructive exercise, we are led to the formula in the statement.

(3) So, we have our definition-theorem for the length of the curves, and for the discussion to be complete, we just need an illustration for this. But here, we have the following standard computation, for the length of the ellipses $(x/a)^2 + (y/b)^2 = 1$:

$$\begin{aligned} L &= 4 \int_0^{\pi/2} \sqrt{\left(\frac{da \cos t}{dt}\right)^2 + \left(\frac{db \sin t}{dt}\right)^2} dt \\ &= 4 \int_0^{\pi/2} \sqrt{a^2 \sin^2 t + b^2 \cos^2 t} dt \end{aligned}$$

There are of course many other applications of the formula in the statement, and we will leave some thinking and computations here as an instructive exercise. \square

Next, let us talk about the contour integrals over curves. We have here:

THEOREM 8.8. *Given a path $\gamma \subset \mathbb{R}^3$, we can talk about integrals of type*

$$I = \int_{\gamma} f(x)dx_1 + g(x)dx_2 + h(x)dx_3$$

with $f, g, h : \mathbb{R}^3 \rightarrow \mathbb{R}$, which are independent on the chosen parametrization of the path.

PROOF. This is something quite straightforward, the idea being as follows:

(1) Regarding the statement itself, assume indeed that we have a path in \mathbb{R}^3 , which can be best thought of as corresponding to a function as follows:

$$\gamma : [a, b] \rightarrow \mathbb{R}^3$$

Observe that this function γ is not exactly the path itself, for instance because the following functions produce the same path, parametrized differently:

$$\delta : [0, b-a] \rightarrow \mathbb{R}^3 \quad , \quad \delta(t) = \gamma(t+a)$$

$$\varepsilon : [0, 1] \rightarrow \mathbb{R}^3 \quad , \quad \varepsilon(t) = \delta((b-a)t)$$

$$\varphi : [0, 1] \rightarrow \mathbb{R}^3 \quad , \quad \varphi(t) = \varepsilon(t^2)$$

$$\psi : [0, 1] \rightarrow \mathbb{R}^3 \quad , \quad \psi(t) = \varepsilon(1-t)$$

⋮

Our claim, however, is that we can talk about integrals as follows, with $f, g, h : \mathbb{R}^3 \rightarrow \mathbb{R}$, which are independent on the chosen parametrization of our path:

$$I = \int_{\gamma} f(x)dx_1 + g(x)dx_2 + h(x)dx_3$$

(2) In order to prove this, let us choose a parametrization $\gamma : [a, b] \rightarrow \mathbb{R}^3$ as above. This parametrization has as components three functions $\gamma_1, \gamma_2, \gamma_3$, given by:

$$\gamma = (\gamma_1, \gamma_2, \gamma_3) : [a, b] \rightarrow \mathbb{R}^3$$

In order to construct the integral I , it is quite clear, by suitably cutting our path into pieces, that we can restrict the attention to the case where all three components $\gamma_1, \gamma_2, \gamma_3 : [a, b] \rightarrow \mathbb{R}$ are increasing, or decreasing. Thus, we can assume that these three components are as follows, increasing or decreasing, and bijective on their images:

$$\gamma_i : [a, b] \rightarrow [a_i, b_i]$$

(3) Moreover, by using the obvious symmetry between the coordinates x_1, x_2, x_3 , in order to construct I , we just need to construct integrals of the following type:

$$I_1 = \int_{\gamma} f(x)dx_1$$

(4) So, let us construct this latter integral I_1 , under the assumptions in (2). The simplest case is when the first path, $\gamma_1 : [a, b] \rightarrow [a_1, b_1]$, is the identity:

$$\gamma_1 : [a, b] \rightarrow [a, b] \quad , \quad \gamma_1(x) = x$$

In other words, the simplest case is when our path is of the following form, with $\gamma_2, \gamma_3 : [a, b] \rightarrow \mathbb{R}$ being certain functions, that should be increasing or decreasing, as per our conventions (2) above, but in what follows we will not need this assumption:

$$\gamma(x_1) = (x_1, \gamma_2(x_1), \gamma_3(x_1))$$

But now, with this convention made, we can define our contour integral, or rather its first component, as explained above, as a usual one-variable integral, as follows:

$$I_1 = \int_a^b f(x_1, \gamma_2(x_1), \gamma_3(x_1)) dx_1$$

(5) With this understood, let us examine now the general case, where the first path, $\gamma_1 : [a, b] \rightarrow [a_1, b_1]$, is arbitrary, increasing or decreasing, and bijective on its image. In this case we can reparametrize our curve, as to have it as in (4) above, as follows:

$$\tilde{\gamma} = (id, \gamma_2 \gamma_1^{-1}, \gamma_3 \gamma_1^{-1}) : [a_1, b_1] \rightarrow \mathbb{R}^3$$

Now since we want our integral $I_1 = \int_{\gamma} f(x) dx_1$ to be independent of the parametrization, we are led to the following formula for it, coming from the formula in (4):

$$\begin{aligned} I_1 &= \int_{\tilde{\gamma}} f(x) dx_1 \\ &= \int_{a_1}^{b_1} f(x_1, \gamma_2 \gamma_1^{-1}(x_1), \gamma_3 \gamma_1^{-1}(x_1)) dx_1 \\ &= \int_a^b f(\gamma_1(y_1), \gamma_2(y_1), \gamma_3(y_1)) \gamma_1'(y_1) dy_1 \end{aligned}$$

Here we have used at the end the change of variable formula, with $x_1 = \gamma_1(y_1)$.

(6) Thus, job done, we have our definition for the contour integrals, with the formula being as follows, obtained by using (5) for all three coordinates x_1, x_2, x_3 :

$$\begin{aligned} I &= \int_a^b f(\gamma_1(y_1), \gamma_2(y_1), \gamma_3(y_1)) \gamma_1'(y_1) dy_1 \\ &\quad + \int_a^b g(\gamma_1(y_2), \gamma_2(y_2), \gamma_3(y_2)) \gamma_2'(y_2) dy_2 \\ &\quad + \int_a^b h(\gamma_1(y_3), \gamma_2(y_3), \gamma_3(y_3)) \gamma_3'(y_3) dy_3 \end{aligned}$$

And with this, we are led to the conclusion in the statement. □

Let us record as well the following more compact form of Theorem 8.8:

THEOREM 8.9. *The contour integrals over a curve $\gamma : [a, b] \rightarrow \mathbb{R}^3$ are given by*

$$\int_{\gamma} \langle F(x), dx \rangle = \int_a^b \langle F(\gamma(y)), \gamma'(y) dy \rangle$$

valid for any $F : \mathbb{R}^3 \rightarrow \mathbb{R}^3$, where on the right $\gamma'(y) dy = (\gamma'_i(y_i) dy_i)_i$.

PROOF. This is a fancy reformulation of what we did in Theorem 8.8 and its proof. Indeed, with the notation $F = (F_1, F_2, F_3) = (f, g, h)$, the integral computed there is:

$$\int_{\gamma} \langle F(x), dx \rangle = \int_{\gamma} F_1(x) dx_1 + F_2(x) dx_2 + F_3(x) dx_3$$

As for the value of this integral, according to the proof of Theorem 8.8, this is:

$$\begin{aligned} \int_{\gamma} \langle F(x), dx \rangle &= \int_a^b F_1(\gamma_1(y_1), \gamma_2(y_1), \gamma_3(y_1)) \gamma'_1(y_1) dy_1 \\ &\quad + \int_a^b F_2(\gamma_1(y_2), \gamma_2(y_2), \gamma_3(y_2)) \gamma'_2(y_2) dy_2 \\ &\quad + \int_a^b F_3(\gamma_1(y_3), \gamma_2(y_3), \gamma_3(y_3)) \gamma'_3(y_3) dy_3 \end{aligned}$$

Now observe that we can write this in a more compact way, as follows:

$$\begin{aligned} \int_{\gamma} \langle F(x), dx \rangle &= \int_a^b F_1(\gamma(y_1)) \gamma'(y_1) dy_1 \\ &\quad + \int_a^b F_2(\gamma(y_2)) \gamma'(y_2) dy_2 \\ &\quad + \int_a^b F_3(\gamma(y_3)) \gamma'(y_3) dy_3 \end{aligned}$$

And we can do even better. Indeed, we have only one integral here, \int_a^b , and in order to best express the integrand, consider the formal vector in the statement, namely:

$$\gamma'(y) dy = \begin{pmatrix} \gamma'_1(y_1) dy_1 \\ \gamma'_2(y_2) dy_2 \\ \gamma'_3(y_3) dy_3 \end{pmatrix}$$

Our integrand appears then as the scalar product of $F(\gamma(y))$ with this latter vector $\gamma'(y) dy$, so our formula above for the contour integral takes the following form:

$$\int_{\gamma} \langle F(x), dx \rangle = \int_a^b \langle F(\gamma(y)), \gamma'(y) dy \rangle$$

Thus, we are led to the conclusion in the statement. □

More concretely now, let us temporarily forget about the paths γ , and have a look at the quantities which are to be integrated, namely:

$$\alpha = F_1(x)dx_1 + F_2(x)dx_2 + F_3(x)dx_3$$

Obviously, these are something rather mathematical, and many things can be said here. However, we can have some physical intuition on them. Assume indeed that we are given a function as follows, that you can think for instance as corresponding to an external force, with $F(x) \in \mathbb{R}^3$ being the force vector applied at a given point $x \in \mathbb{R}^3$:

$$F : \mathbb{R}^3 \rightarrow \mathbb{R}^3$$

By writing $F = (F_1, F_2, F_3)$, we can then consider the following quantity, and when $F : \mathbb{R}^3 \rightarrow \mathbb{R}^3$ varies, we obtain exactly the abstract quantities α considered above:

$$\langle F(x), dx \rangle = F_1(x)dx_1 + F_2(x)dx_2 + F_3(x)dx_3$$

Thus, all in all, what we have done in the above with our construction of contour integrals, was to define quantities as follows, with γ being a path in \mathbb{R}^3 , and with $F : \mathbb{R}^3 \rightarrow \mathbb{R}^3$ being a certain function, that we can think of, if we want, as being a force:

$$I = \int_{\gamma} \langle F(x), dx \rangle$$

Which brings us into physics. Indeed, by assuming now that $F : \mathbb{R}^3 \rightarrow \mathbb{R}^3$ does correspond to a force, we can formulate the following definition:

DEFINITION 8.10. *The work done by a force $F = F(x)$ for moving a particle from point $p \in \mathbb{R}^3$ to point $q \in \mathbb{R}^3$ via a given path $\gamma : p \rightarrow q$ is the following quantity:*

$$W(\gamma) = \int_{\gamma} \langle F(x), dx \rangle$$

We say that F is conservative if this work quantity $W(\gamma)$ does not depend on the chosen path $\gamma : p \rightarrow q$, and in this case we denote this quantity by $W(p, q)$.

This definition is something quite subtle, and as a first comment, assume that we have two paths $\gamma : p \rightarrow q$ and $\delta : p \rightarrow q$. We can then consider the path $\circ : p \rightarrow p$ obtained by going along $\gamma : p \rightarrow q$, and then along δ reversed, $\delta^{-1} : q \rightarrow p$, and we have:

$$W(\circ) = W(\gamma) - W(\delta)$$

Thus F is conservative precisely when, for any loop $\circ : p \rightarrow p$, we have:

$$W(\circ) = 0$$

Intuitively, this means that F is some sort of “clean”, ideal force, with no dirty things like friction involved. As we will soon see, gravity is such a clean force, with a simple example coming from throwing a rock up in the sky. That rock will travel on a loop $p \rightarrow q \rightarrow p$, and will come back here to p unchanged, save for the fact that its speed

vector is reversed. Thus, and assuming now that work has something to do with energy, which is intuitive, there has been no overall work of gravity on this loop, $W(\circ) = 0$.

As a first result now, regarding the conservative forces, we have:

THEOREM 8.11. *The work done by a conservative force F on a mass m object is*

$$W(p, q) = T(q) - T(p)$$

with $T = m||v||^2/2$ standing as usual for the kinetic energy of the object.

PROOF. Assuming that F is conservative, and acts via the usual formula $F = ma$ on our object of mass m , we have the following computation, as desired:

$$\begin{aligned} W(p, q) &= \int_p^q \langle F(x), dx \rangle \\ &= m \int_p^q \langle a(x), dx \rangle \\ &= m \int_p^q \left\langle \frac{dv(x)}{dt}, v(x) dt \right\rangle \\ &= \frac{m}{2} \int_p^q \frac{d \langle v(x), v(x) \rangle}{dt} dt \\ &= \frac{m}{2} \int_p^q \frac{d||v(x)||^2}{dt} dt \\ &= \frac{m}{2} (||v(q)||^2 - ||v(p)||^2) \\ &= T(q) - T(p) \end{aligned}$$

Here we have used in the middle the fact that the time derivative of a scalar product of functions $\langle v, w \rangle$ consists of two terms, which are equal when $v = w$. \square

In order to formulate our next result, observe that we have the following computation for the contour integrals of gradients, which is independent on the chosen path:

$$\begin{aligned} \int_p^q \langle \nabla V(x), dx \rangle &= \int_p^q \frac{dV}{dx_1} \cdot dx_1 + \frac{dV}{dx_2} \cdot dx_2 + \frac{dV}{dx_3} \cdot dx_3 \\ &= \int_p^q dV \\ &= V(q) - V(p) \end{aligned}$$

To be more precise, this computation certainly works when V is a function of just one variable, x_1 , x_2 or x_3 , thanks to the fundamental theorem of calculus, and the general case follows from this, by using the chain rule for derivatives.

Now with this formula in hand, we can formulate the following result, which is quite conceptual, and which includes some basic gravitation physics too, at the end:

THEOREM 8.12. *A force F is conservative precisely when it is of the form*

$$F = -\nabla V$$

for a certain function V , and in this case the work done by it is given by:

$$W(p, q) = V(p) - V(q)$$

Also, the gravitation force is conservative, coming from $V = -C/||x||$, with $C > 0$.

PROOF. This is something quite tricky, the idea being as follows:

(1) In one sense, assume that F is conservative. Since the work $W(p, q) = W(\gamma)$ is independent of the chosen path $\gamma : p \rightarrow q$, we can find a function V such that:

$$W(p, q) = V(p) - V(q)$$

Observe that this function V is well-defined up to an additive constant. Now with this formula in hand, we further obtain, as desired:

$$\begin{aligned} W(p, q) = V(p) - V(q) &\implies \int_p^q \langle F(x), dx \rangle = - \int_p^q \langle \nabla V(x), dx \rangle \\ &\implies \langle F(x), dx \rangle = - \langle \nabla V(x), dx \rangle \\ &\implies F_i(x) dx_i = - \frac{dV}{dx_i} \cdot dx_i \\ &\implies F_i(x) = - \frac{dV}{dx_i} \\ &\implies F = -\nabla V \end{aligned}$$

(2) In the other sense now, assuming $F = -\nabla V$, we have the following computation, valid for any loop $\circ : p \rightarrow p$, which shows that F is indeed conservative:

$$W(\circ) = - \int_{\circ} \nabla V = 0$$

More generally, regarding the work done by such a force $F = -\nabla V$, along a path $\gamma : p \rightarrow q$, which is independent on this path γ , this is given by:

$$W(p, q) = - \int_p^q \nabla V = V(p) - V(q)$$

(3) Finally, regarding the last assertion, we will leave this as an exercise. □

We can now put everything together, and we have the following result:

THEOREM 8.13. *Given a conservative force F , appearing as follows, with V being uniquely determined up to an additive constant,*

$$F = -\nabla V$$

the movements of a particle under F preserve the total energy, given by

$$E = T + V$$

with $T = m||v||^2/2$ being the kinetic energy, and with V being called potential energy.

PROOF. By using Theorem 8.11 and Theorem 8.12, we have:

$$W(p, q) = T(q) - T(p) \quad , \quad W(p, q) = V(p) - V(q)$$

Now observe that these equalities give the following formula:

$$T(p) + V(p) = T(q) + V(q)$$

Thus, the total energy $E = T + V$ is conserved, as claimed. \square

8c. Green and Stokes

At the level of mathematics now, in relation with the above, a useful result is:

THEOREM 8.14 (Green). *Given a plane curve $C \subset \mathbb{R}^2$, we have the formula*

$$\int_C Pdx + Qdy = \int_D \left(\frac{dQ}{dx} - \frac{dP}{dy} \right) dxdy$$

where $D \subset \mathbb{R}^2$ is the domain enclosed by C .

PROOF. Assume indeed that we have a plane curve $C \subset \mathbb{R}^2$, without self-intersections, which is piecewise C^1 , and is assumed to be counterclockwise oriented. In order to prove the formula regarding P , we can parametrize the enclosed domain D as follows:

$$D = \left\{ (x, y) \mid a \leq x \leq b, f(x) \leq y \leq g(x) \right\}$$

We have then the following computation, which gives the result for P :

$$\begin{aligned} \int_D -\frac{dP}{dy} dxdy &= \int_a^b \left(\int_{f(x)}^{g(x)} -\frac{dP}{dy}(x, y) dy \right) dx \\ &= \int_a^b P(x, f(x)) - P(x, g(x)) dx \\ &= \int_a^b P(x, f(x)) dx + \int_b^a P(x, g(x)) dx \\ &= \int_C Pdx \end{aligned}$$

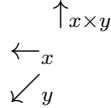
As for the result for the Q term, the computation here is similar. \square

Moving forward to 3D, let us start with a standard definition, as follows:

DEFINITION 8.15. *The vector product of two vectors in \mathbb{R}^3 is given by*

$$x \times y = \|x\| \cdot \|y\| \cdot \sin \theta \cdot n$$

where $n \in \mathbb{R}^3$ with $n \perp x, y$ and $\|n\| = 1$ is constructed using the right-hand rule:



Alternatively, in usual vertical linear algebra notation for all vectors,

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \times \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} x_2 y_3 - x_3 y_2 \\ x_3 y_1 - x_1 y_3 \\ x_1 y_2 - x_2 y_1 \end{pmatrix}$$

the rule being that of computing 2×2 determinants, and adding a middle sign.

In practice now, in order to get familiar with the vector products, nothing better than doing some classical mechanics. We have here the following key result:

THEOREM 8.16. *In the gravitational 2-body problem, the angular momentum*

$$J = x \times p$$

with $p = mv$ being the usual momentum, is conserved.

PROOF. There are several things to be said here, the idea being as follows:

(1) First of all the usual momentum, $p = mv$, is not conserved, because the simplest solution is the circular motion, where the moment gets turned around. But this suggests precisely that, in order to fix the lack of conservation of the momentum p , what we have to do is to make a vector product with the position x . Leading to J , as above.

(2) Regarding now the proof, consider indeed a particle m moving under the gravitational force of a particle M , assumed, as usual, to be fixed at 0. By using the fact that for two proportional vectors, $p \sim q$, we have $p \times q = 0$, we obtain:

$$\begin{aligned} \dot{J} &= \dot{x} \times p + x \times \dot{p} \\ &= v \times mv + x \times ma \\ &= m(v \times v + x \times a) \\ &= m(0 + 0) \\ &= 0 \end{aligned}$$

Now since the derivative of J vanishes, this quantity is constant, as stated. \square

At the mathematical level now, we have the following key result:

THEOREM 8.17. *The area of a surface $S \subset \mathbb{R}^3$, parametrized as $S = r(D)$ with $r : D \rightarrow \mathbb{R}^3$ and $D \subset \mathbb{R}^2$, is given by*

$$A(S) = \int_D \|r_x \times r_y\| \, dx dy$$

with $r_x, r_y : D \rightarrow \mathbb{R}^3$ being the partial derivatives of r , and $\|\cdot\|$ being the norm of \mathbb{R}^3 .

PROOF. Thus is something quite similar to Theorem 8.7, and many things can be said here, at the theoretical level, in analogy with those for the curves. Among others, let us mention that we can talk, more generally, about surface integrals, defined as follows:

$$\int_S f(s) ds = \int_D f(r(x, y)) \|r_x \times r_y\| \, dx dy$$

As for the basic illustrations of the formula in the statement, consider a surface of type $z = f(x, y)$. Here we have $r(x, y) = (x, y, f(x, y))$, and we obtain:

$$\begin{aligned} A(S) &= \int_D \left\| \begin{pmatrix} 1 \\ 0 \\ f_x \end{pmatrix} \times \begin{pmatrix} 0 \\ 1 \\ f_y \end{pmatrix} \right\| \, dx dy \\ &= \int_D \left\| \begin{pmatrix} -f_x \\ -f_y \\ 1 \end{pmatrix} \right\| \, dx dy \\ &= \int_D \sqrt{f_x^2 + f_y^2 + 1} \, dx dy \end{aligned}$$

There are of course many other applications of the formula in the statement, and we will leave some thinking and computations here as an instructive exercise. \square

Many other things can be said, as a continuation of the above, but let us not deviate too much, from what we wanted to do here. Next, comes the following definition:

DEFINITION 8.18. *We can talk about the divergence of $\varphi : \mathbb{R}^3 \rightarrow \mathbb{R}^3$, as being*

$$\langle \nabla, \varphi \rangle = \left\langle \begin{pmatrix} \frac{d}{dx} \\ \frac{d}{dy} \\ \frac{d}{dz} \end{pmatrix}, \begin{pmatrix} \varphi_x \\ \varphi_y \\ \varphi_z \end{pmatrix} \right\rangle = \frac{d\varphi_x}{dx} + \frac{d\varphi_y}{dy} + \frac{d\varphi_z}{dz}$$

as well as about the curl of the same function $\varphi : \mathbb{R}^3 \rightarrow \mathbb{R}^3$, as being

$$\nabla \times \varphi = \begin{vmatrix} u_x & \frac{d}{dx} & \varphi_x \\ u_y & \frac{d}{dy} & \varphi_y \\ u_z & \frac{d}{dz} & \varphi_z \end{vmatrix} = \begin{pmatrix} \frac{d\varphi_z}{dy} - \frac{d\varphi_y}{dz} \\ \frac{d\varphi_x}{dz} - \frac{d\varphi_z}{dx} \\ \frac{d\varphi_y}{dx} - \frac{d\varphi_x}{dy} \end{pmatrix}$$

where u_x, u_y, u_z are the unit vectors along the coordinate directions x, y, z .

Getting back now to calculus tools, in 3 dimensions, we have the following result:

THEOREM 8.19 (Stokes). *Given a smooth oriented surface $S \subset \mathbb{R}^3$, with boundary $C \subset \mathbb{R}^3$, and a vector field F , we have the following formula:*

$$\int_S \langle (\nabla \times F)(x), n(x) \rangle dx = \int_C \langle F(x), dx \rangle$$

In other words, the line integral of a vector field F over a loop C equals the surface integral of the curl of the vector field φ , over the enclosed surface S .

PROOF. This basically follows from the Green theorem, the idea being as follows:

(1) Let us first parametrize our surface S , and its boundary C . We can assume that we are in the situation where we have a closed oriented curve $\gamma : [a, b] \rightarrow \mathbb{R}^2$, with interior $D \subset \mathbb{R}^2$, and where the surface appears as $S = \psi(D)$, with $\psi : D \rightarrow \mathbb{R}^3$. In this case, the function $\delta = \psi \circ \gamma$ parametrizes the boundary of our surface, $C = \delta[a, b]$.

(2) Let us first look at the integral on the right in the statement. We have the following formula, with $J_y(\psi)$ standing for the Jacobian of ψ at the point $y = \gamma(t)$:

$$\begin{aligned} \int_C \langle F(x), dx \rangle &= \int_\gamma \langle F(\psi(\gamma)), d\psi(\gamma) \rangle \\ &= \int_\gamma \langle F(\psi(y)), J_y(\psi) d\gamma \rangle \end{aligned}$$

In order to further process this formula, let us introduce the following function:

$$P(u, v) = \left\langle F(\psi(u, v)), \frac{d\psi}{du}(u, v) \right\rangle e_u + \left\langle F(\psi(u, v)), \frac{d\psi}{dv}(u, v) \right\rangle e_v$$

In terms of this function, we have the following formula for our line integral:

$$\int_C \langle F(x), dx \rangle = \int_\gamma \langle P(y), dy \rangle$$

(3) In order to compute now the other integral in the statement, we first have:

$$\begin{aligned}
& \frac{dP_v}{du} - \frac{dP_u}{dv} \\
&= \left\langle \frac{d(F\psi)}{du}, \frac{d\psi}{dv} \right\rangle + \left\langle F\psi, \frac{d^2\psi}{dudv} \right\rangle - \left\langle \frac{d(F\psi)}{dv}, \frac{d\psi}{du} \right\rangle - \left\langle F\psi, \frac{d^2\psi}{dvdu} \right\rangle \\
&= \left\langle \frac{d(F\psi)}{du}, \frac{d\psi}{dv} \right\rangle - \left\langle \frac{d(F\psi)}{dv}, \frac{d\psi}{du} \right\rangle \\
&= \left\langle \frac{d\psi}{dv}, (J_{\psi(u,v)}F - (J_{\psi(u,v)}F)^t) \frac{d\psi}{du} \right\rangle \\
&= \left\langle \frac{d\psi}{dv}, (\nabla \times F) \times \frac{d\psi}{du} \right\rangle \\
&= \left\langle \nabla \times F, \frac{d\psi}{du} \times \frac{d\psi}{dv} \right\rangle
\end{aligned}$$

We conclude that the integral on the left in the statement is given by:

$$\begin{aligned}
& \int_S \langle (\nabla \times F)(x), n(x) \rangle dx \\
&= \int_D \left\langle (\nabla \times F)(\psi(u, v)), \frac{d\psi}{du}(u, v) \times \frac{d\psi}{dv}(u, v) \right\rangle dudv \\
&= \int_D \left(\frac{dP_v}{du} - \frac{dP_u}{dv} \right) dudv
\end{aligned}$$

(4) But with this, we are done, because the integrals computed in (2) and (3) are indeed equal, due to the Green theorem. Thus, the Stokes formula holds indeed. \square

As a conclusion to what we did so far, we have the following statement:

THEOREM 8.20. *The following results hold, in 3 dimensions:*

(1) *Fundamental theorem for gradients, namely*

$$\int_a^b \langle \nabla f, dx \rangle = f(b) - f(a)$$

(2) *Fundamental theorem for divergences, or Gauss or Green formula,*

$$\int_B \langle \nabla, \varphi \rangle = \int_S \langle \varphi(x), n(x) \rangle dx$$

(3) *Fundamental theorem for curls, or Stokes formula,*

$$\int_A \langle (\nabla \times \varphi)(x), n(x) \rangle dx = \int_P \langle \varphi(x), dx \rangle$$

where S is the boundary of the body B , and P is the boundary of the area A .

PROOF. This follows indeed from the various formulae established above. \square

8d. Gauss, Maxwell

We would like to discuss now, following Gauss and others, some applications of the above to electrostatics, and why not to electrodynamics too. Let us start with:

FACT 8.21 (Coulomb law). *Any pair of charges $q_1, q_2 \in \mathbb{R}$ is subject to a force as follows, which is attractive if $q_1 q_2 < 0$ and repulsive if $q_1 q_2 > 0$,*

$$||F|| = K \cdot \frac{|q_1 q_2|}{d^2}$$

where $d > 0$ is the distance between the charges, and $K > 0$ is a certain constant.

Observe the amazing similarity with the Newton law for gravity. However, as we will soon see, passed a few simple facts, things will be more complicated here.

In analogy with our study of gravity, let us start with:

DEFINITION 8.22. *Given charges $q_1, \dots, q_k \in \mathbb{R}$ located at positions $x_1, \dots, x_k \in \mathbb{R}^3$, we define their electric field to be the vector function*

$$E(x) = K \sum_i \frac{q_i(x - x_i)}{||x - x_i||^3}$$

so that their force applied to a charge $Q \in \mathbb{R}$ positioned at $x \in \mathbb{R}^3$ is given by $F = QE$.

Observe the analogy with gravity, save for the fact that instead of masses $m > 0$ we have now charges $q \in \mathbb{R}$, and that at the level of constants, G gets replaced by K .

More generally, we will be interested in electric fields of various non-discrete charge configurations, such as charged curves, surfaces and solid bodies. So, let us formulate:

DEFINITION 8.23. *The electric field of a charge configuration $L \subset \mathbb{R}^3$, with charge density function $\rho : L \rightarrow \mathbb{R}$, is the vector function*

$$E(x) = K \int_L \frac{\rho(z)(x - z)}{||x - z||^3} dz$$

so that the force of L applied to a charge Q positioned at x is given by $F = QE$.

With the above definitions in hand, it is most convenient now to forget about the charges, and focus on the study of the corresponding electric fields E .

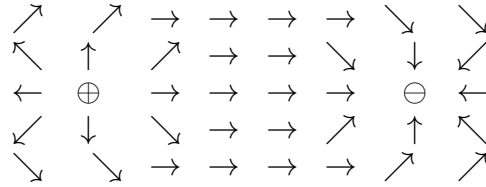
These fields are by definition vector functions $E : \mathbb{R}^3 \rightarrow \mathbb{R}^3$, with the convention that they take $\pm\infty$ values at the places where the charges are located, and intuitively, are best represented by their field lines, which are constructed as follows:

DEFINITION 8.24. *The field lines of an electric field $E : \mathbb{R}^3 \rightarrow \mathbb{R}^3$ are the oriented curves $\gamma \subset \mathbb{R}^3$ pointing at every point $x \in \mathbb{R}^3$ at the direction of the field, $E(x) \in \mathbb{R}^3$.*

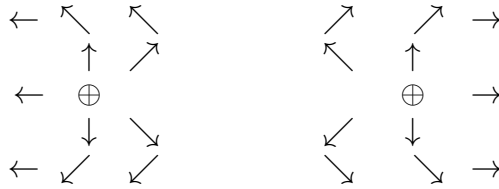
As a basic example here, for one charge the field lines are the half-lines emanating from its position, oriented according to the sign of the charge:



For two charges now, if these are of opposite signs, $+$ and $-$, you get a picture that you are very familiar with, namely that of the field lines of a bar magnet:



If the charges are $+, +$ or $-, -$, you get something of similar type, but repulsive this time, with the field lines emanating from the charges being no longer shared:



These pictures, and notably the last one, with $+, +$ charges, are quite interesting, because the repulsion situation does not appear in the context of gravity. Thus, we can only expect our geometry here to be far more complicated than that of gravity.

In general now, the first thing that can be said about the field lines is that, by definition, they do not cross. Thus, what we have here is some sort of oriented 1D foliation of \mathbb{R}^3 , in the sense that \mathbb{R}^3 is smoothly decomposed into oriented curves $\gamma \subset \mathbb{R}^3$.

The field lines, as constructed in Definition 8.24, obviously do not encapsulate the whole information about the field, with the direction of each vector $E(x) \in \mathbb{R}^3$ being there, but with the magnitude $\|E(x)\| \geq 0$ of this vector missing. However, say when drawing, when picking up uniformly radially spaced field lines around each charge, and with the number of these lines proportional to the magnitude of the charge, and then completing the picture, the density of the field lines around each point $x \in \mathbb{R}$ will give you the magnitude $\|E(x)\| \geq 0$ of the field there, up to a scalar. So, let us formulate:

PROPOSITION 8.25. *Given an electric field $E : \mathbb{R}^3 \rightarrow \mathbb{R}^3$, the knowledge of its field lines is the same as the knowledge of the composition*

$$nE : \mathbb{R}^3 \rightarrow \mathbb{R}^3 \rightarrow S$$

where $S \subset \mathbb{R}^3$ is the unit sphere, and $n : \mathbb{R}^3 \rightarrow S$ is the rescaling map, namely:

$$n(x) = \frac{x}{\|x\|}$$

However, in practice, when the field lines are accurately drawn, the density of the field lines gives you the magnitude of the field, up to a scalar.

PROOF. We have two assertions here, the idea being as follows:

(1) The first assertion is clear from definitions, with of course our usual convention that the electric field and its problematics take place outside the locations of the charges, which makes everything in the statement to be indeed well-defined.

(2) Regarding now the last assertion, which is of course a bit informal, this follows from the above discussion. It is possible to be a bit more mathematical here, with a definition, formula and everything, but we will not need this, in what follows. \square

Let us introduce now a key definition, as follows:

DEFINITION 8.26. *The flux of an electric field $E : \mathbb{R}^3 \rightarrow \mathbb{R}^3$ through a surface $S \subset \mathbb{R}^3$, assumed to be oriented, is the quantity*

$$\Phi_E(S) = \int_S \langle E(x), n(x) \rangle dx$$

with $n(x)$ being unit vectors orthogonal to S , following the orientation of S . Intuitively, the flux measures the signed number of field lines crossing S .

Here by orientation of S we mean precisely the choice of unit vectors $n(x)$ as above, orthogonal to S , which must vary continuously with x . For instance a sphere has two possible orientations, one with all these vectors $n(x)$ pointing inside, and one with all these vectors $n(x)$ pointing outside. More generally, any surface has locally two possible orientations, so if it is connected, it has two possible orientations. In what follows the convention is that the closed surfaces are oriented with each $n(x)$ pointing outside.

Regarding the last sentence of Definition 8.26, this is of course something informal, meant to help, coming from the interpretation of the field lines from Proposition 8.25. However, we will see later that this simple interpretation can be of great use.

As a first illustration, let us do a basic computation, as follows:

PROPOSITION 8.27. *For a point charge $q \in \mathbb{R}$ at the center of a sphere S ,*

$$\Phi_E(S) = \frac{q}{\varepsilon_0}$$

where the constant is $\varepsilon_0 = 1/(4\pi K)$, independently of the radius of S .

PROOF. Assuming that S has radius r , we have the following computation:

$$\begin{aligned} \Phi_E(S) &= \int_S \langle E(x), n(x) \rangle dx \\ &= \int_S \left\langle \frac{Kqx}{r^3}, \frac{x}{r} \right\rangle dx \\ &= \int_S \frac{Kq}{r^2} dx \\ &= \frac{Kq}{r^2} \times 4\pi r^2 \\ &= 4\pi Kq \end{aligned}$$

Thus with $\varepsilon_0 = 1/(4\pi K)$ as above, we obtain the result. \square

More generally now, we have the following key result, due to Gauss, which is the foundation of advanced electrostatics, and of everything following from it, namely electrodynamics, and then quantum mechanics, and particle physics:

THEOREM 8.28 (Gauss law). *The flux of a field E through a surface S is given by*

$$\Phi_E(S) = \frac{Q_{enc}}{\varepsilon_0}$$

where Q_{enc} is the total charge enclosed by S , and $\varepsilon_0 = 1/(4\pi K)$.

PROOF. This basically follows from Proposition 8.27, a bit modified, by adding to the computation there a number of standard ingredients. We refer here for instance to Feynman [30], but we will be back to this right next, with a more advanced proof. \square

In relation now with our previous mathematics, we have the following result:

THEOREM 8.29. *Given an electric potential E , its divergence is given by*

$$\langle \nabla, E \rangle = \frac{\rho}{\varepsilon_0}$$

where ρ denotes as usual the charge distribution. Also, we have

$$\nabla \times E = 0$$

meaning that the curl of E vanishes.

PROOF. We have several assertions here, the idea being as follows:

(1) The first formula, called Gauss law in differential form, follows from:

$$\begin{aligned}
 \int_B \langle \nabla, E \rangle &= \int_S \langle E(x), n(x) \rangle dx \\
 &= \Phi_E(S) \\
 &= \frac{Q_{enc}}{\varepsilon_0} \\
 &= \int_B \frac{\rho}{\varepsilon_0}
 \end{aligned}$$

Now since this must hold for any B , this gives the formula in the statement.

(2) Regarding the curl, by discretizing and linearity we can assume that we are dealing with a single charge q , positioned at 0. We have, by using spherical coordinates r, s, t :

$$\begin{aligned}
 \int_a^b \langle E(x), dx \rangle &= \int_a^b \left\langle \frac{Kqx}{||x||^3}, dx \right\rangle \\
 &= \int_a^b \left\langle \frac{Kq}{r^2} \cdot \frac{x}{||x||}, dx \right\rangle \\
 &= \int_a^b \frac{Kq}{r^2} dr \\
 &= \left[-\frac{Kq}{r} \right]_a^b \\
 &= Kq \left(\frac{1}{r_a} - \frac{1}{r_b} \right)
 \end{aligned}$$

In particular the integral of E over any closed loop vanishes, and by using now the Stokes formula, we conclude that the curl of E vanishes, as stated. \square

So long for electrostatics, which provide a good motivation and illustration for our mathematics. When upgrading to electrodynamics, things become even more interesting, because our technology can be used in order to understand the Maxwell equations:

THEOREM 8.30. *Electrodynamics is governed by the formulae*

$$\langle \nabla, E \rangle = \frac{\rho}{\varepsilon_0} \quad , \quad \langle \nabla, B \rangle = 0$$

$$\nabla \times E = -\dot{B} \quad , \quad \nabla \times B = \mu_0 J + \mu_0 \varepsilon_0 \dot{E}$$

called Maxwell equations.

PROOF. This is something fundamental, coming as a tricky mixture of physics and mathematics. To be more precise, the first formula is the Gauss law, ρ being the charge, and ε_0 being a constant, and with this Gauss law more or less replacing the Coulomb law from electrostatics. The second formula is something basic, and anonymous. The third formula is the Faraday law. As for the fourth formula, this is the Ampère law, as modified by Maxwell, with J being the volume current density, and μ_0 being a constant. \square

However, the above is not all. Quite remarkably, the constants μ_0, ε_0 are related by the following formula, due to Biot-Savart, with c being the speed of light:

$$\mu_0 \varepsilon_0 = \frac{1}{c^2}$$

So, what has light to do with all this? The idea is that accelerating or decelerating charges produce electromagnetic radiation, of various wavelengths, called light, of various colors, and with all this coming from the mathematics of the Maxwell equations.

8e. Exercises

We had an exciting geometric chapter here, and as exercises on this, we have:

EXERCISE 8.31. *Clarify all the details in the classification of the conics.*

EXERCISE 8.32. *Learn about the stereographic projection, and cartography in general.*

EXERCISE 8.33. *Compute the lengths of some curves, of your choice.*

EXERCISE 8.34. *Learn more about conservative forces, and their properties.*

EXERCISE 8.35. *Compute the areas of some surfaces, of your choice.*

EXERCISE 8.36. *Clarify all the details in the proof of the Stokes formula.*

EXERCISE 8.37. *Try to compute the field lines, for simple charge configurations.*

EXERCISE 8.38. *Learn more about the Gauss law, and the Maxwell equations.*

As bonus exercise, read some systematic differential geometry, and why not some algebraic geometry too, for curves, surfaces, and other manifolds.

Part III

Function spaces

Wanted man in California
Wanted man in Buffalo
Wanted man in Kansas City
Wanted man in Ohio

CHAPTER 9

Banach spaces

9a. Normed spaces

Welcome to function space theory, also known as functional analysis. Although, for being fully honest, the basics here, which are quite algebraic, rather deserve the name “functional algebra”. But do not worry, we will keep things as analytic as possible.

The idea is that various types of functions form various types of infinite dimensional complex spaces, that we can study with the usual methods from linear algebra, complemented for dealing with ∞ by our favorite logic weapon, which is the Zorn lemma.

Let us start with something very general, as follows:

DEFINITION 9.1. *A normed space is a complex vector space V , which can be finite or infinite dimensional, together with a map*

$$||\cdot|| : V \rightarrow \mathbb{R}_+$$

called norm, subject to the following conditions:

- (1) $||x|| = 0$ implies $x = 0$.
- (2) $||\lambda x|| = |\lambda| \cdot ||x||$, for any $x \in V$, and $\lambda \in \mathbb{C}$.
- (3) $||x + y|| \leq ||x|| + ||y||$, for any $x, y \in V$.

As a basic example here, which is finite dimensional, we have the space $V = \mathbb{C}^N$, with the norm on it being the usual length of the vectors, namely:

$$||x|| = \sqrt{\sum_i |x_i|^2}$$

Indeed, for this space (1) is clear, (2) is clear too, and (3) is something well-known, which is equivalent to the triangle inequality in \mathbb{C}^N , and which can be deduced from the Cauchy-Schwarz inequality. More on this, with some generalizations, in a moment.

Getting back now to the general case, we have the following result:

PROPOSITION 9.2. *Any normed vector space V is a metric space, with*

$$d(x, y) = ||x - y||$$

as distance. If this metric space is complete, we say that V is a Banach space.

PROOF. This follows from the definition of the metric spaces, as follows:

(1) The first distance axiom, $d(x, y) \geq 0$, and $d(x, y) = 0$ precisely when $x = y$, follows from the fact that the norm takes values in \mathbb{R}_+ , and from $\|x\| = 0 \implies x = 0$.

(2) The second distance axiom, which is the symmetry one, $d(x, y) = d(y, x)$, follows from our condition $\|\lambda x\| = |\lambda| \cdot \|x\|$, with $\lambda = -1$.

(3) As for the third distance axiom, which is the triangle inequality $d(x, y) \leq d(x, z) + d(y, z)$, this follows from our third norm axiom, namely $\|x + y\| \leq \|x\| + \|y\|$. \square

Very nice all this, and it is possible to develop some general theory here, but before everything, however, we need more examples, besides \mathbb{C}^N with its usual norm.

However, these further examples are actually quite tricky to construct, needing some inequality know-how. Let us start with a very basic result, as follows:

THEOREM 9.3 (Jensen). *Given a convex function $f : \mathbb{R} \rightarrow \mathbb{R}$, we have the following inequality, for any $x_1, \dots, x_N \in \mathbb{R}$, and any $\lambda_1, \dots, \lambda_N > 0$ summing up to 1,*

$$f(\lambda_1 x_1 + \dots + \lambda_N x_N) \leq \lambda_1 f(x_1) + \dots + \lambda_N f(x_N)$$

with equality when $x_1 = \dots = x_N$. In particular, by taking the weights λ_i to be all equal, we obtain the following inequality, valid for any $x_1, \dots, x_N \in \mathbb{R}$,

$$f\left(\frac{x_1 + \dots + x_N}{N}\right) \leq \frac{f(x_1) + \dots + f(x_N)}{N}$$

and once again with equality when $x_1 = \dots = x_N$. We have a similar statement holds for the concave functions, with all the inequalities being reversed.

PROOF. This is indeed something quite routine, the idea being as follows:

(1) First, we can talk about convex functions in a usual, intuitive way, with this meaning by definition that the following inequality must be satisfied:

$$f\left(\frac{x+y}{2}\right) \leq \frac{f(x) + f(y)}{2}$$

(2) But this means, via a simple argument, by approximating numbers $t \in [0, 1]$ by sums of powers 2^{-k} , that for any $t \in [0, 1]$ we must have:

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y)$$

Alternatively, via yet another simple argument, this time by doing some geometry with triangles, this means that we must have:

$$f\left(\frac{x_1 + \dots + x_N}{N}\right) \leq \frac{f(x_1) + \dots + f(x_N)}{N}$$

But then, again alternatively, by combining the above two simple arguments, the following must happen, for any $\lambda_1, \dots, \lambda_N > 0$ summing up to 1:

$$f(\lambda_1 x_1 + \dots + \lambda_N x_N) \leq \lambda_1 f(x_1) + \dots + \lambda_N f(x_N)$$

(3) Summarizing, all our Jensen inequalities, at $N = 2$ and at $N \in \mathbb{N}$ arbitrary, are equivalent. The point now is that, if we look at what the first Jensen inequality, that we took as definition for the convexity, means, this is simply equivalent to:

$$f''(x) \geq 0$$

(4) Thus, we are led to the conclusions in the statement, regarding the convex functions. As for the concave functions, the proof here is similar. Alternatively, we can say that f is concave precisely when $-f$ is convex, and get the results from what we have. \square

As a basic application of the Jensen inequality, we have:

PROPOSITION 9.4. *For $p \in (1, \infty)$ we have the following inequality,*

$$\left| \frac{x_1 + \dots + x_N}{N} \right|^p \leq \frac{|x_1|^p + \dots + |x_N|^p}{N}$$

and for $p \in (0, 1)$ we have the following reverse inequality,

$$\left| \frac{x_1 + \dots + x_N}{N} \right|^p \geq \frac{|x_1|^p + \dots + |x_N|^p}{N}$$

with in both cases equality precisely when $|x_1| = \dots = |x_N|$.

PROOF. This follows indeed from Theorem 9.3, because we have:

$$(x^p)'' = p(p-1)x^{p-2}$$

Thus x^p is convex for $p > 1$ and concave for $p < 1$, which gives the results. \square

Observe that, more generally, Theorem 9.3 applied with $f(x) = x^p$ with $p \in (1, \infty)$ shows that we have the following inequality, whenever $\lambda_i > 0$ sum up to 1:

$$(\lambda_1 x_1 + \dots + \lambda_N x_N)^p \leq \lambda_1 x_1^p + \dots + \lambda_N x_N^p$$

For exponents $p \in (0, 1)$ we have the same inequality, reversed, namely:

$$(\lambda_1 x_1 + \dots + \lambda_N x_N)^p \geq \lambda_1 x_1^p + \dots + \lambda_N x_N^p$$

Observe also that at $p = 2$ we obtain as particular case of the above inequality the Cauchy-Schwarz inequality, or rather something equivalent to it, namely:

$$\left(\frac{x_1 + \dots + x_N}{N} \right)^2 \leq \frac{x_1^2 + \dots + x_N^2}{N}$$

We will be back to this later on in this book, when talking scalars products and Hilbert spaces, with some more conceptual proofs for such inequalities.

As another basic application of the Jensen inequality, we have:

THEOREM 9.5 (Young). *We have the following inequality,*

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q}$$

valid for any $a, b \geq 0$, and any exponents $p, q > 1$ satisfying $\frac{1}{p} + \frac{1}{q} = 1$.

PROOF. We use the logarithm function, which is concave on $(0, \infty)$, due to:

$$(\log x)'' = \left(-\frac{1}{x}\right)' = -\frac{1}{x^2}$$

Thus we can apply the Jensen inequality, and we obtain in this way:

$$\begin{aligned} \log \left(\frac{a^p}{p} + \frac{b^q}{q} \right) &\geq \frac{\log(a^p)}{p} + \frac{\log(b^q)}{q} \\ &= \log(a) + \log(b) \\ &= \log(ab) \end{aligned}$$

Now by exponentiating, we obtain the Young inequality. \square

Observe that for the simplest exponents, namely $p = q = 2$, the Young inequality gives something which is trivial, but is very useful and basic, namely:

$$ab \leq \frac{a^2 + b^2}{2}$$

In general, the Young inequality is something non-trivial, and the idea with it is that “when stuck with a problem, and with $ab \leq \frac{a^2+b^2}{2}$ not working, try Young”. We will be back to this general principle, later in this book, with some illustrations.

Moving forward now, as a consequence of the Young inequality, we have:

THEOREM 9.6 (Hölder). *Assuming that $p, q \geq 1$ are conjugate, in the sense that*

$$\frac{1}{p} + \frac{1}{q} = 1$$

we have the following inequality, valid for any two vectors $x, y \in \mathbb{C}^N$,

$$\sum_i |x_i y_i| \leq \left(\sum_i |x_i|^p \right)^{1/p} \left(\sum_i |y_i|^q \right)^{1/q}$$

with the convention that an ∞ exponent produces a $\max |x_i|$ quantity.

PROOF. This is something very standard, the idea being as follows:

(1) Assume first that we are dealing with finite exponents, $p, q \in (1, \infty)$. By linearity we can assume that x, y are normalized, in the following way:

$$\sum_i |x_i|^p = \sum_i |y_i|^q = 1$$

In this case, we want to prove that the following inequality holds:

$$\sum_i |x_i y_i| \leq 1$$

For this purpose, we use the Young inequality, which gives, for any i :

$$|x_i y_i| \leq \frac{|x_i|^p}{p} + \frac{|y_i|^q}{q}$$

By summing now over $i = 1, \dots, N$, we obtain from this, as desired:

$$\begin{aligned} \sum_i |x_i y_i| &\leq \sum_i \frac{|x_i|^p}{p} + \sum_i \frac{|y_i|^q}{q} \\ &= \frac{1}{p} + \frac{1}{q} \\ &= 1 \end{aligned}$$

(2) In the case $p = 1$ and $q = \infty$, or vice versa, the inequality holds too, trivially, with the convention that an ∞ exponent produces a max quantity, according to:

$$\lim_{p \rightarrow \infty} \left(\sum_i |x_i|^p \right)^{1/p} = \max |x_i|$$

Thus, we are led to the conclusion in the statement. □

As a consequence now of the Hölder inequality, we have:

THEOREM 9.7 (Minkowski). *Assuming $p \in [1, \infty]$, we have the inequality*

$$\left(\sum_i |x_i + y_i|^p \right)^{1/p} \leq \left(\sum_i |x_i|^p \right)^{1/p} + \left(\sum_i |y_i|^p \right)^{1/p}$$

for any two vectors $x, y \in \mathbb{C}^N$, with our usual conventions at $p = \infty$.

PROOF. We have indeed the following estimate, using the Hölder inequality, and the conjugate exponent $q \in [1, \infty]$, given by $1/p + 1/q = 1$:

$$\begin{aligned}
\sum_i |x_i + y_i|^p &= \sum_i |x_i + y_i| \cdot |x_i + y_i|^{p-1} \\
&\leq \sum_i |x_i| \cdot |x_i + y_i|^{p-1} + \sum_i |y_i| \cdot |x_i + y_i|^{p-1} \\
&\leq \left(\sum_i |x_i|^p \right)^{1/p} \left(\sum_i |x_i + y_i|^{(p-1)q} \right)^{1/q} \\
&\quad + \left(\sum_i |y_i|^p \right)^{1/p} \left(\sum_i |x_i + y_i|^{(p-1)q} \right)^{1/q} \\
&= \left[\left(\sum_i |x_i|^p \right)^{1/p} + \left(\sum_i |y_i|^p \right)^{1/p} \right] \left(\sum_i |x_i + y_i|^p \right)^{1-1/p}
\end{aligned}$$

Here we have used the following fact, at the end:

$$\frac{1}{p} + \frac{1}{q} = 1 \implies \frac{1}{q} = \frac{p-1}{p} \implies (p-1)q = p$$

Now by dividing both sides by the last quantity at the end, we obtain:

$$\left(\sum_i |x_i + y_i|^p \right)^{1/p} \leq \left(\sum_i |x_i|^p \right)^{1/p} + \left(\sum_i |y_i|^p \right)^{1/p}$$

Thus, we are led to the conclusion in the statement. \square

Good news, done with inequalities, and as a consequence of the above results, and more specifically of the Minkowski inequality obtained above, we can formulate:

THEOREM 9.8. *Given an exponent $p \in [1, \infty]$, the formula*

$$\|x\|_p = \left(\sum_i |x_i|^p \right)^{1/p}$$

with usual conventions at $p = \infty$, defines a norm on \mathbb{C}^N , making it a Banach space.

PROOF. Here the normed space assertion follows from the Minkowski inequality, established above, and the Banach space assertion is trivial, because our space being finite dimensional, by standard linear algebra all the Cauchy sequences converge. \square

We will see later in this chapter that Theorem 9.8 has several generalizations, with some of them dealing with functions. Before that, however, you might have another question regarding Theorem 9.8, namely, what is the underlying geometry there.

So, let us discuss this latter question. We will actually see that, with a bit of geometric know-how, not only we will reach to a better understanding of the p -norm on \mathbb{C}^N , but we will also be able to recover the Hölder inequality for it, without any computation.

It is convenient to talk about \mathbb{R}^N instead of \mathbb{C}^N . Let us start with:

PROPOSITION 9.9. *For any exponent $p > 1$, the following set*

$$S_p = \left\{ x \in \mathbb{R}^N \mid \sum_i |x_i|^p = 1 \right\}$$

is a smooth submanifold of \mathbb{R}^N .

PROOF. We know from chapter 8 that the unit sphere in \mathbb{R}^N is a smooth manifold. In our terms, this solves our problem at $p = 2$, because this unit sphere is:

$$S_2 = \left\{ x \in \mathbb{R}^N \mid \sum_i x_i^2 = 1 \right\}$$

Now observe that we have a bijection $S_p \simeq S_2$, at least on the part where all the coordinates are positive, $x_i > 0$, given by the following function:

$$x_i \rightarrow x_i^{2/p}$$

Thus we obtain that S_p is indeed a manifold, as claimed. \square

We already know that the manifold S_p constructed above is the unit sphere, in the case $p = 2$. In order to have a better geometric picture of what is going on, in general, observe that S_p can be constructed as well at $p = 1$, as follows:

$$S_1 = \left\{ x \in \mathbb{R}^N \mid \sum_i |x_i| = 1 \right\}$$

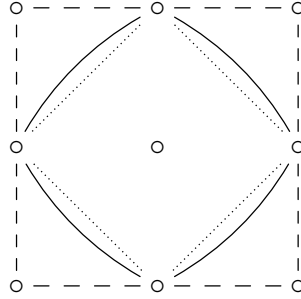
However, this is no longer a manifold, as we can see for instance at $N = 2$, where we obtain a square. Now observe that we can talk as well about $p = \infty$, as follows:

$$S_\infty = \left\{ x \in \mathbb{R}^N \mid \sup_i |x_i| = 1 \right\}$$

This latter set is no longer a manifold either, as we can see for instance at $N = 2$, where we obtain again a square, containing the previous square, the one at $p = 1$.

With these limiting constructions in hand, we can now have a better geometric picture of what is going on, in the general context of Proposition 9.9. Indeed, let us draw, at

$N = 2$ for simplifying, our sets S_p at the values $p = 1, 2, \infty$ of the exponent:



We can see that what we have is a small square, at $p = 1$, becoming smooth and inflating towards the circle, in the parameter range $p \in (1, 2]$, and then further inflating, in the parameter range $p \in [2, \infty)$, towards the big square appearing at $p = \infty$.

Now let us get to our second objective, proving Hölder, without computations. We will use Lagrange multipliers, whose basic theory can be summarized as follows:

FACT 9.10. *In order for a function $f : X \rightarrow \mathbb{R}$ defined on a manifold X to have a local extremum at $x \in X$, we must have, as usual*

$$f'(x) = 0$$

but with this taking into account the fact that the equations defining the manifold count as well as “zero”, and so must be incorporated into the formula $f'(x) = 0$.

To be more precise, we must have a formula as follows, with g_i being the constraint functions for X , and with $\lambda_i \in \mathbb{R}$ being certain scalars, called Lagrange multipliers:

$$f'(x) = \sum_i \lambda_i g'_i(x)$$

With these preliminaries in hand, we can formulate our result, as follows:

THEOREM 9.11. *The local extrema over S_p of the function*

$$f(x) = \sum_i x_i y_i$$

can be computed by using Lagrange multipliers, and this gives

$$\left| \sum_i x_i y_i \right| \leq \left(\sum_i |x_i|^p \right)^{1/p} \left(\sum_i |y_i|^q \right)^{1/q}$$

with $1/p + 1/q = 1$, that is, the Hölder inequality, with a purely geometric proof.

PROOF. We can restrict the attention to the case where all the coordinates are positive, $x_i > 0$ and $y_i > 0$. The derivative of the function in the statement is:

$$f'(x) = (y_1, \dots, y_N)$$

On the other hand, we know that the manifold S_p appears by definition as the set of zeroes of the function $\varphi(x) = \sum_i x_i^p - 1$, having derivative as follows:

$$\varphi'(x) = p(x_1^{p-1}, \dots, x_N^{p-1})$$

Thus, by using Lagrange multipliers, the critical points of f must satisfy:

$$(y_1, \dots, y_N) \sim (x_1^{p-1}, \dots, x_N^{p-1})$$

In other words, the critical points must satisfy $x_i = \lambda y_i^{1/(p-1)}$, for some $\lambda > 0$, and by using now $\sum_i x_i^p = 1$ we can compute the precise value of λ , and we get:

$$\lambda = \left(\sum_i y_i^{p/(p-1)} \right)^{-1/p}$$

Now let us see what this means. Since the critical point is unique, this must be a maximum of our function, and we conclude that for any $x \in S_p$, we have:

$$\begin{aligned} \sum_i x_i y_i &\leq \sum_i \lambda y_i^{1/(p-1)} \cdot y_i \\ &= \left(\sum_i y_i^{p/(p-1)} \right)^{1-1/p} \\ &= \left(\sum_i y_i^q \right)^{1/q} \end{aligned}$$

Thus we have Hölder, and the general case follows from this, by rescaling. \square

9b. Banach spaces

Very nice all the above, but you might wonder at this point, what is the relation of all this with functions. In answer, Theorem 9.8 can be reformulated as follows:

THEOREM 9.12. *Given an exponent $p \in [1, \infty]$, the formula*

$$\|f\|_p = \left(\int |f(x)|^p \right)^{1/p}$$

with usual conventions at $p = \infty$, defines a norm on the space of functions

$$f : \{1, \dots, N\} \rightarrow \mathbb{C}$$

making it a Banach space.

PROOF. This is a just fancy reformulation of Theorem 9.8, by using the fact that the space formed by the functions $f : \{1, \dots, N\} \rightarrow \mathbb{C}$ is canonically isomorphic to \mathbb{C}^N , in the obvious way, and by replacing the sums from the \mathbb{C}^N context with integrals with respect to the counting measure on $\{1, \dots, N\}$, in the function context. \square

Moving now towards infinite dimensions and more standard analysis, the idea will be that of extending Theorem 9.12 to the arbitrary measured spaces. Let us start with:

THEOREM 9.13. *Given an exponent $p \in [1, \infty]$, the formula*

$$\|x\|_p = \left(\sum_i |x_i|^p \right)^{1/p}$$

with usual conventions at $p = \infty$, defines a norm on the space of sequences

$$l^p = \left\{ (x_i)_{i \in \mathbb{N}} \mid \sum_i |x_i|^p < \infty \right\}$$

making it a Banach space.

PROOF. As before with the finite sequences, the normed space assertion follows from the Minkowski inequality, established above, which extends without problems to the case of the infinite sequences, and with the Banach space assertion being clear too. \square

We can unify and generalize what we have, in the following way:

THEOREM 9.14. *Given a discrete measured space X , and an exponent $p \in [1, \infty]$,*

$$\|f\|_p = \left(\int_X |f(x)|^p \right)^{1/p}$$

with usual conventions at $p = \infty$, defines a norm on the space of functions

$$l^p(X) = \left\{ f : X \rightarrow \mathbb{C} \mid \int_X |f(x)|^p < \infty \right\}$$

making it a Banach space.

PROOF. This is just a fancy reformulation of what we have:

(1) The case where X is finite corresponds to Theorem 9.12.

(2) The case where X is countable corresponds to Theorem 9.13.

(3) Finally, the case where X is uncountable is easy to deal with too, by using the same arguments as in the countable case. \square

Before going further, let us mention that the l^p spaces above are not the only interesting Banach spaces of sequences. Indeed, we have as well the following result:

THEOREM 9.15. *The following space of complex sequences,*

$$c_0 = \left\{ (x_i)_{i \in \mathbb{N}} \mid x_i \rightarrow 0 \right\}$$

is a Banach space, with the sup norm, and we have $c_0 \subset l^\infty$, proper inclusion.

PROOF. We can define indeed c_0 as above, as being the space of sequences $x_i \in \mathbb{C}$ satisfying $x_i \rightarrow 0$, with the sup norm. We have then an inclusion as follows:

$$c_0 \subset l^\infty$$

It is then clear that c_0 is closed, so it is indeed a Banach space, as claimed. \square

Now back to the l^p spaces, in order to further extend the above results, to the case of the arbitrary measured spaces X , which are not necessarily discrete, let us start with:

THEOREM 9.16. *Given two functions $f, g : X \rightarrow \mathbb{C}$ and an exponent $p \geq 1$, we have*

$$\left(\int_X |f + g|^p \right)^{1/p} \leq \left(\int_X |f|^p \right)^{1/p} + \left(\int_X |g|^p \right)^{1/p}$$

called Minkowski inequality. Also, assuming that $p, q \geq 1$ satisfy $1/p + 1/q = 1$, we have

$$\int_X |fg| \leq \left(\int_X |f|^p \right)^{1/p} \left(\int_X |g|^q \right)^{1/q}$$

called Hölder inequality. These inequalities hold as well for ∞ values of the exponents.

PROOF. This is very standard, exactly as in the case of sequences, finite or not, but since the above inequalities are really very general and final, here are the details:

(1) Let us first prove Hölder, in the case of finite exponents, $p, q \in (1, \infty)$. By linearity we can assume that f, g are normalized, in the following way:

$$\int_X |f|^p = \int_X |g|^q = 1$$

In this case, we want to prove that the following inequality holds:

$$\int_X |fg| \leq 1$$

For this purpose, we use the Young inequality, which gives, for any $x \in X$:

$$|f(x)g(x)| \leq \frac{|f(x)|^p}{p} + \frac{|g(x)|^q}{q}$$

By integrating now over $x \in X$, we obtain from this, as desired:

$$\begin{aligned} \int_X |fg| &\leq \int_X \frac{|f(x)|^p}{p} + \int_X \frac{|g(x)|^q}{q} \\ &= \frac{1}{p} + \frac{1}{q} \\ &= 1 \end{aligned}$$

(2) Let us prove now Minkowski, again in the finite exponent case, $p \in (1, \infty)$. We have the following estimate, using the Hölder inequality, and the conjugate exponent:

$$\begin{aligned} \int_X |f+g|^p &= \int_X |f+g| \cdot |f+g|^{p-1} \\ &\leq \int_X |f| \cdot |f+g|^{p-1} + \int_X |g| \cdot |f+g|^{p-1} \\ &\leq \left(\int_X |f|^p \right)^{1/p} \left(\int_X |f+g|^{(p-1)q} \right)^{1/q} \\ &\quad + \left(\int_X |g|^p \right)^{1/p} \left(\int_X |f+g|^{(p-1)q} \right)^{1/q} \\ &= \left[\left(\int_X |f|^p \right)^{1/p} + \left(\int_X |g|^p \right)^{1/p} \right] \left(\int_X |f+g|^p \right)^{1-1/p} \end{aligned}$$

Thus, we are led to the Minkowski inequality in the statement.

(3) Finally, in the infinite exponent cases we have similar results, with the convention that an ∞ exponent produces an essential supremum, according to:

$$\lim_{p \rightarrow \infty} \left(\int_X |f|^p \right)^{1/p} = \text{ess sup } |f|$$

Thus, we are led to the conclusion in the statement. □

We can now extend Theorem 9.14, into something very general, as follows:

THEOREM 9.17. *Given a measured space X , and $p \in [1, \infty]$, the following space, with the convention that functions are identified up to equality almost everywhere,*

$$L^p(X) = \left\{ f : X \rightarrow \mathbb{C} \mid \int_I |f(x)|^p dx < \infty \right\}$$

is a vector space, and the following quantity

$$\|f\|_p = \left(\int_X |f(x)|^p \right)^{1/p}$$

is a norm on it, making it a Banach space.

PROOF. This follows indeed from Theorem 9.16, with due attention to the null sets, and this because of the first normed space axiom, namely:

$$||x|| = 0 \implies x = 0$$

To be more precise, in order for this axiom to hold, we must identify the functions up to equality almost everywhere, as indicated in the statement. \square

Summarizing, very nice all this. So, we have our examples of Banach spaces, which look definitely interesting, and related to analysis. In the remainder of this chapter we will develop some general Banach space theory, and apply it to the above L^p spaces.

9c. Linear forms

Getting now to work, as a first result about the abstract normed spaces, we would like to talk about the linear maps $T : V \rightarrow W$, which can be thought of as being some kind of infinite matrices, when V, W are infinite dimensional. We first have here:

PROPOSITION 9.18. *For a linear map $T : V \rightarrow W$, the following conditions are equivalent, and if they hold, we say that T is bounded:*

- (1) *T is continuous.*
- (2) *T is continuous at 0.*
- (3) *T maps the unit ball of V into something bounded.*
- (4) *T is bounded, in the sense that $||T|| = \sup_{||x||=1} ||Tx||$ is finite.*

PROOF. This is something elementary, the idea being as follows:

(1) \iff (2) This is indeed clear from definitions, by using the linearity of T , and performing various rescalings.

(2) \iff (3) This is clear from definitions too, again by using the linearity of T , and performing various rescalings.

(3) \iff (4) This is clear from definitions too, with the number $||T||$ needed in (4) being the bound coming from (3). \square

With the above result in hand, we can now formulate:

THEOREM 9.19. *Given two Banach spaces V, W , the bounded linear maps*

$$T : V \rightarrow W$$

form a linear space $B(V, W)$, on which the following quantity is a norm,

$$||T|| = \sup_{||x||=1} ||Tx||$$

making $B(V, W)$ a Banach space. When $V = W$, we obtain a Banach algebra.

PROOF. All this is very standard, and in the case $V = W$, the proof goes as follows:

(1) The fact that we have indeed an algebra, satisfying the product condition in the statement, follows from the following estimates, which are all elementary:

$$\|S + T\| \leq \|S\| + \|T\|$$

$$\|\lambda T\| = |\lambda| \cdot \|T\|$$

$$\|ST\| \leq \|S\| \cdot \|T\|$$

(2) Regarding now the last assertion, if $\{T_n\} \subset B(V)$ is Cauchy then $\{T_n x\}$ is Cauchy for any $x \in V$, so we can define the limit $T = \lim_{n \rightarrow \infty} T_n$ by setting:

$$Tx = \lim_{n \rightarrow \infty} T_n x$$

Let us first check that the application $x \rightarrow Tx$ is linear. We have:

$$\begin{aligned} T(x + y) &= \lim_{n \rightarrow \infty} T_n(x + y) \\ &= \lim_{n \rightarrow \infty} T_n(x) + T_n(y) \\ &= \lim_{n \rightarrow \infty} T_n(x) + \lim_{n \rightarrow \infty} T_n(y) \\ &= T(x) + T(y) \end{aligned}$$

Similarly, we have as well the following computation:

$$\begin{aligned} T(\lambda x) &= \lim_{n \rightarrow \infty} T_n(\lambda x) \\ &= \lambda \lim_{n \rightarrow \infty} T_n(x) \\ &= \lambda T(x) \end{aligned}$$

Thus we have a linear map $T : A \rightarrow A$. It remains to prove that we have $T \in B(V)$, and that we have $T_n \rightarrow T$ in norm. For this purpose, observe that we have:

$$\begin{aligned} &\|T_n - T_m\| \leq \varepsilon, \quad \forall n, m \geq N \\ \implies &\|T_n x - T_m x\| \leq \varepsilon, \quad \forall \|x\| = 1, \quad \forall n, m \geq N \\ \implies &\|T_n x - T x\| \leq \varepsilon, \quad \forall \|x\| = 1, \quad \forall n \geq N \\ \implies &\|T_N x - T x\| \leq \varepsilon, \quad \forall \|x\| = 1 \\ \implies &\|T_N - T\| \leq \varepsilon \end{aligned}$$

As a first consequence, we obtain $T \in B(V)$, because we have:

$$\begin{aligned} \|T\| &= \|T_N + (T - T_N)\| \\ &\leq \|T_N\| + \|T - T_N\| \\ &\leq \|T_N\| + \varepsilon \\ &< \infty \end{aligned}$$

As a second consequence, we obtain $T_N \rightarrow T$ in norm, and we are done. \square

As a basic example for the above construction, in the case where both our spaces are finite dimensional, $V = \mathbb{C}^N$ and $W = \mathbb{C}^M$, with $N, M < \infty$, we obtain a matrix space:

$$B(\mathbb{C}^N, \mathbb{C}^M) = M_{M \times N}(\mathbb{C})$$

More on this later. On the other hand, of particular interest is as well the case $W = \mathbb{C}$ of the above construction, which leads to the following result:

THEOREM 9.20. *Given a Banach space V , its dual space, constructed as*

$$V^* = \left\{ f : V \rightarrow \mathbb{C}, \text{ linear and bounded} \right\}$$

is a Banach space too, with norm given by:

$$\|f\| = \sup_{\|x\|=1} |f(x)|$$

When V is finite dimensional, we have $V \simeq V^$.*

PROOF. This is clear indeed from Theorem 9.19, because we have:

$$V^* = B(V, \mathbb{C})$$

Thus, we are led to the conclusions in the statement. □

In order to better understand the linear forms, we will need:

THEOREM 9.21 (Hahn-Banach). *Given a Banach space V , the following happen:*

- (1) *Given $x \in V - \{0\}$, there exists $f \in V^*$ with $f(x) \neq 0$.*
- (2) *Given a subspace $W \subset V$, any $f \in W^*$ extends into a $\tilde{f} \in V^*$, of same norm.*

PROOF. This is something quite tricky, the idea being as follows:

(1) As a first observation, the assertion (1) is a particular case of the assertion (2). Indeed, given a nonzero vector $x \in V - \{0\}$, we can consider the linear space $W = \mathbb{C}x$, and construct a linear form $f \in W^*$ by the following formula:

$$f(\lambda x) = \lambda$$

But with this done, any extension $\tilde{f} \in V^*$ of this linear form $f \in W^*$, as provided by (2), does the job for (1), and this because we have $\tilde{f}(x) = f(x) = 1$.

(2) As a comment now, you might wonder why keeping both (1,2) in the statement of the theorem. In answer, the Hahn-Banach theorem is (2), while (1) is the main consequence of this theorem, which is something very useful, in practice. So, with our theorem formulated as above, we have there all we need to know, about Hahn-Banach.

(3) As a second remark now, assuming that we are in finite dimensions, given a subspace $W \subset V$ as in (2), we can always find a direct sum decomposition, as follows:

$$V = W \oplus U$$

But once we have such a decomposition, given a linear form $f \in W^*$ as in (2), we can define an extension $\tilde{f} \in V^*$ by the following formula:

$$\tilde{f}(w, u) = f(w)$$

Thus we have proved a weak form of (2), without reference to the norm of the extension, which is still stronger than (1), in finite dimensions.

(4) Still staying in finite dimensions, it is possible to further build on this, as to have the equality $\|\tilde{f}\| = \|f\|$ happening too, as required by (2). For instance in the case of the space $V = \mathbb{C}^N$, with its usual norm, we can use the following decomposition, and we will leave the proof of the equality $\|\tilde{f}\| = \|f\|$ in this case as an easy exercise:

$$V = W \oplus W^\perp$$

However, in the general finite dimensional case, that of the space $V = \mathbb{C}^N$, but with an arbitrary norm, this simple, geometric argument is no longer valid. We will be back to this in a moment, with explanations on how to do this, directly in a more general setting, that where the codimension of $W \subset V$ is 1, or more generally, is finite.

(5) Getting now to the general case, where V can be infinite dimensional, we are facing two problems here, which can be summarized as follows:

– First, there is a problem coming from infinite dimensionality. Indeed, inspired by the above, we can try to extend linear forms by adding 1, 2, 3, ... dimensions to our space. But if we want to add ∞ dimensions, we will probably need the Zorn lemma.

– Second, at the technical level, when trying to compute norms, and arrange as for our extensions to satisfy $\|\tilde{f}\| = \|f\|$, we have a problem with real vs complex. Indeed, and you will have to believe me here, the real problem is easier than the complex one.

In short, we are in need of a good plan here. In what follows we will first focus on the real problem, and solve it by adding 1 dimension, and then invoking the Zorn lemma. And then, we will use a trick, in order to convert our real result into a complex result.

(6) So, consider a real Banach space V , having a subspace $W \subset V$, and assume in addition that this subspace is of codimension one, in the sense that we have:

$$V = W \oplus \mathbb{R}x$$

Our claim is that any real linear form $f : W \rightarrow \mathbb{R}$ can be extended into a real linear form $f : V \rightarrow \mathbb{R}$, having the same norm. In order to prove this, let us set:

$$\tilde{f}(w + \lambda x) = f(w) + \lambda \alpha$$

This linear form $\tilde{f} : W \rightarrow \mathbb{R}$ extends then $f : V \rightarrow \mathbb{R}$, and the problem is that of finding the correct scalar $\alpha \in \mathbb{R}$, making the equality $\|\tilde{f}\| = \|f\|$ hold.

(7) In order to solve this latter question, we can assume, by linearity, that we have $\|f\| = 1$. In this case, we want to find $\alpha \in \mathbb{R}$ such that the following holds:

$$|f(w) + \lambda\alpha| \leq \|w + \lambda x\|$$

By replacing $w \rightarrow -\lambda w$, we would like the following inequality to hold:

$$|-\lambda f(w) + \lambda\alpha| \leq \|-\lambda w + \lambda x\|$$

Equivalently, by dividing by λ , we would like the following inequality to hold:

$$|f(w) - \alpha| \leq \|w - x\|$$

But this is the same as asking for the following inequality to hold:

$$-\|w - x\| \leq f(w) - \alpha \leq \|w - x\|$$

Thus, our scalar $\alpha \in \mathbb{R}$ must be subject to the following inequalities:

$$f(w) - \|w - x\| \leq \alpha \leq f(w) + \|w - x\|$$

(8) To summarize, we can find indeed our scalar $\alpha \in \mathbb{R}$, making our extension \tilde{f} to have norm 1 as desired, provided that the following inequality holds:

$$\sup_{w \in W} f(w) - \|w - x\| \leq \inf_{w \in W} f(w) + \|w - x\|$$

But, in order to establish this latter inequality, observe that our assumption $\|f\| = 1$, and then the triangle inequality give, for any two vectors $v, w \in W$:

$$\begin{aligned} f(v) - f(w) &= f(v - w) \\ &\leq \|v - w\| \\ &\leq \|v - x\| + \|w - x\| \end{aligned}$$

The point now is that this estimate can be written in the following way:

$$f(v) - \|v - x\| \leq f(w) + \|w - x\|$$

Thus we can indeed pick $\alpha \in \mathbb{R}$ to be between the quantities on the left, and those on the right, with $v, w \in W$ varying, and with this choice, we have, as needed:

$$\|\tilde{f}\| = \|f\|$$

(9) Next, with this done, we can now prove Hahn-Banach in general, in the real case, by invoking the Zorn lemma. Assume indeed that we have a real Banach space V , a real subspace $W \subset V$, and a real linear form $f : W \rightarrow \mathbb{R}$. We can then consider the family of extensions $\tilde{f} : \tilde{W} \rightarrow \mathbb{R}$ of our linear form f , having the same norm as it, $\|\tilde{f}\| = \|f\|$. This family is then partially ordered, in the obvious way, and by using the Zorn lemma, we conclude that it must have a maximal element. But, in view of what we found in (8), this maximal element $\tilde{f} : \tilde{W} \rightarrow \mathbb{R}$ must be such that $\tilde{W} = V$, and we are done.

(10) Finally, getting back now to the complex case, assume that we have a complex Banach space V , a complex subspace $W \subset V$, and a linear form $f \in W^*$. We can then apply the real result, established above, to the following real linear form:

$$g = \operatorname{Re}(f) : W \rightarrow \mathbb{R}$$

We obtain in this way a certain extension $\tilde{g} : V \rightarrow \mathbb{R}$, and then we can set:

$$\tilde{f}(x) = \tilde{g}(x) - i\tilde{g}(ix)$$

And with this, our claim is that we are done, with this being the needed extension.

(11) Indeed, as a first observation, $\tilde{f} : V \rightarrow \mathbb{C}$ as constructed above is indeed a linear form, in the complex sense, with this coming from the following computation:

$$\begin{aligned} \tilde{f}((a + ib)x) &= \tilde{g}((a + ib)x) - i\tilde{g}(i(a + ib)x) \\ &= \tilde{g}((a + ib)x) - i\tilde{g}((ia - b)x) \\ &= a\tilde{g}(x) + b\tilde{g}(ix) - i(a\tilde{g}(ix) - b\tilde{g}(x)) \\ &= a(\tilde{g}(x) - i\tilde{g}(ix)) + ib(\tilde{g}(x) - i\tilde{g}(ix)) \\ &= (a + ib)(\tilde{g}(x) - i\tilde{g}(ix)) \\ &= (a + ib)\tilde{f}(x) \end{aligned}$$

(12) Next, since we have $z = \operatorname{Re}(z) - i\operatorname{Re}(iz)$ for any complex number $z \in \mathbb{C}$, by linearity we obtain that $\tilde{f} \in V^*$ extends indeed the original linear form $f \in W^*$. As for the same norm assertion, this follows from the following equalities:

$$\|\tilde{f}\| = \|\tilde{g}\| = \|g\| = \|f\|$$

Thus, theorem proved, in its general complex form, the one in the statement. □

9d. Duality results

With the above understood, we can now formulate a key result, as follows:

THEOREM 9.22. *Given a Banach space V , we have an embedding as follows,*

$$V \subset V^{**}$$

which is an isomorphism in finite dimensions, and for the l^p and L^p spaces too.

PROOF. There are several things going on here, the idea being as follows:

(1) The fact that we have indeed a vector space embedding $V \subset V^{**}$ is clear from definitions, the formula of this embedding being as follows:

$$i(v)[f] = f(v)$$

(2) However, the fact that this embedding $V \subset V^{**}$ is isometric is something more subtle, which requires the use of the Hahn-Banach result from Theorem 9.21.

(3) Next, the fact that we have $V = V^{**}$ in finite dimensions is clear.

(4) Regarding now the formula $V = V^{**}$ for the various l^p and L^p spaces, this is something quite tricky. Let us start with the simplest case, that of the space $V = l^2$. We know that this space is given by definition by the following formula:

$$l^2 = \left\{ (x_i)_{i \in \mathbb{N}} \mid \sum_i x_i^2 < \infty \right\}$$

Now let us look for linear forms $f : l^2 \rightarrow \mathbb{C}$. By linearity such a linear form must appear as follows, for certain scalars $a_i \in \mathbb{C}$, which must be such that f is well-defined:

$$f((x_i)_i) = \sum_i a_i x_i$$

But, what does the fact that f is well-defined mean? In answer, this means that the values of f must all converge, which in practice means that we must have:

$$\sum_i x_i^2 < \infty \implies \left| \sum_i a_i x_i \right| < \infty$$

Moreover, we would like our linear form $f : l^2 \rightarrow \mathbb{C}$ to be bounded, and by denoting by $A = \|f\| < \infty$ the minimal bound, this means that we must have:

$$\left| \sum_i a_i x_i \right| \leq A \sqrt{\sum_i x_i^2}$$

Now recall that the Cauchy-Schwarz inequality tells us that we have:

$$\left| \sum_i a_i x_i \right| \leq \sqrt{\sum_i a_i^2} \cdot \sqrt{\sum_i x_i^2}$$

Thus, the linear form $f : l^2 \rightarrow \mathbb{C}$ associated to any $a = (a_i) \in l^2$ will do. Moreover, conversely, by examining the proof of Cauchy-Schwarz, we conclude that this condition $a = (a_i) \in l^2$ is in fact necessary. Thus, we have proved that we have:

$$(l^2)^* = l^2$$

But this gives the $V = V^{**}$ result in the statement for our space $V = l^2$, because by dualizing one more time we obtain, as desired:

$$(l^2)^{**} = (l^2)^* = l^2$$

(5) Getting now to more complicated spaces, let us look, more generally, at $L^2(X)$. We know that this space is given by definition by the following formula:

$$L^2(X) = \left\{ f : X \rightarrow \mathbb{C} \mid \int_X f(x)^2 dx < \infty \right\}$$

As before, when looking for linear forms $\varphi : L^2(X) \rightarrow \mathbb{C}$, by linearity, and with some measure theory helping, our forms must appear via a formula as follows:

$$\varphi(f) = \int_X f(x)\varphi(x)dx$$

Now in order for this integral to converge, as for our map $\varphi : L^2(X) \rightarrow \mathbb{C}$ to be well-defined, and with the additional requirement that φ must be actually bounded, we must have an inequality as follows, for a certain positive constant $A < \infty$:

$$\left| \int_X f(x)\varphi(x)dx \right| \leq A \sqrt{\int_X f(x)^2 dx}$$

Now recall that the Cauchy-Schwarz inequality tells us that we have:

$$\left| \int_X f(x)\varphi(x)dx \right| \leq \sqrt{\int_X \varphi(x)^2 dx} \cdot \sqrt{\int_X f(x)^2 dx}$$

Thus, the linear form $\varphi : L^2(X) \rightarrow \mathbb{C}$ associated to any $\varphi \in L^2(X)$ will do. Moreover, conversely, by examining the proof of Cauchy-Schwarz, we conclude that this condition $\varphi \in L^2(X)$ is in fact necessary. Thus, we have proved that we have:

$$(L^2)^* = L^2$$

But this gives the $V = V^{**}$ result in the statement for our space $V = L^2$, because by dualizing one more time we obtain, as desired:

$$(L^2)^{**} = (L^2)^* = L^2$$

(6) Moving ahead now, let us go back to the l^p spaces, as in (4), but now with general exponents $p \in [1, \infty]$, instead of $p = 2$. The space l^p is by definition given by:

$$l^p = \left\{ (x_i)_{i \in \mathbb{N}} \mid \sum_i |x_i|^p < \infty \right\}$$

Now by arguing as in (4), a linear form $f : l^p \rightarrow \mathbb{C}$ must come as follows:

$$f((x_i)_i) = \sum_i a_i x_i$$

To be more precise, here $a_i \in \mathbb{C}$ are certain scalars, which are subject to an inequality as follows, for a certain constant $A < \infty$, making f well-defined, and bounded:

$$\left| \sum_i a_i x_i \right| \leq A \left(\sum_i |x_i|^p \right)^{1/p}$$

Now recall that the Hölder inequality tells us that we have, with $\frac{1}{p} + \frac{1}{q} = 1$:

$$\left| \sum_i a_i x_i \right| \leq \left(\sum_i |x_i|^p \right)^{1/p} \left(\sum_i |a_i|^q \right)^{1/q}$$

Thus, the linear form $f : l^p \rightarrow \mathbb{C}$ associated to any element $a = (a_i) \in l^q$ will do. Moreover, conversely, by examining the proof of Hölder, we conclude that this condition $a = (a_i) \in l^q$ is in fact necessary. Thus, we have proved that we have:

$$(l^p)^* = l^q$$

But this gives the $V = V^{**}$ result in the statement for our space $V = l^p$, because by dualizing one more time we obtain, as desired:

$$(l^p)^{**} = (l^q)^* = l^q$$

(7) All this is very nice, and time now to generalize everything that we know, by looking at the general spaces $L^p(X)$, with $p \in [1, \infty]$. These spaces are given by:

$$L^p(X) = \left\{ f : X \rightarrow \mathbb{C} \mid \int_X |f(x)|^p dx < \infty \right\}$$

As before in (5), when looking for linear forms $\varphi : L^p(X) \rightarrow \mathbb{C}$, by linearity, and with some measure theory helping, our forms must appear via a formula as follows:

$$\varphi(f) = \int_X f(x) \varphi(x) dx$$

Now in order for this integral to converge, as for our map $\varphi : L^p(X) \rightarrow \mathbb{C}$ to be well-defined, and with the additional requirement that φ must be actually bounded, we must have an inequality as follows, for a certain positive constant $A < \infty$:

$$\left| \int_X f(x) \varphi(x) dx \right| \leq A \left(\int_X |f(x)|^p dx \right)^{1/p}$$

Now recall that the Hölder inequality tells us that we have, with $\frac{1}{p} + \frac{1}{q} = 1$:

$$\left| \int_X f(x) \varphi(x) dx \right| \leq \left(\int_X |f(x)|^p dx \right)^{1/p} \left(\int_X |\varphi(x)|^q dx \right)^{1/q}$$

Thus, the linear form $\varphi : L^p(X) \rightarrow \mathbb{C}$ associated to any function $\varphi \in L^q(X)$ will do. Moreover, conversely, by examining the proof of Hölder, we conclude that this condition $\varphi \in L^q(X)$ is in fact necessary. Thus, we have proved that we have:

$$(L^p)^* = L^q$$

But this gives the $V = V^{**}$ result in the statement for our space $V = L^p$, because by dualizing one more time we obtain, as desired:

$$(L^p)^{**} = (L^q)^* = L^p$$

(8) Finally, let us mention that not all Banach spaces satisfy $V = V^{**}$, with a basic counterexample here being the space c_0 of sequences $x_n \in \mathbb{C}$ satisfying $x_n \rightarrow 0$, with the sup norm. Indeed, computations show that we have the following formulae:

$$c_0^* = l^1, \quad (l^1)^* = l^\infty$$

Thus, in this case $V \subset V^{**}$ is the embedding $c_0 \subset l^\infty$, which is not an isomorphism.

(9) To be more precise here, let us define c_0 as above, as being the space of sequences $x_n \in \mathbb{C}$ satisfying $x_n \rightarrow 0$, with the sup norm. We have then an inclusion as follows:

$$c_0 \subset l^\infty$$

It is then clear that c_0 is closed, so it is indeed a Banach space. Now regarding the duals, we can use here the standard fact that given an inclusion of Banach spaces $V \subset W$, by restricting the linear forms from W to V we obtain a surjection $W^* \rightarrow V^*$. By using this for the above inclusion $c_0 \subset l^\infty$, we obtain a surjection as follows:

$$(l^\infty)^* = l^1 \rightarrow c_0^*$$

But this surjection is clearly injective, so we have indeed $c_0^* = l^1$, as claimed above. \square

9e. Exercises

This was a mixed algebraic and analytic chapter, and as exercises on this, we have:

EXERCISE 9.23. *Review if needed the theory of convex functions, and Jensen.*

EXERCISE 9.24. *Learn about the Lagrange multipliers, that we used in the above.*

EXERCISE 9.25. *Find some other Banach space norms, in finite dimensions.*

EXERCISE 9.26. *Find some other examples of Banach spaces of sequences.*

EXERCISE 9.27. *Learn about bases of Banach spaces, theorems and questions.*

EXERCISE 9.28. *Learn about the tensor products of Banach spaces too.*

EXERCISE 9.29. *Learn more about the reflexive Banach spaces.*

EXERCISE 9.30. *Learn about Banach algebras, and what can be done with them.*

As bonus exercise, find a good Banach space book, and start reading it.

CHAPTER 10

Hilbert spaces

10a. Linear algebra

We discuss in this chapter the Hilbert spaces, which are the most important types of Banach spaces. The idea is very simple, namely that the most basic Banach spaces should be those whose norm comes from a scalar product, according to the following formula:

$$||x|| = \sqrt{\langle x, x \rangle}$$

In fact, we have already met such spaces in chapter 9, because the standard norm on \mathbb{C}^N comes as above, from the standard scalar product on \mathbb{C}^N , given by:

$$\langle x, y \rangle = \sum_i x_i \bar{y}_i$$

More generally, the norm on any $l^2(I)$ comes as above, with the scalar product on $l^2(I)$ being given the same formula. And even more generally, the norm on any space of type $L^2(X)$ comes as above, with the scalar product on $L^2(X)$ being given by:

$$\langle f, g \rangle = \int_X f(x) \overline{g(x)} dx$$

Summarizing, we have some beginning of theory here, and it is tempting at this point to axiomatize the scalar products and the Hilbert spaces, and then develop the theory of these Hilbert spaces, as a continuation of what we did in chapter 9.

However, this is just one point of view on the Hilbert spaces. Alternatively, we can build everything from scratch, simply by starting with what we know about \mathbb{C}^N and its scalar product \langle, \rangle , from linear algebra, and generalizing this to infinite dimensions. And with this being a quite nice approach, because everything that we learned in chapter 9 about the general Banach spaces heavily simplifies in the case of the Hilbert spaces.

Well, hope you get my point, we are here in front of a dilemma, as follows:

DILEMMA 10.1. *We have two ways of viewing the theory of Hilbert spaces, as a generalization of linear algebra, or as a particular case of Banach space theory.*

So, what to do, in this kind of situation? Ask the cats of course, that we have not seen since chapter 4. And the cats, both more in shape than ever, declare:

CATS 10.2. *We cats view the theory of Hilbert spaces as being part of quantum mechanics. But for a human introduction, better go with the linear algebra way.*

Okay, thanks cats, I was sort of expecting that, but the advice at the end is really useful. So, we will go this way, and forgetting now everything that we learned in chapter 9 about Banach spaces, and no worries about this, we will manage to reprove all that things, in the Hilbert space case, let us start with a discussion in finite dimensions.

You certainly know from linear algebra and calculus many things about \mathbb{C}^N and its scalar product \langle, \rangle , and we will review now this. Let us start with:

DEFINITION 10.3. *We endow \mathbb{C}^N with its usual scalar product, namely*

$$\langle x, y \rangle = \sum_i x_i \bar{y}_i$$

so that the norm of the vectors is given by $\|x\| = \sqrt{\langle x, x \rangle}$.

Here the convention for linearity, which is at left, and the notation \langle, \rangle for the scalar product are both taken from physics, and more specifically from cat quantum mechanics. Not to be confused with human quantum mechanics, where the conventions are usually different, with the scalar product being denoted $\langle | \rangle$, and taken linear at right.

Getting now to what can be done with scalar products, as a basic result, we have:

THEOREM 10.4. *The linear maps $T : \mathbb{C}^N \rightarrow \mathbb{C}^N$ are in correspondence with the square matrices $A \in M_N(\mathbb{C})$, with the linear map associated to such a matrix being*

$$Tx = Ax$$

and with the matrix associated to a linear map being $A_{ij} = \langle Te_j, e_i \rangle$.

PROOF. The first assertion is clear, because a linear map $T : \mathbb{C}^N \rightarrow \mathbb{C}^N$ must send a vector $x \in \mathbb{C}^N$ to a certain vector $Tx \in \mathbb{C}^N$, all whose components are linear combinations of the components of x . Thus, we can write, for certain complex numbers $A_{ij} \in \mathbb{C}$:

$$T \begin{pmatrix} x_1 \\ \vdots \\ x_N \end{pmatrix} = \begin{pmatrix} A_{11}x_1 + \dots + A_{1N}x_N \\ \vdots \\ A_{N1}x_1 + \dots + A_{NN}x_N \end{pmatrix}$$

But this means $Tx = Ax$, as claimed. Regarding now the second assertion, with $Tx = Ax$, if we denote by e_1, \dots, e_N the standard basis of \mathbb{C}^N , we have:

$$Te_j = \begin{pmatrix} A_{1j} \\ \vdots \\ A_{Nj} \end{pmatrix}$$

But this gives the second formula, $\langle Te_j, e_i \rangle = A_{ij}$, as desired. □

Our claim now is that, no matter what we want to do with T or A , of advanced type, we will run at some point into their adjoints T^* and A^* , constructed as follows:

PROPOSITION 10.5. *The adjoint operator $T^* : \mathbb{C}^N \rightarrow \mathbb{C}^N$, which is given by*

$$\langle Tx, y \rangle = \langle x, T^*y \rangle$$

corresponds to the adjoint matrix $A^ \in M_N(\mathbb{C})$, given by*

$$(A^*)_{ij} = \bar{A}_{ji}$$

via the correspondence between linear maps and matrices constructed above.

PROOF. Given a linear map $T : \mathbb{C}^N \rightarrow \mathbb{C}^N$, fix $y \in \mathbb{C}^N$, and consider the linear form $\varphi(x) = \langle Tx, y \rangle$. This form must be as follows, for a certain vector $T^*y \in \mathbb{C}^N$:

$$\varphi(x) = \langle x, T^*y \rangle$$

Thus, we have constructed a map $y \rightarrow T^*y$ as in the statement, which is obviously linear, and that we can call T^* . Now by taking the vectors $x, y \in \mathbb{C}^N$ to be elements of the standard basis of \mathbb{C}^N , our defining formula for T^* reads:

$$\langle Te_i, e_j \rangle = \langle e_i, T^*e_j \rangle$$

By reversing the scalar product on the right, this formula can be written as:

$$\langle T^*e_j, e_i \rangle = \overline{\langle Te_i, e_j \rangle}$$

But this means that the matrix of T^* is given by $(A^*)_{ij} = \bar{A}_{ji}$, as desired. \square

Getting back to our claim, the adjoints $*$ are indeed ubiquitous, as shown by:

THEOREM 10.6. *The following happen:*

- (1) $T(x) = Ux$ with $U \in M_N(\mathbb{C})$ is an isometry precisely when $U^* = U^{-1}$.
- (2) $T(x) = Px$ with $P \in M_N(\mathbb{C})$ is a projection precisely when $P^2 = P^* = P$.

PROOF. Let us first recall that the lengths, or norms, of the vectors $x \in \mathbb{C}^N$ can be recovered from the knowledge of the scalar products, as follows:

$$\|x\| = \sqrt{\langle x, x \rangle}$$

Conversely, we can recover the scalar products out of norms, by using the following difficult to remember formula, called complex polarization identity:

$$4 \langle x, y \rangle = \|x + y\|^2 - \|x - y\|^2 + i\|x + iy\|^2 - i\|x - iy\|^2$$

Finally, we will use Proposition 10.5, and more specifically the following formula coming from there, valid for any matrix $A \in M_N(\mathbb{C})$ and any two vectors $x, y \in \mathbb{C}^N$:

$$\langle Ax, y \rangle = \langle x, A^*y \rangle$$

(1) Given a matrix $U \in M_N(\mathbb{C})$, we have indeed the following equivalences, with the first one coming from the polarization identity, and the other ones being clear:

$$\begin{aligned}
 \|Ux\| = \|x\| &\iff \langle Ux, Uy \rangle = \langle x, y \rangle \\
 &\iff \langle x, U^*Uy \rangle = \langle x, y \rangle \\
 &\iff U^*Uy = y \\
 &\iff U^* = U^{-1}
 \end{aligned}$$

(2) Given a matrix $P \in M_N(\mathbb{C})$, in order for $x \rightarrow Px$ to be an oblique projection, we must have $P^2 = P$. Now observe that this projection is orthogonal when:

$$\begin{aligned}
 \langle Px - x, Py \rangle = 0 &\iff \langle P^*Px - P^*x, y \rangle = 0 \\
 &\iff P^*Px - P^*x = 0 \\
 &\iff P^*P - P^* = 0 \\
 &\iff P^*P = P^*
 \end{aligned}$$

By conjugating, $P^*P = P$, so we must have $P = P^*$, which gives the result. \square

At a more advanced level now, diagonalization, we first have the following result:

THEOREM 10.7. *Any matrix $A \in M_N(\mathbb{C})$ which is self-adjoint, $A = A^*$, is diagonalizable, with the diagonalization being of the following type,*

$$A = UDU^*$$

with $U \in U_N$, and with $D \in M_N(\mathbb{R})$ diagonal. The converse holds too.

PROOF. As a first remark, the converse trivially holds, because if we take a matrix of the form $A = UDU^*$, with U unitary and D diagonal and real, then we have:

$$A^* = (UDU^*)^* = UDU^* = A$$

In the other sense now, assume that A is self-adjoint, $A = A^*$. Our first claim is that the eigenvalues are real. Indeed, assuming $Av = \lambda v$, we have:

$$\begin{aligned}
 \lambda \langle v, v \rangle &= \langle Av, v \rangle \\
 &= \langle v, Av \rangle \\
 &= \bar{\lambda} \langle v, v \rangle
 \end{aligned}$$

Thus we obtain $\lambda \in \mathbb{R}$, as claimed. Our next claim now is that the eigenspaces corresponding to different eigenvalues are pairwise orthogonal. Assume indeed that we have $Av = \lambda v$, $Aw = \mu w$. We have then the following computation, using $\lambda, \mu \in \mathbb{R}$:

$$\begin{aligned}
 \lambda \langle v, w \rangle &= \langle Av, w \rangle \\
 &= \langle v, Aw \rangle \\
 &= \mu \langle v, w \rangle
 \end{aligned}$$

Thus $\lambda \neq \mu$ implies $v \perp w$, as claimed. In order now to finish, it remains to prove that the eigenspaces of A span the whole \mathbb{C}^N . For this purpose, let us pick an eigenvector, $Av = \lambda v$. Assuming that we have a vector w orthogonal to it, $v \perp w$, we have:

$$\begin{aligned} \langle Aw, v \rangle &= \langle w, Av \rangle \\ &= \langle w, \lambda v \rangle \\ &= \lambda \langle w, v \rangle \\ &= 0 \end{aligned}$$

Thus, if v is an eigenvector, then the vector space v^\perp is invariant under A . Moreover, since a matrix A is self-adjoint precisely when $\langle Av, v \rangle \in \mathbb{R}$ for any vector $v \in \mathbb{C}^N$, as one can see by expanding the scalar product, the restriction of A to the subspace v^\perp is self-adjoint. Thus, we can proceed by recurrence, and we obtain the result. \square

Let us discuss now the case of the unitary matrices. We have here:

THEOREM 10.8. *Any matrix $U \in M_N(\mathbb{C})$ which is unitary, $U^* = U^{-1}$, is diagonalizable, with the eigenvalues on \mathbb{T} . More precisely we have*

$$U = VDV^*$$

with $V \in U_N$, and with $D \in M_N(\mathbb{T})$ diagonal. The converse holds too.

PROOF. As a first remark, the converse trivially holds, because given a matrix of type $U = VDV^*$, with $V \in U_N$, and with $D \in M_N(\mathbb{T})$ being diagonal, we have:

$$U^* = (VDV^*)^* = VD^{-1}V^{-1} = (VDV^*)^{-1} = U^{-1}$$

Let us prove now the first assertion, stating that the eigenvalues of a unitary matrix $U \in U_N$ belong to the unit circle \mathbb{T} . Assuming $Uv = \lambda v$, we have:

$$\begin{aligned} \langle v, v \rangle &= \langle U^*Uv, v \rangle \\ &= \langle Uv, Uv \rangle \\ &= \langle \lambda v, \lambda v \rangle \\ &= |\lambda|^2 \langle v, v \rangle \end{aligned}$$

Thus $\lambda \in \mathbb{T}$, as claimed. Our next claim now is that the eigenspaces corresponding to different eigenvalues are pairwise orthogonal. Assume indeed that $Uv = \lambda v$, $Uw = \mu w$. We have then the following computation, using $U^* = U^{-1}$ and $\lambda, \mu \in \mathbb{T}$:

$$\begin{aligned} \lambda \langle v, w \rangle &= \langle Uv, w \rangle \\ &= \langle v, U^*w \rangle \\ &= \langle v, U^{-1}w \rangle \\ &= \langle v, \mu^{-1}w \rangle \\ &= \mu \langle v, w \rangle \end{aligned}$$

Thus $\lambda \neq \mu$ implies $v \perp w$, as claimed. In order now to finish, it remains to prove that the eigenspaces of U span the whole \mathbb{C}^N . For this purpose, let us pick an eigenvector, $Uv = \lambda v$. Assuming that we have a vector w orthogonal to it, $v \perp w$, we have:

$$\begin{aligned} \langle Uw, v \rangle &= \langle w, U^*v \rangle \\ &= \langle w, U^{-1}v \rangle \\ &= \langle w, \lambda^{-1}v \rangle \\ &= \lambda \langle w, v \rangle \\ &= 0 \end{aligned}$$

Thus, if v is an eigenvector, then the vector space v^\perp is invariant under U . Now since U is an isometry, so is its restriction to this space v^\perp . Thus this restriction is a unitary, and so we can proceed by recurrence, and we obtain the result. \square

The self-adjoint matrices and the unitary matrices are particular cases of the general notion of a “normal matrix”, and we have here the following result:

THEOREM 10.9. *Any matrix $A \in M_N(\mathbb{C})$ which is normal, $AA^* = A^*A$, is diagonalizable, with the diagonalization being of the following type,*

$$A = UDU^*$$

with $U \in U_N$, and with $D \in M_N(\mathbb{C})$ diagonal. The converse holds too.

PROOF. As a first remark, the converse trivially holds, because if we take a matrix of the form $A = UDU^*$, with U unitary and D diagonal, then we have:

$$AA^* = UDD^*U^* = UD^*DU^* = A^*A$$

In the other sense now, this is something more technical. Our first claim is that a matrix A is normal precisely when the following happens, for any vector v :

$$\|Av\| = \|A^*v\|$$

Indeed, the above equality can be written as follows:

$$\langle AA^*v, v \rangle = \langle A^*Av, v \rangle$$

But this is equivalent to $AA^* = A^*A$, by expanding the scalar products. Our next claim is that A, A^* have the same eigenvectors, with conjugate eigenvalues:

$$Av = \lambda v \implies A^*v = \bar{\lambda}v$$

Indeed, this follows from the following computation, and from the trivial fact that if a matrix A is normal, then so is any matrix of type $A - \lambda 1_N$:

$$\begin{aligned} \|(A^* - \bar{\lambda}1_N)v\| &= \|(A - \lambda 1_N)^*v\| \\ &= \|(A - \lambda 1_N)v\| \\ &= 0 \end{aligned}$$

Let us prove now that the eigenspaces of A are pairwise orthogonal. Assume that we have two eigenvectors, $Av = \lambda v$, $Aw = \mu w$, corresponding to different eigenvalues, $\lambda \neq \mu$. We have the following computation, which shows that $\lambda \neq \mu$ implies $v \perp w$:

$$\begin{aligned}\lambda \langle v, w \rangle &= \langle Av, w \rangle \\ &= \langle v, A^*w \rangle \\ &= \langle v, \bar{\mu}w \rangle \\ &= \mu \langle v, w \rangle\end{aligned}$$

In order to finish, it remains to prove that the eigenspaces of A span the whole \mathbb{C}^N . For this purpose, let us pick two eigenvectors v, w of the matrix AA^* , corresponding to different eigenvalues, $\lambda \neq \mu$. The eigenvalue equations are then as follows:

$$AA^*v = \lambda v \quad , \quad AA^*w = \mu w$$

We have the following computation, using the normality condition $AA^* = A^*A$, and the fact that the eigenvalues of AA^* , and in particular μ , are real:

$$\begin{aligned}\lambda \langle Av, w \rangle &= \langle \lambda Av, w \rangle \\ &= \langle A\lambda v, w \rangle \\ &= \langle AAA^*v, w \rangle \\ &= \langle AA^*Av, w \rangle \\ &= \langle Av, AA^*w \rangle \\ &= \langle Av, \mu w \rangle \\ &= \mu \langle Av, w \rangle\end{aligned}$$

We conclude that we have $\langle Av, w \rangle = 0$. But this reformulates as follows:

$$\lambda \neq \mu \implies A(E_\lambda) \perp E_\mu$$

Now since the eigenspaces of AA^* are pairwise orthogonal, and span the whole \mathbb{C}^N , we deduce from this that these eigenspaces are invariant under A :

$$A(E_\lambda) \subset E_\lambda$$

But with this result in hand, we can finish. Indeed, we can decompose the problem, and the matrix A itself, following these eigenspaces of AA^* , which in practice amounts in saying that we can assume that we only have 1 eigenspace. Now by rescaling, this is the same as assuming that we have $AA^* = 1$. But with this, we are now into the unitary case, that we know how to solve, as explained in Theorem 10.8, and so done. \square

10b. Hilbert spaces

Getting now to infinite dimensions, there are many things to be done here, in order to extend the above linear algebra material. We will do this slowly, as follows:

(1) In the remainder of this chapter we discuss the notion of Hilbert space, with axioms, general theory, examples, and a look into linear operators too.

(2) And in what regards diagonalization in infinite dimensions, which is something trickier than in finite dimensions, we will leave this for later, for chapter 12 below.

Getting started now, we have the following key definition, generalizing what we know from finite dimensions, about the space \mathbb{C}^N , and its scalar product $\langle \cdot, \cdot \rangle$:

DEFINITION 10.10. *A Hilbert space is a complex vector space H with a scalar product $\langle x, y \rangle$, which will be linear at left and antilinear at right,*

$$\langle \lambda x, y \rangle = \lambda \langle x, y \rangle, \quad \langle x, \lambda y \rangle = \bar{\lambda} \langle x, y \rangle$$

which is complete with respect to corresponding norm

$$\|x\| = \sqrt{\langle x, x \rangle}$$

in the sense that any sequence which is Cauchy with respect to $\|\cdot\|$ converges.

In other words, in terms of the Banach space theory from chapter 9, a Hilbert space is a Banach space whose norm comes from a scalar product, via $\|x\| = \sqrt{\langle x, x \rangle}$. However, for reasons discussed in the beginning of this chapter, we would rather prefer to develop the Hilbert space theory independently from what we know from chapter 9.

Getting now to Definition 10.10 as stated, this is something quite subtle, and there is some mathematics encapsulated there, needing some discussion. Indeed, we must prove that $\|x\| = \sqrt{\langle x, x \rangle}$ is a norm. But this comes from the following result:

THEOREM 10.11. *Consider a complex vector space H with a scalar product $\langle \cdot, \cdot \rangle$, and set $\|x\| = \sqrt{\langle x, x \rangle}$. The following happen:*

- (1) *We have the Cauchy-Schwarz inequality $|\langle x, y \rangle| \leq \|x\| \cdot \|y\|$.*
- (2) *We have the Minkowski inequality $\|x + y\| \leq \|x\| + \|y\|$.*
- (3) *Thus, $\|\cdot\|$ is indeed a norm, and $d(x, y) = \|x - y\|$ is a distance.*

PROOF. This is something very standard, the idea being as follows:

(1) In order to prove Cauchy-Schwarz, consider the following quantity, depending on a real variable $t \in \mathbb{R}$, and on a variable on the unit circle, $w \in \mathbb{T}$:

$$f(t) = \|twx + y\|^2$$

By developing f , we see that this is a degree 2 polynomial in t :

$$\begin{aligned} f(t) &= \langle twx + y, twx + y \rangle \\ &= t^2 \langle x, x \rangle + tw \langle x, y \rangle + t\bar{w} \langle y, x \rangle + \langle y, y \rangle \\ &= t^2 \|x\|^2 + 2t \operatorname{Re}(w \langle x, y \rangle) + \|y\|^2 \end{aligned}$$

Since f is obviously positive, its discriminant must be negative:

$$4\operatorname{Re}(w < x, y >) - 4\|x\|^2 \cdot \|y\|^2 \leq 0$$

But this is equivalent to the following condition:

$$|\operatorname{Re}(w < x, y >)| \leq \|x\| \cdot \|y\|$$

Now the point is that we can arrange for the number $w \in \mathbb{T}$ to be such that the quantity $w < x, y >$ is real. Thus, we obtain the following inequality, as desired:

$$| < x, y > | \leq \|x\| \cdot \|y\|$$

(2) This follows indeed from the Cauchy-Schwarz inequality, as follows:

$$\begin{aligned} & \|x + y\| \leq \|x\| + \|y\| \\ \iff & \|x + y\|^2 \leq (\|x\| + \|y\|)^2 \\ \iff & \|x\|^2 + \|y\|^2 + 2\operatorname{Re} < x, y > \leq \|x\|^2 + \|y\|^2 + 2\|x\| \cdot \|y\| \\ \iff & \operatorname{Re} < x, y > \leq \|x\| \cdot \|y\| \end{aligned}$$

(3) This follows indeed from the Minkowski inequality, which corresponds to the triangle inequality, the other two axioms for a distance being trivially satisfied. \square

Summarizing, Definition 10.10 is fully justified. Before getting into basic examples and theory, let us record as well the following result, which further clarifies the relation between Hilbert and Banach spaces, and which is something very useful on its own:

THEOREM 10.12. *The scalar products can be recovered from distances, via the following formula, called complex polarization identity:*

$$4 < x, y > = \|x + y\|^2 - \|x - y\|^2 + i\|x + iy\|^2 - i\|x - iy\|^2$$

Thus, a Banach space is a Hilbert space precisely when this polarization formula defines, out of the norm $\|\cdot\|$, a scalar product $<, >$, satisfying $< x, x > = \|x\|^2$.

PROOF. In what regards the polarization formula in the statement, this is certainly difficult to memorize, but easy to prove, as follows:

$$\begin{aligned} & \|x + y\|^2 - \|x - y\|^2 + i\|x + iy\|^2 - i\|x - iy\|^2 \\ = & \|x\|^2 + \|y\|^2 - \|x\|^2 - \|y\|^2 + i\|x\|^2 + i\|y\|^2 - i\|x\|^2 - i\|y\|^2 \\ & + 2\operatorname{Re}(< x, y >) + 2\operatorname{Re}(< x, y >) + 2i\operatorname{Im}(< x, y >) + 2i\operatorname{Im}(< x, y >) \\ = & 4 < x, y > \end{aligned}$$

As for the last assertion, that is something rather philosophical, obtained by carefully examining Definition 10.10, with the complex polarization identity in mind. \square

At the level of examples now, we first have the following result:

THEOREM 10.13. *Given an index set I , the space of square-summable sequences*

$$l^2(I) = \left\{ (x_i)_{i \in I} \mid \sum_i |x_i|^2 < \infty \right\}$$

is a Hilbert space, with scalar product as follows:

$$\langle x, y \rangle = \sum_i x_i \bar{y}_i$$

When I is finite, $I = \{1, \dots, N\}$, we obtain in this way the usual space $H = \mathbb{C}^N$.

PROOF. We have already met such things before, but let us quickly recall this:

(1) We know that $l^2(I) \subset \mathbb{C}^I$ is the space of vectors satisfying $\|x\| < \infty$. We want to prove that $l^2(I)$ is a vector space, that $\langle x, y \rangle$ is a scalar product on it, that $l^2(I)$ is complete with respect to $\|\cdot\|$, and finally that for $|I| < \infty$ we have $l^2(I) = \mathbb{C}^{|I|}$.

(2) The last assertion, $l^2(I) = \mathbb{C}^{|I|}$ for $|I| < \infty$, is clear, because in this case the sums are finite, so the condition $\|x\| < \infty$ is automatic. So, we know at least one thing.

(3) Regarding the other assertions, everything follows from Cauchy-Schwarz. We first obtain, by raising to the square and expanding, that for any $x, y \in l^2(I)$ we have:

$$\|x + y\|^2 \leq \|x\|^2 + \|y\|^2$$

Thus $l^2(I)$ is indeed a vector space, the other vector space conditions being trivial.

(4) Also, $\langle x, y \rangle$ is surely a scalar product on this vector space, because all the conditions for a scalar product are trivially satisfied.

(5) Finally, the fact that our space $l^2(I)$ is indeed complete with respect to its norm $\|\cdot\|$ follows in the obvious way, the limit of a Cauchy sequence $\{x_n\}$ being the vector $y = (y_i)$ given by $y_i = \lim_{n \rightarrow \infty} x_{ni}$, with all the verifications here being trivial. \square

Going now a bit abstract, we have, more generally, the following result:

THEOREM 10.14. *Given an arbitrary space X with a positive measure μ on it, the space of square-summable complex functions on it, namely*

$$L^2(X) = \left\{ f : X \rightarrow \mathbb{C} \mid \int_X |f(x)|^2 d\mu(x) < \infty \right\}$$

is a Hilbert space, with scalar product as follows:

$$\langle f, g \rangle = \int_X f(x) \overline{g(x)} d\mu(x)$$

When $X = I$ is discrete, meaning that the measure μ on it is the counting measure, $\mu(\{x\}) = 1$ for any $x \in X$, we obtain in this way the previous spaces $l^2(I)$.

PROOF. This is something routine, remake of Theorem 10.13, as follows:

(1) The proof of the first, and main assertion is something perfectly similar to the proof of Theorem 10.13, by replacing everywhere the sums by integrals.

(2) With the remark that we forgot to say in the statement that the L^2 functions are by definition taken up to equality almost everywhere, $f = g$ when $\|f - g\| = 0$.

(3) As for the last assertion, when μ is the counting measure all our integrals here become usual sums, and so we recover in this way Theorem 10.13. \square

As a conclusion to what we did so far, the Hilbert spaces are now axiomatized, and the main examples discussed. In order to do now some geometry on our spaces, in analogy with what we know from finite dimensions, let us start with the following definition:

DEFINITION 10.15. *Let H be a Hilbert space.*

- (1) *We call two vectors orthogonal, $x \perp y$, when $\langle x, y \rangle = 0$.*
- (2) *Given a subset $S \subset H$, we set $S^\perp = \{x \in H \mid x \perp y, \forall y \in S\}$.*

As a first comment here, in relation with (1), it follows from Cauchy-Schwarz that given any two vectors x, y we have $\langle x, y \rangle = \|x\| \cdot \|y\| \cdot \alpha$, with $\alpha \in [-1, 1]$. In the case of the real Hilbert spaces we have $\alpha = \cos t$, with t being the angle between x, y , but here we will only deal with complex Hilbert spaces, where this formula is not available.

In what regards (2), this is something very familiar too, and as an observation here, the subset $S^\perp \subset H$ constructed above is a closed linear space. In finite dimensions a useful, well-known formula here is $S^{\perp\perp} = S$, in case $S \subset H$ itself is a linear space. As explained below, this generalizes to the infinite dimensional setting as $S^{\perp\perp} = \bar{S}$.

Getting now to what can be done with orthogonality, we have here:

THEOREM 10.16. *Let H be a Hilbert space, and $E \subset H$ be a closed subspace.*

- (1) *Given $x \in H$, we can find a unique $y \in E$, minimizing $\|x - y\|$.*
- (2) *With x, y as above, we have $x = y + z$, for a certain $z \in E^\perp$.*
- (3) *Thus, we have a direct sum decomposition $H = E \oplus E^\perp$.*
- (4) *In terms of $H = E \oplus E^\perp$, the projection $x \rightarrow y$ is given by $P(x, y) = x$.*

PROOF. This is something very standard, the idea being as follows:

(1) Given $x \in H$ and two vectors $v, w \in E$, we have the following estimate:

$$\begin{aligned} \|x - v\|^2 + \|x - w\|^2 &= 2 \left(\left\| x - \frac{v + w}{2} \right\|^2 + \left\| \frac{v - w}{2} \right\|^2 \right) \\ &\geq 2d(x, E)^2 + \frac{\|v - w\|^2}{2} \end{aligned}$$

But this shows that any sequence in E realizing the inf in the definition of $d(x, E)$ is Cauchy, so it converges to a vector y . Since E is closed we have $y \in E$, so y realizes the inf. Moreover, again from the above inequality, such a y realizing the inf is unique.

(2) In order to prove $x - y \in E^\perp$, let $v \in E$ and choose $w \in \mathbb{T}$ such that $w \langle x - y, v \rangle$ is a real number. For any $t \in \mathbb{R}$ we have the following equality:

$$\|x - y + twv\|^2 = \|x - y\|^2 + 2tw \langle x - y, v \rangle + t^2\|v\|^2$$

By construction of the vector y we know that this function has a minimum at $t = 0$. But this function is a degree 2 polynomial, so the middle term must vanish:

$$2w \langle x - y, v \rangle = 0$$

Now since this must hold for any $v \in E$, we must have $x - y \in E^\perp$, as desired.

(3) This is consequence of what we found in (1,2).

(4) This is also a consequence of what we found in (1,2). \square

Many other things can be said, as a continuation of the above, as for instance with the formula $E^{\perp\perp} = \bar{E}$, when $E \subset H$ is a closed subspace, which must be adjusted to $E^{\perp\perp} = \bar{E}$, in the case of the arbitrary subspaces $E \subset H$. Also, importantly, we have:

THEOREM 10.17. *Given a Hilbert space H and a closed subspace $E \subset H$, any linear form $f : E \rightarrow \mathbb{C}$ can be extended into a linear form $\tilde{f} : H \rightarrow \mathbb{C}$, having the same norm, and this simply by using $H = E \oplus E^\perp$, and setting $\tilde{f} = 0$ on E^\perp .*

PROOF. This is indeed something self-explanatory. Observe that what we have here is the Hahn-Banach theorem, for the Hilbert spaces, coming with a trivial proof. \square

Still talking abstract functional analysis, the few Banach space findings left from chapter 9 trivialize in the case of Hilbert spaces, as shown by the following result:

THEOREM 10.18. *Let H be a Hilbert space.*

- (1) *Any linear form $f : H \rightarrow \mathbb{C}$ must be of type $f(y) = \langle z, y \rangle$, with $z \in H$.*
- (2) *Thus, we have a Banach space isomorphism $H^* \simeq \bar{H}$.*
- (3) *In particular, H is reflexive as Banach space, $H^{**} = H$.*

PROOF. This is something that we already know from chapter 9, coming from our general discussion there, but we have an elementary proof for this, as follows:

(1) Consider a linear form $f : H \rightarrow \mathbb{C}$. Choose $v \in H$ such that $f(v) \neq 0$. By linearity we may assume $f(v) = 1$. Then each $z \in H$ decomposes in the following way:

$$z = (z - f(z)v) + f(z)v$$

This shows that we have a direct sum decomposition of H , as follows:

$$H = \ker(f) \oplus \mathbb{C}v$$

Now pick $z \in \ker(f)^\perp$ and consider the kernel of the linear form $f_z(y) = \langle z, y \rangle$:

$$\text{Ker}(f_z) = \{y \in H \mid \langle z, y \rangle = 0\} \supset \text{Ker}(f)$$

The linear forms f_z and f are then given by the following formulae:

$$f_z(a + \lambda v) = \lambda f_z(v) \quad , \quad f(a + \lambda v) = \lambda$$

It follows that we have $f = \mu f_z$, with $\mu = f_z(v)^{-1}$, and so that we have, as desired:

$$f = f_{\mu z}$$

(2) This is just an abstract reformulation of what we found in (1).

(3) This follows from (2), because we have $H^{**} = \bar{H}^* = \bar{\bar{H}} = H$. □

As a conclusion to all this, the various general Banach space results from chapter 9 are all clear in the Hilbert space setting. However, do not worry, the Hilbert spaces have their own amount of mystery, that we will explore in what follows.

10c. Bases, separability

At a more advanced level now, we can talk about orthonormal bases, and the related notion of dimension of a Hilbert space. We have here the following result:

THEOREM 10.19. *Any Hilbert space H has an orthonormal basis $\{e_i\}_{i \in I}$, which is by definition a set of vectors whose span is dense in H , and which satisfy*

$$\langle e_i, e_j \rangle = \delta_{ij}$$

with δ being a Kronecker symbol. The cardinality $|I|$ of the index set, which can be finite, countable, or uncountable, depends only on H , and is called dimension of H . We have

$$H \simeq l^2(I)$$

in the obvious way, mapping $\sum \lambda_i e_i \rightarrow (\lambda_i)$. The Hilbert spaces with $\dim H = |I|$ being countable, such as $l^2(\mathbb{N})$, are all isomorphic, and are called separable.

PROOF. We have many assertions here, the idea being as follows:

(1) In finite dimensions an orthonormal basis $\{e_i\}_{i \in I}$ can be constructed by starting with any vector space basis $\{f_i\}_{i \in I}$, and using the Gram-Schmidt procedure. As for the other assertions, these are all clear, from basic linear algebra.

(2) In general, the same method works, namely Gram-Schmidt, with a subtlety coming from the fact that the basis $\{e_i\}_{i \in I}$ will not span in general the whole H , but just a dense subspace of it, as it is in fact obvious by looking at the standard basis of $l^2(\mathbb{N})$.

(3) And there is a second subtlety as well, coming from the fact that the recurrence procedure needed for Gram-Schmidt must be replaced by some sort of “transfinite recurrence”, using standard tools from logic, and more specifically the Zorn lemma.

(4) Finally, everything at the end, regarding our notion of separability for the Hilbert spaces, is clear from definitions, and from our various results above. \square

According to Theorem 10.19, there is only one separable Hilbert space, up to isomorphism. There are many interesting things that can be said, about this magic and unique Hilbert space. As a first result such result, which is something theoretical, we have:

THEOREM 10.20. *The following happen, in relation with separability:*

- (1) *The Hilbert space $H = L^2[-1, 1]$ is separable, with orthonormal basis coming by applying Gram-Schmidt to the basis $\{x^k\}_{k \in \mathbb{N}}$, coming from Weierstrass.*
- (2) *In fact, any $H = L^2(\mathbb{R}, \mu)$, with $d\mu(x) = f(x)dx$, is separable, and the same happens in higher dimensions, for $H = L^2(\mathbb{R}^N, \mu)$, with $d\mu(x) = f(x)dx$.*
- (3) *More generally, given a separable abstract measured space X , the associated Hilbert space of square-summable functions $H = L^2(X)$ is separable.*

PROOF. Many things can be said here, the idea being as follows:

(1) The fact that $H = L^2[-1, 1]$ is separable is clear indeed from the Weierstrass density theorem, which provides us with the algebraic basis $g_k = x^k$, which can be orthogonalized by using the Gram-Schmidt procedure, as explained in Theorem 10.19.

(2) Regarding now more general spaces, of type $H = L^2(\mathbb{R}, \mu)$, we can use here the same argument, after modifying if needed our measure μ , in order for the functions $g_k = x^k$ to be indeed square-summable. As for higher dimensions, the situation here is similar, because we can use here the multivariable polynomials $g_k(x) = x_1^{k_1} \dots x_N^{k_N}$.

(3) Finally, the last assertion, regarding the general spaces of type $H = L^2(X)$, which generalizes all this, comes as a consequence of the general measure theory developed in Part II, and we will leave working out the details here as an instructive exercise. \square

As a conclusion to all this, which is a bit philosophical, we have:

CONCLUSION 10.21. *We are interested in one space, namely the unique separable Hilbert space H , but due to various technical reasons, it is often better to forget that we have $H = l^2(\mathbb{N})$, and say instead that we have $H = L^2(X)$, with X being a separable measured space, or simply say that H is an abstract separable Hilbert space.*

It is also possible to make some physics comments here, with the unique separable Hilbert space H from Conclusion 10.21, that we will be presumably obsessed with, in what follows, being, and no surprise here, the space that we live in.

Let us go back now to Theorem 10.20 and its proof, which is something quite subtle. That material leads us into orthogonal polynomials, which are defined as follows:

DEFINITION 10.22. *The orthogonal polynomials with respect to $d\mu(x) = f(x)dx$ are polynomials $P_k \in \mathbb{R}[x]$ of degree $k \in \mathbb{N}$, which are orthogonal inside $H = L^2(\mathbb{R}, \mu)$:*

$$\int_{\mathbb{R}} P_k(x)P_l(x)f(x)dx = 0 \quad , \quad \forall k \neq l$$

Equivalently, these orthogonal polynomials $\{P_k\}_{k \in \mathbb{N}}$, which are each unique modulo scalars, appear from the Weierstrass basis $\{x^k\}_{k \in \mathbb{N}}$, by doing Gram-Schmidt.

As a first observation, the orthogonal polynomials exist indeed for any real measure $d\mu(x) = f(x)dx$, because we can obtain them from the monomials x^k via Gram-Schmidt, as indicated above. It is possible to be a bit more explicit here, as follows:

THEOREM 10.23. *The orthogonal polynomials with respect to μ are given by*

$$P_k = c_k \begin{vmatrix} M_0 & M_1 & \dots & M_k \\ M_1 & M_2 & \dots & M_{k+1} \\ \vdots & \vdots & & \vdots \\ M_{k-1} & M_k & \dots & M_{2k-1} \\ 1 & x & \dots & x^k \end{vmatrix}$$

where $M_k = \int_{\mathbb{R}} x^k d\mu(x)$ are the moments of μ , and $c_k \in \mathbb{R}^*$ can be any numbers.

PROOF. Let us first see what happens at small values of $k \in \mathbb{N}$. At $k = 0$ our formula is as follows, stating that the first polynomial P_0 must be a constant, as it should:

$$P_0 = c_0 |M_0| = c_0$$

At $k = 1$ now, again by using $M_0 = 1$, the formula is as follows:

$$P_1 = c_1 \begin{vmatrix} M_0 & M_1 \\ 1 & x \end{vmatrix} = c_1(x - M_1)$$

But this is again the good formula, because the degree is 1, and we have:

$$\begin{aligned} \langle 1, P_1 \rangle &= c_1 \langle 1, x - M_1 \rangle \\ &= c_1(\langle 1, x \rangle - \langle 1, M_1 \rangle) \\ &= c_1(M_1 - M_1) \\ &= 0 \end{aligned}$$

At $k = 2$ now, things get more complicated, with the formula being as follows:

$$P_2 = c_2 \begin{vmatrix} M_0 & M_1 & M_2 \\ M_1 & M_2 & M_3 \\ 1 & x & x^2 \end{vmatrix}$$

However, no need for big computations here, in order to check the orthogonality, because by using the fact that x^k integrates up to M_k , we obtain:

$$\langle 1, P_2 \rangle = \int_{\mathbb{R}} P_2(x) d\mu(x) = c_2 \begin{vmatrix} M_0 & M_1 & M_2 \\ M_1 & M_2 & M_3 \\ M_0 & M_1 & M_2 \end{vmatrix} = 0$$

Similarly, again by using the fact that x^k integrates up to M_k , we have as well:

$$\langle x, P_2 \rangle = \int_{\mathbb{R}} x P_2(x) d\mu(x) = c_2 \begin{vmatrix} M_0 & M_1 & M_2 \\ M_1 & M_2 & M_3 \\ M_1 & M_2 & M_3 \end{vmatrix} = 0$$

Thus, result proved at $k = 0, 1, 2$, and the proof in general is similar. \square

In practice now, all this leads us to a lot of interesting combinatorics, and countless things can be said. For the simplest measured space $X \subset \mathbb{R}$, which is the interval $[-1, 1]$, with its uniform measure, the orthogonal basis problem can be solved as follows:

THEOREM 10.24. *The orthonormal polynomials for $L^2[-1, 1]$, subject to*

$$\int_{-1}^1 P_k(x) P_l(x) dx = \delta_{kl}$$

and called Legendre polynomials, satisfy the equation

$$(1 - x^2)P_k''(x) - 2xP_k'(x) + k(k+1)P_k(x) = 0$$

which is the Legendre equation from physics. Moreover, we have the formula

$$P_k(x) = \frac{1}{2^k k!} \cdot \frac{d^k}{dx^k} (1 - x^2)^k$$

called Rodrigues formula for the Legendre polynomials.

PROOF. The idea here is that thinking at what Gram-Schmidt does, this is certainly something by recurrence. And examining the recurrence leads to the Legendre equation. As for the Rodrigues formula, we have two choices here, either by verifying that $\{P_k\}$ is orthonormal, or by verifying the Legendre equation. And both methods work. \square

The above result is just the tip of the iceberg, and as a continuation, we have:

THEOREM 10.25. *The orthogonal polynomials for $L^2[-1, 1]$, with measure*

$$d\mu(x) = (1 - x)^\alpha (1 + x)^\beta dx$$

called Jacobi polynomials, satisfy as well a degree 2 equation, and are given by:

$$P_k(x) = \frac{(-1)^k}{2^k k!} (1 - x)^{-\alpha} (1 + x)^{-\beta} \frac{d^k}{dx^k} [(1 - x)^\alpha (1 + x)^\beta (1 - x^2)^k]$$

At $\alpha = \beta = 0$ we recover the Legendre polynomials, and at $\alpha = \beta = \pm \frac{1}{2}$ we recover the Chebycheff polynomials of the first and second kind, from trigonometry.

PROOF. Obviously, many things going on here, and much more can be added, but the idea is quite simple, namely that this appears as a generalization of Theorem 10.24. We will leave learning more about all this as an interesting exercise. \square

Getting now to other spaces $X \subset \mathbb{R}$, of particular interest here is the following result, which complements well Theorem 10.24, for the needs of basic quantum mechanics:

THEOREM 10.26. *The orthogonal polynomials for $L^2[0, \infty)$, with scalar product*

$$\langle f, g \rangle = \int_0^\infty f(x)g(x)e^{-x} dx$$

are the Laguerre polynomials $\{P_k\}$, given by the following formula,

$$P_k(x) = \frac{e^x}{k!} \cdot \frac{d^k}{dx^k} (e^{-x} x^k)$$

called Rodrigues formula for the Laguerre polynomials.

PROOF. The story here is very similar to that of the Legendre polynomials, and many further things can be said here, with exercise for you to learn a bit about all this. \square

Finally, regarding the space $X = \mathbb{R}$ itself, we have here the following result:

THEOREM 10.27. *The orthogonal polynomials for $L^2(\mathbb{R})$, with scalar product*

$$\langle f, g \rangle = \int_{-\infty}^\infty f(x)g(x)e^{-x^2} dx$$

are the Hermite polynomials $\{P_k\}$, given by the following formula,

$$P_k(x) = (-1)^k e^{x^2} \cdot \frac{d^k}{dx^k} (e^{-x^2})$$

called Rodrigues formula for the Hermite polynomials.

PROOF. As before, the story here is quite similar to that of the Legendre and other orthogonal polynomials, and exercise for you to learn a bit about all this. \square

And with this, good news, end of the story with the orthogonal polynomials, at least at the very introductory level, and this due to the following fact, which is something quite technical, and that we will not attempt to prove, or even explain in detail here:

FACT 10.28. *From an abstract point of view, coming from degree 2 equations, and Rodrigues formulae for the solutions, there are only three types of “classical” orthogonal polynomials, namely the Jacobi, Laguerre and Hermite ones, discussed above.*

Finally, as already mentioned, the above results are very useful in the context of basic quantum mechanics, and more specifically, for solving the hydrogen atom, following Heisenberg and Schrödinger. Again, exercise for you to learn a bit about this.

10d. Linear operators

Time now to slow down, after the above exciting material regarding the orthogonal polynomials. Remember indeed from the beginning of this chapter that our main goal is that of generalizing linear algebra, to the case of infinite dimensions. But for generalizing the linear algebra basics, what we have in Conclusion 10.21 will basically do.

Let us start with a basic result regarding the linear operators, as follows:

THEOREM 10.29. *Given a Hilbert space H , consider the linear operators $T : H \rightarrow H$, and for each such operator define its norm by the following formula:*

$$\|T\| = \sup_{\|x\|=1} \|Tx\|$$

The operators which are bounded, $\|T\| < \infty$, form then a complex algebra $B(H)$, which is complete with respect to $\|\cdot\|$. When H comes with a basis $\{e_i\}_{i \in I}$, we have

$$B(H) \subset \mathcal{L}(H) \subset M_I(\mathbb{C})$$

where $\mathcal{L}(H)$ is the algebra of all linear operators $T : H \rightarrow H$, and $\mathcal{L}(H) \subset M_I(\mathbb{C})$ is the correspondence $T \rightarrow M$ obtained via the usual linear algebra formulae, namely:

$$T(x) = Mx \quad , \quad M_{ij} = \langle Te_j, e_i \rangle$$

In infinite dimensions, none of the above two inclusions is an equality.

PROOF. In what regards the first assertion, we already know this from chapter 9. Regarding now the second assertion, given a bounded operator $T : H \rightarrow H$, let us associate to it a matrix $M \in M_I(\mathbb{C})$ as in the statement, by the following formula:

$$M_{ij} = \langle Te_j, e_i \rangle$$

The correspondence $T \rightarrow M$ is then linear, having kernel is $\{0\}$, so we have an embedding $B(H) \subset M_I(\mathbb{C})$. Our claim now is that this embedding is multiplicative. But this is clear too, because if we denote by $T \rightarrow M_T$ our correspondence, we have:

$$\begin{aligned} (M_{ST})_{ij} &= \langle STE_j, e_i \rangle \\ &= \left\langle S \sum_k \langle Te_j, e_k \rangle e_k, e_i \right\rangle \\ &= \sum_k \langle Se_k, e_i \rangle \langle Te_j, e_k \rangle \\ &= \sum_k (M_S)_{ik} (M_T)_{kj} \\ &= (M_S M_T)_{ij} \end{aligned}$$

Next, we must prove that the original operator $T : H \rightarrow H$ can be recovered from its matrix $M \in M_I(\mathbb{C})$ via the formula in the statement, namely $Tx = Mx$. But it is enough to check this on the vectors of the basis, $x = e_j$, and here we have, indeed:

$$(Te_j)_i = \langle Te_j, e_i \rangle = M_{ij} = (Me_j)_i$$

Finally, in finite dimensions the inclusions $B(H) \subset \mathcal{L}(H) \subset M_I(\mathbb{C})$ are of course both isomorphisms, but in general this does not hold. Indeed, over the space $H = L^2(\mathbb{N})$, a basic example of matrix not coming from a linear operator is the all-one matrix:

$$M = \begin{pmatrix} 1 & 1 & \dots \\ 1 & 1 & \dots \\ \vdots & \vdots & \end{pmatrix}$$

As for the examples of linear operators which are not bounded, these are more complicated, coming from logic, and we will not really need them in what follows. \square

As in the finite dimensional case, we can talk about adjoint operators, in this setting, the definition and main properties of the construction $T \rightarrow T^*$ being as follows:

THEOREM 10.30. *Each operator $T \in B(H)$ has an adjoint $T^* \in B(H)$, given by:*

$$\langle Tx, y \rangle = \langle x, T^*y \rangle$$

The operation $T \rightarrow T^$ is antilinear, antimultiplicative, involutive, and satisfies:*

$$\|T\| = \|T^*\| \quad , \quad \|TT^*\| = \|T\|^2$$

When H comes with a basis $\{e_i\}_{i \in I}$, the operation $T \rightarrow T^$ corresponds to*

$$(M^*)_{ij} = \overline{M_{ji}}$$

at the level of the associated matrices $M \in M_I(\mathbb{C})$.

PROOF. There are several things to be proved here, the idea being as follows:

(1) Given $y \in H$, the formula $\varphi(x) = \langle Tx, y \rangle$ defines a linear map $H \rightarrow \mathbb{C}$. Thus, we must have a formula as follows, for a certain vector $T^*y \in H$:

$$\varphi(x) = \langle x, T^*y \rangle$$

Moreover, this vector $T^*y \in H$ is unique with this property, and we conclude from this that the formula $y \rightarrow T^*y$ defines a certain map $T^* : H \rightarrow H$, which is unique.

(2) Let us prove that we have $T^* \in B(H)$. By using once again the uniqueness of T^* , we conclude that we have the following formulae, which show that T^* is linear:

$$T^*(x + y) = T^*x + T^*y \quad , \quad T^*(\lambda x) = \lambda T^*x$$

Observe also that T^* is bounded as well, because we have:

$$\begin{aligned} \|T\| &= \sup_{\|x\|=1} \sup_{\|y\|=1} \langle Tx, y \rangle \\ &= \sup_{\|y\|=1} \sup_{\|x\|=1} \langle x, T^*y \rangle \\ &= \|T^*\| \end{aligned}$$

(3) The fact that the correspondence $T \rightarrow T^*$ is antilinear, antimultiplicative, and is an involution comes from the following formulae, coming from uniqueness:

$$(S + T)^* = S^* + T^* \quad , \quad (\lambda T)^* = \bar{\lambda} T^*$$

$$(ST)^* = T^* S^* \quad , \quad (T^*)^* = T$$

As for $\|T\| = \|T^*\|$, this is something that we already know, from the proof of (2).

(4) In order to prove now the formula $\|TT^*\| = \|T\|^2$, observe first that we have:

$$\|TT^*\| \leq \|T\| \cdot \|T^*\| = \|T\|^2$$

On the other hand, we have as well the following estimate:

$$\begin{aligned} \|T\|^2 &= \sup_{\|x\|=1} | \langle Tx, Tx \rangle | \\ &= \sup_{\|x\|=1} | \langle x, T^*Tx \rangle | \\ &\leq \|T^*T\| \end{aligned}$$

By replacing $T \rightarrow T^*$ we obtain from this $\|T\|^2 \leq \|TT^*\|$, and we are done.

(5) Regarding the last assertion, we can compute the matrix $M^* \in M_I(\mathbb{C})$ associated to the operator $T^* \in B(H)$, by using $\langle Tx, y \rangle = \langle x, T^*y \rangle$, in the following way:

$$\begin{aligned} (M^*)_{ij} &= \langle T^*e_j, e_i \rangle \\ &= \overline{\langle e_i, T^*e_j \rangle} \\ &= \overline{\langle Te_i, e_j \rangle} \\ &= \overline{M}_{ji} \end{aligned}$$

Thus, we have reached to the usual formula for the adjoints of matrices, as desired. \square

As in finite dimensions, the operators T, T^* can be thought of as being “twin brothers”, and there is a lot more mathematics connecting them, as for instance:

THEOREM 10.31. *Given a bounded operator $T \in B(H)$, the following happen:*

- (1) $\ker T^* = (Im T)^\perp$.
- (2) $\overline{Im T^*} = (\ker T)^\perp$.

PROOF. Both these assertions are elementary, as follows:

(1) Let us first prove “ \subset ”. Assuming $T^*x = 0$, we have indeed $x \perp \text{Im}T$, because:

$$\langle x, Ty \rangle = \langle T^*x, y \rangle = 0$$

As for “ \supset ”, assuming $\langle x, Ty \rangle = 0$ for any y , we have $T^*x = 0$, because:

$$\langle T^*x, y \rangle = \langle x, Ty \rangle = 0$$

(2) This can be deduced from (1), applied to the operator T^* , as follows:

$$(\ker T)^\perp = (\text{Im}T^*)^{\perp\perp} = \overline{\text{Im}T^*}$$

Here we have used the formula $K^{\perp\perp} = \bar{K}$, valid for any linear subspace $K \subset H$ of a Hilbert space, which for K closed reads $K^{\perp\perp} = K$, and comes from $H = K \oplus K^\perp$, and which in general follows from $K^{\perp\perp} \subset \bar{K}^{\perp\perp} = \bar{K}$, the reverse inclusion being clear. \square

At the level of examples now, let us begin with the rotations. Things here are quite tricky in arbitrary dimensions. We first have the following result:

THEOREM 10.32. *For a linear operator $U \in B(H)$ the following conditions are equivalent, and if they are satisfied, we say that U is an isometry:*

- (1) *U is a metric space isometry, $d(Ux, Uy) = d(x, y)$.*
- (2) *U is a normed space isometry, $\|Ux\| = \|x\|$.*
- (3) *U preserves the scalar product, $\langle Ux, Uy \rangle = \langle x, y \rangle$.*
- (4) *U satisfies the isometry condition $U^*U = 1$.*

In finite dimensions, we recover in this way the usual unitary transformations.

PROOF. The proofs are similar to those in finite dimensions, as follows:

(1) \iff (2) This follows indeed from the formula of the distances, namely:

$$d(x, y) = \|x - y\|$$

(2) \iff (3) This is again standard, because we can pass from scalar products to distances, and vice versa, by using $\|x\| = \sqrt{\langle x, x \rangle}$, and the polarization formula:

$$4\langle x, y \rangle = \|x + y\|^2 - \|x - y\|^2 + i\|x + iy\|^2 - i\|x - iy\|^2$$

(3) \iff (4) We have indeed the following equivalences, by using the standard formula $\langle Tx, y \rangle = \langle x, T^*y \rangle$, which defines the adjoint operator:

$$\begin{aligned} \langle Ux, Uy \rangle = \langle x, y \rangle &\iff \langle x, U^*Uy \rangle = \langle x, y \rangle \\ &\iff U^*Uy = y \\ &\iff U^*U = 1 \end{aligned}$$

Thus, we are led to the conclusions in the statement. \square

The point now is that the condition $U^*U = 1$ does not imply in general $UU^* = 1$, the simplest counterexample here being the shift operator on $l^2(\mathbb{N})$:

PROPOSITION 10.33. *The shift operator on the space $l^2(\mathbb{N})$, given by*

$$S(e_i) = e_{i+1}$$

*is an isometry, $S^*S = 1$. However, we have $SS^* \neq 1$.*

PROOF. The adjoint of the shift is given by the following formula:

$$S^*(e_i) = \begin{cases} e_{i-1} & \text{if } i > 0 \\ 0 & \text{if } i = 0 \end{cases}$$

When composing S, S^* , in one sense we obtain the following formula:

$$S^*S(e_i) = e_i$$

In other other sense now, we obtain the following formula:

$$SS^*(e_i) = \begin{cases} e_i & \text{if } i > 0 \\ 0 & \text{if } i = 0 \end{cases}$$

Summarizing, the compositions are given by the following formulae:

$$S^*S = 1 \quad , \quad SS^* = Proj(e_0^\perp)$$

Thus, we are led to the conclusions in the statement. \square

As a conclusion, the notion of isometry is not the correct infinite dimensional analogue of the notion of unitary. However, the unitary operators can be introduced as follows:

THEOREM 10.34. *For a linear operator $U \in B(H)$ the following conditions are equivalent, and if they are satisfied, we say that U is a unitary:*

- (1) *U is an isometry, which is invertible.*
- (2) *U, U^{-1} are both isometries.*
- (3) *U, U^* are both isometries.*
- (4) *$UU^* = U^*U = 1$.*
- (5) *$U^* = U^{-1}$.*

Moreover, the unitary operators form a group $U(H) \subset B(H)$.

PROOF. There are several statements here, the idea being as follows:

(1) The various equivalences in the statement are all clear from definitions, and from Theorem 10.32 in what regards the various possible notions of isometries which can be used, by using the formula $(ST)^* = T^*S^*$ for the adjoints of the products of operators.

(2) The fact that the products and inverses of unitaries are unitaries is also clear, and we conclude that the unitary operators form a group $U(H) \subset B(H)$, as stated. \square

Let us discuss now the projections. We have here the following result:

THEOREM 10.35. *For a linear operator $P \in B(H)$ the following conditions are equivalent, and if they are satisfied, we say that P is a projection:*

- (1) *P is the orthogonal projection on a closed subspace $K \subset H$.*
- (2) *P satisfies the projection equations $P^2 = P^* = P$.*

PROOF. As in finite dimensions, P is an abstract projection, not necessarily orthogonal, when it is an idempotent, algebraically speaking, in the sense that we have:

$$P^2 = P$$

The point now is that this projection is orthogonal when:

$$\begin{aligned} \langle Px - x, Py \rangle = 0 &\iff \langle P^*Px - P^*x, y \rangle = 0 \\ &\iff P^*Px - P^*x = 0 \\ &\iff P^*P - P^* = 0 \\ &\iff P^*P = P^* \end{aligned}$$

Now observe that by conjugating, we obtain $P^*P = P$. Thus, we must have $P = P^*$, and so we have shown that any orthogonal projection must satisfy, as claimed:

$$P^2 = P^* = P$$

Conversely, if this condition is satisfied, $P^2 = P$ shows that P is a projection, and $P = P^*$ shows via the above computation that P is indeed orthogonal. \square

There is a relation between the projections and the general isometries, such as the shift S that we met before, and we have the following result:

PROPOSITION 10.36. *Given an isometry $U \in B(H)$, the operator*

$$P = UU^*$$

is a projection, namely the orthogonal projection on $\text{Im}(U)$.

PROOF. Assume indeed that we have an isometry, $U^*U = 1$. The fact that $P = UU^*$ is indeed a projection can be checked abstractly, as follows:

$$(UU^*)^* = UU^*$$

$$UU^*UU^* = UU^*$$

As for the last assertion, this is something that we already met, for the shift, and the situation in general is similar, with the result itself being clear. \square

More generally now, along the same lines, and clarifying the whole situation with the unitaries and isometries, we have the following result:

THEOREM 10.37. *An operator $U \in B(H)$ is a partial isometry, in the usual geometric sense, when the following two operators are projections:*

$$P = UU^* \quad , \quad Q = U^*U$$

Moreover, the isometries, adjoints of isometries and unitaries are respectively characterized by the conditions $Q = 1$, $P = 1$, $P = Q = 1$.

PROOF. The first assertion is a straightforward extension of Proposition 10.36, and the second assertion follows from various results regarding isometries established above. \square

So long for basic operator theory, fundamentals and examples. We will be back to this in chapter 12 below, with suitable diagonalization results for the self-adjoint operators, $T = T^*$, and then for the general normal operators too, $TT^* = T^*T$, in analogy with what we learned in the beginning of this chapter, in finite dimensions.

10e. Exercises

We had an exciting chapter here, and as exercises on this, we have:

EXERCISE 10.38. *What is the formula of the rank one projection on $x \in \mathbb{C}^N$?*

EXERCISE 10.39. *Prove that the diagonalizable matrices in $M_N(\mathbb{C})$ are dense.*

EXERCISE 10.40. *Work out the equality case in Cauchy-Schwarz, and Minkowski.*

EXERCISE 10.41. *Review the full basic Banach space theory, for the Hilbert spaces.*

EXERCISE 10.42. *Learn if needed the full proof of the Zorn lemma.*

EXERCISE 10.43. *Learn more about various families of orthogonal polynomials.*

EXERCISE 10.44. *What should symmetry mean, in the operator theory setting?*

EXERCISE 10.45. *Try to axiomatize the Banach $*$ -algebras, with $B(H)$ in mind.*

As bonus exercise, and no surprise here, learn some quantum mechanics.

CHAPTER 11

Function spaces

11a. Distributions

Welcome to advanced functional analysis. We have seen in chapters 9-10 the basic theory of Banach spaces and Hilbert spaces, and normally, with the knowledge of the L^p spaces, and in particular of the L^2 spaces, we can do many things. However, there are far more things that must be known, and this even at the beginner level, and this chapter and the next one will be an introduction to this. Our plan will be as follows:

(1) In this chapter, which will be basically a continuation of chapter 9, we discuss a number of fundamental functional analysis topics, namely convolution, distributions, complex analysis, Fourier analysis, and some basic applications to physics.

(2) And in the next chapter, which will be basically a continuation of chapter 10, we discuss a fundamental question that we still have left from there, namely the diagonalization of the self-adjoint operators, and of the normal operators.

Getting started, let us first talk about convolution. We have already talked in Part I about the convolution of Dirac masses, according to the formula $\delta_x * \delta_y = \delta_{x+y}$, and of more general discrete measures, according to this formula, and linearity. In the continuous measure setting, at the level of densities, the relevant definition is as follows:

DEFINITION 11.1. *The convolution of two functions $f, g : \mathbb{R} \rightarrow \mathbb{C}$ is the function*

$$(f * g)(x) = \int_{\mathbb{R}} f(x - y)g(y)dy$$

provided that the function $y \rightarrow f(x - y)g(y)$ is indeed integrable, for any x .

There are many reasons for introducing this operation, that we will gradually discover, in what follows. As a basic example, let us take $g = \chi_{[0,1]}$. We have then:

$$(f * g)(x) = \int_0^1 f(x - y)dy$$

Thus, with this choice of g , the operation $f \rightarrow f * g$ has some sort of “regularizing effect”, that can be useful for many purposes. We will be back to this, later.

Go on ahead with more theory, let us try to understand when the convolution operation is well-defined. We have here the following basic result:

PROPOSITION 11.2. *The convolution operation is well-defined on the space*

$$C_c(\mathbb{R}) = \left\{ f \in C(\mathbb{R}) \mid \text{supp}(f) = \text{compact} \right\}$$

of continuous functions $f : \mathbb{R} \rightarrow \mathbb{C}$ having compact support.

PROOF. We have several things to be proved, the idea being as follows:

(1) First we must show that given two functions $f, g \in C_c(\mathbb{R})$, their convolution $f * g$ is well-defined, as a function $f * g : \mathbb{R} \rightarrow \mathbb{C}$. But this follows from the following estimate, where l denotes the length of the compact subsets of \mathbb{R} :

$$\begin{aligned} \int_{\mathbb{R}} |f(x-y)g(y)| dy &= \int_{\text{supp}(g)} |f(x-y)g(y)| dy \\ &\leq \max(g) \int_{\text{supp}(g)} |f(x-y)| dy \\ &\leq \max(g) \cdot l(\text{supp}(g)) \cdot \max(f) \\ &< \infty \end{aligned}$$

(2) Next, we must show that the function $f * g : \mathbb{R} \rightarrow \mathbb{C}$ that we constructed is indeed continuous. But this follows from the following estimate, where K_f is the constant of uniform continuity for the function $f \in C_c(\mathbb{R})$:

$$\begin{aligned} |(f * g)(x + \varepsilon) - (f * g)(x)| &= \left| \int_{\mathbb{R}} f(x + \varepsilon - y)g(y) dy - \int_{\mathbb{R}} f(x - y)g(y) dy \right| \\ &= \left| \int_{\mathbb{R}} (f(x + \varepsilon - y) - f(x - y))g(y) dy \right| \\ &\leq \int_{\mathbb{R}} |f(x + \varepsilon - y) - f(x - y)| \cdot |g(y)| dy \\ &\leq K_f \cdot \varepsilon \cdot \int_{\mathbb{R}} |g| \end{aligned}$$

(3) Finally, we must show that the function $f * g \in C(\mathbb{R})$ that we constructed has indeed compact support. For this purpose, our claim is that we have:

$$\text{supp}(f * g) \subset \text{supp}(f) + \text{supp}(g)$$

In order to prove this claim, observe that we have, by definition of $f * g$:

$$(f * g)(x) = \int_{\mathbb{R}} f(x - y)g(y) dy = \int_{\text{supp}(g)} f(x - y)g(y) dy$$

But this latter quantity being 0 for $x \notin \text{supp}(f) + \text{supp}(g)$, this gives the result. \square

Here are now a few remarkable properties of the convolution operation:

PROPOSITION 11.3. *The following hold, for the functions in $C_c(\mathbb{R})$:*

- (1) $f * g = g * f$.
- (2) $f * (g * h) = (f * g) * h$.
- (3) $f * (\lambda g + \mu h) = \lambda f * g + \mu f * h$.

PROOF. These formulae are all elementary, the idea being as follows:

- (1) This follows from the following computation, with $y = x - t$:

$$\begin{aligned}
 (f * g)(x) &= \int_{\mathbb{R}} f(x - y)g(y)dy \\
 &= \int_{\mathbb{R}} f(t)g(x - t)dt \\
 &= \int_{\mathbb{R}} g(x - t)f(t)dt \\
 &= (g * f)(x)
 \end{aligned}$$

- (2) This is clear from definitions.

- (3) Once again, this is clear from definitions. □

In relation with derivatives, and with the “regularizing effect” of the convolution operation mentioned after Definition 11.1, we have the following result:

THEOREM 11.4. *Given two functions $f, g \in C_c(\mathbb{R})$, assuming that g is differentiable, then so is $f * g$, with derivative given by the following formula:*

$$(f * g)' = f * g'$$

*More generally, given $f, g \in C_c(\mathbb{R})$, and assuming that g is k times differentiable, then so is $f * g$, with k -th derivative given by $(f * g)^{(k)} = f * g^{(k)}$.*

PROOF. In what regards the first assertion, with $y = x - t$, then $t = x - y$, we get:

$$\begin{aligned}
 (f * g)'(x) &= \frac{d}{dx} \int_{\mathbb{R}} f(x - y)g(y)dy \\
 &= \frac{d}{dx} \int_{\mathbb{R}} f(t)g(x - t)dt \\
 &= \int_{\mathbb{R}} f(t)g'(x - t)dt \\
 &= \int_{\mathbb{R}} f(x - y)g'(y)dy \\
 &= (f * g')(x)
 \end{aligned}$$

As for the second assertion, this follows from the first one, by recurrence. □

Finally, getting beyond the compactly supported continuous functions, we have the following result, which is of particular theoretical importance:

THEOREM 11.5. *The convolution operation is well-defined on $L^1(\mathbb{R})$, and we have:*

$$\|f * g\|_1 \leq \|f\|_1 \|g\|_1$$

*Thus, if $f \in L^1(\mathbb{R})$ and $g \in C_c^k(\mathbb{R})$, then $f * g$ is well-defined, and $f * g \in C_c^k(\mathbb{R})$.*

PROOF. In what regards the first assertion, this follows from:

$$\begin{aligned} \int_{\mathbb{R}} |(f * g)(x)| dx &\leq \int_{\mathbb{R}} \int_{\mathbb{R}} |f(x-y)g(y)| dy dx \\ &= \int_{\mathbb{R}} \int_{\mathbb{R}} |f(x-y)g(y)| dx dy \\ &= \int_{\mathbb{R}} |f(x)| dx \int_{\mathbb{R}} |g(y)| dy \end{aligned}$$

As for the second assertion, this follows from this, and from Theorem 11.4. \square

Summarizing, we have some theory for the convolution of functions, in analogy with what we knew from Part I about the convolution of Dirac masses, $\delta_x * \delta_y = \delta_{x+y}$.

Let us discuss now a related topic, the mathematical distributions. These are something quite smart, and as an advertisement for what we will be doing, we have:

ADVERTISEMENT 11.6. *With a suitable theory of distributions, covering both the functions and the Dirac masses, the basic step function, namely*

$$H(x) = \begin{cases} 0 & (x \leq 0) \\ 1 & (x > 0) \end{cases}$$

is differentiable when viewed as distribution, with derivative $H' = \delta_0$.

And isn't this crazy, hope you agree with me. Getting started now, there is a price to pay for doing such things, namely formulating a technical definition, as follows:

DEFINITION 11.7. *A distribution on an open interval $I \subset \mathbb{R}$ is a functional*

$$\varphi : C_c^\infty(I) \rightarrow \mathbb{C}$$

such that for any $K \subset I$ compact, there exist $n \in \mathbb{N}$ and $c > 0$ such that

$$|\varphi(f)| \leq c \|f\|_{C^n(K)}$$

for any $f \in C_c^\infty(I)$ having support in K , where $\|f\|_{C^n(K)} = \sup_{x \in K} \sum_{i=0}^n |f^{(i)}(x)|$.

Obviously, this is something quite technical, but believe me, every little thing in the above is there for a reason, in order to the whole theory to work fine.

At the level of main examples of distributions, we have the integration functionals associated to the measures, and in particular to the measures having a density. In view of this, we can consider any function $f \in L^1(I)$, viewed as density, as being a distribution. Other basic examples include the Dirac masses δ_x at the points $x \in I$.

Regarding the general theory of distributions, that is quite similar to the theory of functions. Algebraically, the distributions form a vector space, and are subject to a number of supplementary operations too, such as dilations, translations and so on, and multiplication by functions too. Analytically, we can talk about convergence of distributions, $\varphi_n \rightarrow \varphi$, and about their support too, $\text{supp}(\varphi) \subset I$, in a quite straightforward way.

Getting now to what we wanted to do, derivatives, we have here:

THEOREM 11.8. *We can talk about the derivatives of distributions, given by*

$$\varphi'(f) = -\varphi(f')$$

and with this notion in hand, the following happen:

- (1) *When φ is a usual differentiable function, φ' is the usual derivative.*
- (2) *For the basic step function we have $H' = \delta_0$, as previously advertised.*
- (3) *In fact, for a function $\varphi = g$ with jumps at $\{x_i\}$, we have $\varphi' = g' + \sum_i J_g(x_i)\delta_{x_i}$.*

PROOF. The first assertion, which by the way explains the need for the $-$ sign, follows from the Leibnitz rule for derivatives. Regarding the second assertion, this follows from:

$$\begin{aligned} H'(f) &= -H(f') \\ &= -\int_0^\infty f'(x)dx \\ &= -f(\infty) + f(0) \\ &= f(0) \\ &= \delta_0(f) \end{aligned}$$

As for the third assertion, which generalizes (1,2), we will leave this as an exercise. \square

Summarizing, quite interesting all this. Many other things can be said about distributions, notably with a convolution theory for them too, generalizing what we know about functions and Dirac masses, and we will leave some thinking or learning here as an exercise. Importantly, all this generalizes to several dimensions too, by making a use of the Jacobian where needed, and again, we will leave some learning here as an exercise.

11b. Complex analysis

Time now for some complex analysis. You certainly know about holomorphic functions, and perhaps about harmonic functions too, and time now to review all this. At the beginning, we have the Cauchy formula for complex functions, as follows:

THEOREM 11.9. *The Cauchy formula, namely*

$$f(x) = \frac{1}{2\pi i} \int_{\gamma} \frac{f(y)}{y - x} dy$$

holds for any holomorphic function $f : X \rightarrow \mathbb{C}$.

PROOF. This is something quite tricky, that you surely know, and as a matter of reminding this, let us verify it for polynomials $P \in \mathbb{C}[X]$. By linearity we can assume $P(x) = x^n$. Also, by a cut-and-paste argument, we can assume that we are on a circle:

$$\gamma : [0, 2\pi] \rightarrow \mathbb{C} \quad , \quad \gamma(t) = x + re^{it}$$

The point now is that for a monomial $P(x) = x^n$ we have:

$$\begin{aligned} \int_{\gamma} y^n dy &= \frac{1}{2\pi} \int_0^{2\pi} (x + re^{it})^n dt \\ &= \frac{1}{2\pi} \int_0^{2\pi} \sum_{k=0}^n \binom{n}{k} x^k (re^{it})^{n-k} dt \\ &= \sum_{k=0}^n \binom{n}{k} x^k r^{n-k} \frac{1}{2\pi} \int_0^{2\pi} e^{i(n-k)t} dt \\ &= \sum_{k=0}^n \binom{n}{k} x^k r^{n-k} \delta_{kn} \\ &= x^n \end{aligned}$$

Here we have used the following key identity, valid for any exponent $m \in \mathbb{Z}$:

$$\begin{aligned} \frac{1}{2\pi} \int_0^{2\pi} e^{imt} dt &= \frac{1}{2\pi} \int_0^{2\pi} \cos(mt) + i \sin(mt) dt \\ &= \delta_{m0} + i \cdot 0 \\ &= \delta_{m0} \end{aligned}$$

By using now the above computation, we obtain the following formula:

$$\begin{aligned} \int_{\gamma} \frac{y^n}{y-x} dy &= \int_0^{2\pi} \frac{(x + re^{it})^n}{re^{it}} rie^{it} dt \\ &= i \int_0^{2\pi} (x + re^{it})^n dt \\ &= i \cdot 2\pi x^n \end{aligned}$$

Thus, the Cauchy formula holds indeed for polynomials $P \in \mathbb{C}[X]$. In general, the proof is substantially more complicated, but you surely know it, say from Rudin [76]. \square

As a main application of the Cauchy formula, we have:

THEOREM 11.10. *The following conditions are equivalent, for a function $f : X \rightarrow \mathbb{C}$:*

- (1) *f is holomorphic.*
- (2) *f is infinitely differentiable.*
- (3) *f is analytic.*
- (4) *The Cauchy formula holds for f .*

PROOF. This is routine from what we have, the idea being as follows:

- (1) \implies (4) is non-trivial, but we know this from Theorem 11.9.
- (4) \implies (3) is trivial, by expanding the series in the Cauchy formula.
- (3) \implies (2) \implies (1) are both elementary, and good exercises for you. \square

As another application of the Cauchy formula, we have:

THEOREM 11.11. *Any holomorphic function $f : X \rightarrow \mathbb{C}$ satisfies the maximum principle, in the sense that given a domain D , with boundary γ , we have:*

$$\exists x \in \gamma \quad , \quad |f(x)| = \max_{y \in D} |f(y)|$$

That is, the maximum of $|f|$ over a domain is attained on its boundary.

PROOF. This follows indeed from the Cauchy formula. Observe that the converse is not true, for instance because functions like \bar{x} satisfy too the maximum principle. We will be back to this in a moment, when talking about harmonic functions. \square

Conceptually, a good explanation for the fact that the maximum principle holds could be that the values of our function inside a disk can be recovered from its values on the boundary. And fortunately, this is indeed the case, and we have:

THEOREM 11.12. *Given an holomorphic function $f : X \rightarrow \mathbb{C}$, and a disk D , with boundary γ , the following happen:*

- (1) *f satisfies the plain mean value formula $f(x) = \int_D f(y) dy$.*
- (2) *f satisfies the boundary mean value formula $f(x) = \int_{\gamma} f(y) dy$.*

PROOF. As usual, this follows from the Cauchy formula, which is something truly powerful, and with working out the details being a good exercise for you. \square

Finally, as yet another application of the Cauchy formula, which is something nice-looking and conceptual, we have the following statement, called Liouville theorem:

THEOREM 11.13. *An entire, bounded holomorphic function*

$$f : \mathbb{C} \rightarrow \mathbb{C} \quad , \quad |f| \leq M$$

must be constant. In particular, knowing $f \rightarrow 0$ with $z \rightarrow \infty$ gives $f = 0$.

PROOF. This follows as usual from the Cauchy formula, namely:

$$f(x) = \frac{1}{2\pi i} \int_{\gamma} \frac{f(y)}{y - x} dy$$

Alternatively, we can view this as a consequence of Theorem 11.12, because given two points $x \neq y$, we can view the values of f at these points as averages over big disks centered at these points, say $D = D_x(R)$ and $E = D_y(R)$, with $R \gg 0$:

$$f(x) = \int_D f(z) dz \quad , \quad f(y) = \int_E f(z) dz$$

Indeed, the point is that when the radius goes to ∞ , these averages tend to be equal, and so we have $f(x) \simeq f(y)$, which gives $f(x) = f(y)$ in the limit. \square

In order to reach to a continuation of this, let us formulate the following definition:

DEFINITION 11.14. *The Laplace operator in 2 dimensions is:*

$$\Delta f = \frac{d^2 f}{dx^2} + \frac{d^2 f}{dy^2}$$

A function $f : \mathbb{R}^2 \rightarrow \mathbb{C}$ satisfying $\Delta f = 0$ will be called harmonic.

Here the formula of Δ is the one coming from physics, and more on this later. As for the notion of harmonic function, this is something quite natural. Indeed, we can think of Δ as being a linear operator on the space of functions $f : \mathbb{R}^2 \rightarrow \mathbb{C}$, and previous experience with linear operators, and linear algebra in general, suggests looking first into the eigenvectors of Δ . But the simplest such eigenvectors are those corresponding to the eigenvalue $\lambda = 0$, and these are exactly our harmonic functions, satisfying:

$$\Delta f = 0$$

Getting now to more concrete things, and to some mathematics that we can do, using our knowledge, let us try to find the functions $f : \mathbb{R}^2 \rightarrow \mathbb{C}$ which are harmonic. And here, as a good surprise, we have an interesting link with the holomorphic functions:

THEOREM 11.15. *Any holomorphic function $f : \mathbb{C} \rightarrow \mathbb{C}$, when regarded as real function*

$$f : \mathbb{R}^2 \rightarrow \mathbb{C}$$

is harmonic. Moreover, the conjugates \bar{f} of holomorphic functions are harmonic too.

PROOF. The first assertion follows from the following computation, for the power functions $f(z) = z^n$, with the usual notation $z = x + iy$:

$$\begin{aligned} \Delta z^n &= \frac{d^2 z^n}{dx^2} + \frac{d^2 z^n}{dy^2} \\ &= \frac{d(nz^{n-1})}{dx} + \frac{d(inz^{n-1})}{dy} \\ &= n(n-1)z^{n-2} - n(n-1)z^{n-2} \\ &= 0 \end{aligned}$$

As for the second assertion, this follows from $\Delta \bar{f} = \overline{\Delta f}$, which is clear from definitions, and which shows that if f is harmonic, then so is its conjugate \bar{f} . \square

In view of the above, we can expect the harmonic functions to share many interesting properties with the holomorphic functions. We will be back to this, which holds indeed, in a moment. Moving now to arbitrary $N \in \mathbb{N}$ dimensions, let us first formulate:

DEFINITION 11.16. *The Laplace operator in N dimensions is:*

$$\Delta f = \sum_{i=1}^N \frac{d^2 f}{dx_i^2}$$

A function $f : \mathbb{R}^N \rightarrow \mathbb{C}$ satisfying $\Delta f = 0$ will be called harmonic.

Again, these are notions coming from physics, and more on this later, at the end of the present chapter, where we will discuss the wave equation, involving Δ .

In view of Theorem 11.15, and in analogy with the theory of holomorphic functions, that we know well from the above, we have the following result:

THEOREM 11.17. *The harmonic functions in N dimensions obey to the same general principles as the holomorphic functions, namely:*

- (1) *The plain mean value formula.*
- (2) *The boundary mean value formula.*
- (3) *The maximum modulus principle.*
- (4) *The Liouville theorem.*

PROOF. This is something quite tricky, the idea being as follows:

(1) Regarding the plain mean value formula, here the statement is that given an harmonic function $f : X \rightarrow \mathbb{C}$, and a ball B , the following happens:

$$f(x) = \int_B f(y) dy$$

In order to prove this, we can assume that B is centered at 0, of radius $r > 0$. If we denote by χ_r the characteristic function of this ball, nomalized as to integrate up to 1, in terms of the usual convolution operation for functions, we want to prove that we have:

$$f = f * \chi_r$$

For doing so, let us pick a number $0 < s < r$, and a solution w of the following equation, supported on B , which can be constructed explicitly:

$$\Delta w = \chi_r - \chi_s$$

By using the standard properties of the convolution operation $*$, we have:

$$\begin{aligned} f * \chi_r - f * \chi_s &= f * (\chi_r - \chi_s) \\ &= f * \Delta w \\ &= \Delta f * w \\ &= 0 \end{aligned}$$

Thus $f * \chi_r = f * \chi_s$, and by letting now $s \rightarrow 0$, we get $f * \chi_r = f$, as desired.

(2) Regarding the boundary mean value formula, here the statement is that given an harmonic function $f : X \rightarrow \mathbb{C}$, and a ball B , with boundary γ , the following happens:

$$f(x) = \int_{\gamma} f(y) dy$$

But this follows as a consequence of the plain mean value formula in (1), with our two mean value formulae, the one there and the one here, being in fact equivalent, by using annuli and radial integration for the proof of the equivalence, in the obvious way.

(3) Regarding the maximum modulus principle, the statement here is that any holomorphic function $f : X \rightarrow \mathbb{C}$ has the property that the maximum of $|f|$ over a domain is attained on its boundary. That is, given a domain D , with boundary γ , we have:

$$\exists x \in \gamma \quad , \quad |f(x)| = \max_{y \in D} |f(y)|$$

But this is something which follows again from the mean value formula in (1), first for the balls, and then in general, by using a standard division argument.

(4) Finally, regarding the Liouville theorem, we can view this as a consequence of (1), because given two points $x \neq y$, we can view the values of f at these points as averages

over big balls centered at these points, say $B = B_x(R)$ and $C = B_y(R)$, with $R \gg 0$:

$$f(x) = \int_B f(z)dz \quad , \quad f(y) = \int_C f(z)dz$$

Indeed, the point is that when the radius goes to ∞ , these averages tend to be equal, and so we have $f(x) \simeq f(y)$, which gives $f(x) = f(y)$ in the limit, as desired. \square

In addition to the above, let us mention too that at $N = 2$, locally, the real harmonic functions are known to be the real parts of holomorphic functions. Also, there is an interesting relation between harmonic functions and soap films. Exercise of course for you, to learn more about this, with all this being first class mathematics and physics.

11c. Fourier analysis

Let us discuss now the construction and main properties of the Fourier transform, which is the main tool in analysis, and mathematics in general. We first have:

DEFINITION 11.18. *Given $f \in L^1(\mathbb{R})$, we define a function $\widehat{f} : \mathbb{R} \rightarrow \mathbb{C}$ by*

$$\widehat{f}(\xi) = \int_{\mathbb{R}} e^{ix\xi} f(x)dx$$

and call it Fourier transform of f .

As a first observation, even if f is a real function, \widehat{f} is a complex function, which is not necessarily real. Also, \widehat{f} is obviously well-defined, because $f \in L^1(\mathbb{R})$ and $|e^{ix\xi}| = 1$. Also, the condition $f \in L^1(\mathbb{R})$ is basically needed for constructing \widehat{f} , because:

$$\widehat{f}(0) = \int_{\mathbb{R}} f(x)dx$$

Generally speaking, the Fourier transform is there for helping with various computations, with the above formula $\widehat{f}(0) = \int f$ being something quite illustrating. Here are some basic properties of the Fourier transform, all providing some good motivations:

PROPOSITION 11.19. *The Fourier transform has the following properties:*

- (1) *Linearity: $\widehat{f+g} = \widehat{f} + \widehat{g}$, $\widehat{\lambda f} = \lambda \widehat{f}$.*
- (2) *Regularity: \widehat{f} is continuous and bounded.*
- (3) *If f is even then \widehat{f} is even.*
- (4) *If f is odd then \widehat{f} is odd.*

PROOF. These results are all elementary, as follows:

- (1) The linearity formulae are indeed both clear from definitions.

(2) The continuity of \widehat{f} follows indeed from:

$$\begin{aligned} |\widehat{f}(\xi + \varepsilon) - \widehat{f}(\xi)| &\leq \int_{\mathbb{R}} |(e^{ix(\xi+\varepsilon)} - e^{ix\xi})f(x)| dx \\ &= \int_{\mathbb{R}} |e^{ix\xi}(e^{ix\varepsilon} - 1)f(x)| dx \\ &\leq |e^{ix\varepsilon} - 1| \int_{\mathbb{R}} |f| \end{aligned}$$

As for the boundedness of \widehat{f} , this is clear as well.

(3) This follows from the following computation, assuming that f is even:

$$\begin{aligned} \widehat{f}(-\xi) &= \int_{\mathbb{R}} e^{-ix\xi} f(x) dx \\ &= \int_{\mathbb{R}} e^{ix\xi} f(-x) dx \\ &= \int_{\mathbb{R}} e^{ix\xi} f(x) dx \\ &= \widehat{f}(\xi) \end{aligned}$$

(4) The proof here is similar to the proof of (3), by changing some signs. \square

We will be back to more theory in a moment, but let us explore now the examples. Here are some basic computations of Fourier transforms:

PROPOSITION 11.20. *We have the following Fourier transform formulae,*

$$\begin{aligned} f = \chi_{[-a,a]} &\implies \widehat{f}(\xi) = \frac{2 \sin(a\xi)}{\xi} \\ f = e^{-ax} \chi_{[0,\infty]}(x) &\implies \widehat{f}(\xi) = \frac{1}{a - i\xi} \\ f = e^{ax} \chi_{[-\infty,0]}(x) &\implies \widehat{f}(\xi) = \frac{1}{a + i\xi} \\ f = e^{-a|x|} &\implies \widehat{f}(\xi) = \frac{2a}{a^2 + \xi^2} \\ f = \operatorname{sgn}(x) e^{-a|x|} &\implies \widehat{f}(\xi) = \frac{2i\xi}{a^2 + \xi^2} \end{aligned}$$

valid for any number $a > 0$.

PROOF. All this follows from some calculus, as follows:

(1) In what regards first formula, assuming $f = \chi_{[-a,a]}$, we have, by using the fact that $\sin(x\xi)$ is an odd function, whose integral vanishes on centered intervals:

$$\begin{aligned}
 \widehat{f}(\xi) &= \int_{-a}^a e^{ix\xi} dx \\
 &= \int_{-a}^a \cos(x\xi) dx + i \int_{-a}^a \sin(x\xi) dx \\
 &= \int_{-a}^a \cos(x\xi) dx \\
 &= \left[\frac{\sin(x\xi)}{\xi} \right]_{-a}^a \\
 &= \frac{2 \sin(a\xi)}{\xi}
 \end{aligned}$$

(2) With $f(x) = e^{-ax} \chi_{[0,\infty)}(x)$, the computation goes as follows:

$$\begin{aligned}
 \widehat{f}(\xi) &= \int_0^\infty e^{ix\xi - ax} dx \\
 &= \int_0^\infty e^{(i\xi - a)x} dx \\
 &= \left[\frac{e^{(i\xi - a)x}}{i\xi - a} \right]_0^\infty \\
 &= \frac{1}{a - i\xi}
 \end{aligned}$$

(3) Regarding the third formula, this follows from the second one, by using the following fact, generalizing the parity computations from Proposition 11.19:

$$F(x) = f(-x) \implies \widehat{F}(\xi) = \widehat{f}(-\xi)$$

(4) The last 2 formulae follow from what we have, by making sums and differences, and the linearity properties of the Fourier transform, from Proposition 11.19. \square

We will see many other examples, in what follows. Getting back now to theory, we have the following result, adding to the various general properties in Proposition 11.19, and providing more motivations for the Fourier transform:

PROPOSITION 11.21. *Given $f, g \in L^1(\mathbb{R})$ we have $\widehat{fg}, f\widehat{g} \in L^1(\mathbb{R})$ and*

$$\int_{\mathbb{R}} f(\xi) \widehat{g}(\xi) d\xi = \int_{\mathbb{R}} \widehat{f}(x) g(x) dx$$

called “exchange of hat” formula.

PROOF. Regarding the fact that we have indeed $\widehat{fg}, f\widehat{g} \in L^1(\mathbb{R})$, this is actually a bit non-trivial, but we will be back to this later. Assuming this, we have:

$$\int_{\mathbb{R}} f(\xi)\widehat{g}(\xi)d\xi = \int_{\mathbb{R}} \int_{\mathbb{R}} f(\xi)e^{ix\xi}g(x)dx d\xi$$

On the other hand, we have as well the following formula:

$$\int_{\mathbb{R}} \widehat{f}(x)g(x)dx = \int_{\mathbb{R}} \int_{\mathbb{R}} e^{ix\xi}f(x)g(\xi)dx d\xi$$

Thus, with $x \leftrightarrow \xi$, we are led to the formula in the statement. \square

As a key result now, showing the power of the Fourier transform, we have:

THEOREM 11.22. *Given $f : \mathbb{R} \rightarrow \mathbb{C}$ such that $f, f' \in L^1(\mathbb{R})$, we have:*

$$\widehat{f'}(\xi) = -i\xi\widehat{f}(\xi)$$

More generally, assuming $f, f', f'', \dots, f^{(n)} \in L^1(\mathbb{R})$, we have

$$\widehat{f^{(k)}}(\xi) = (-i\xi)^k \widehat{f}(\xi)$$

for any $k = 1, 2, \dots, n$.

PROOF. These results follow by doing a partial integration, as follows:

(1) Assuming that $f : \mathbb{R} \rightarrow \mathbb{C}$ has compact support, we have indeed:

$$\begin{aligned} \widehat{f'}(\xi) &= \int_{\mathbb{R}} e^{ix\xi} f'(x) dx \\ &= - \int_{\mathbb{R}} i\xi e^{ix\xi} f(x) dx \\ &= -i\xi \int_{\mathbb{R}} e^{ix\xi} f(x) dx \\ &= -i\xi \widehat{f}(\xi) \end{aligned}$$

(2) Regarding the higher derivatives, the formula here follows by recurrence. \square

Importantly, we have a converse statement as well, as follows:

THEOREM 11.23. *Assuming that $f \in L^1(\mathbb{R})$ is such that $F(x) = xf(x)$ belongs to $L^1(\mathbb{R})$ too, the function \widehat{f} is differentiable, with derivative given by:*

$$(\widehat{f})'(\xi) = i\widehat{F}(\xi)$$

More generally, if $F_k(x) = x^k f(x)$ belongs to $L^1(\mathbb{R})$, for $k = 0, 1, \dots, n$, we have

$$(\widehat{f})^{(k)}(\xi) = i^k \widehat{F_k}(\xi)$$

for any $k = 1, 2, \dots, n$.

PROOF. These results are both elementary, as follows:

(1) Regarding the first assertion, the computation here is as follows:

$$\begin{aligned}
 (\widehat{f})'(\xi) &= \frac{d}{d\xi} \int_{\mathbb{R}} e^{ix\xi} f(x) dx \\
 &= \int_{\mathbb{R}} ix e^{ix\xi} f(x) dx \\
 &= i \int_{\mathbb{R}} e^{ix\xi} x f(x) dx \\
 &= i \widehat{F}(\xi)
 \end{aligned}$$

(2) As for the second assertion, this follows from the first one, by recurrence. \square

As a conclusion to all this, we are on a good way with our theory, and we have:

CONCLUSION 11.24. *Modulo normalization factors, the Fourier transform converts the derivatives into multiplications by the variable, and vice versa.*

And isn't this interesting, because isn't computing derivatives a difficult task. Here is now another useful result, of the same type, this time regarding convolutions:

THEOREM 11.25. *Assuming $f, g \in L^1(\mathbb{R})$, the following happens:*

$$\widehat{f * g} = \widehat{f} \cdot \widehat{g}$$

Moreover, under suitable assumptions, the formula $\widehat{fg} = \widehat{f} * \widehat{g}$ holds too.

PROOF. This is something quite subtle, the idea being as follows:

(1) Regarding the first assertion, this is something elementary, as follows:

$$\begin{aligned}
 \widehat{f * g}(\xi) &= \int_{\mathbb{R}} e^{ix\xi} (f * g)(x) dx \\
 &= \int_{\mathbb{R}} \int_{\mathbb{R}} e^{ix\xi} f(x-y) g(y) dx dy \\
 &= \int_{\mathbb{R}} e^{iy\xi} \left(\int_{\mathbb{R}} e^{i(x-y)\xi} f(x-y) dx \right) g(y) dy \\
 &= \int_{\mathbb{R}} e^{iy\xi} \left(\int_{\mathbb{R}} e^{it\xi} f(t) dt \right) g(y) dy \\
 &= \int_{\mathbb{R}} e^{iy\xi} \widehat{f}(\xi) g(y) dy \\
 &= \widehat{f}(\xi) \widehat{g}(\xi)
 \end{aligned}$$

(2) As for the second assertion, this is something more tricky, and we will be back to this later. In the meantime, here is however some sort of proof, not very honest:

$$\begin{aligned}
 (\widehat{f} * \widehat{g})(\xi) &= \int_{\mathbb{R}} \widehat{f}(\xi - \eta) \widehat{g}(\eta) d\eta \\
 &= \int_{\mathbb{R}} \int_{\mathbb{R}} \int_{\mathbb{R}} e^{ix(\xi - \eta)} f(x) e^{iy\eta} g(y) dx dy d\eta \\
 &= \int_{\mathbb{R}} \int_{\mathbb{R}} \int_{\mathbb{R}} e^{ix\eta} e^{i(y-x)\eta} f(x) g(y) dx dy d\eta \\
 &= \int_{\mathbb{R}} e^{ix\eta} f(x) g(x) dx \\
 &= \widehat{fg}(\eta)
 \end{aligned}$$

To be more precise, the point here is that we can pass from the triple to the single integral by arguing that “we must have $x = y$ ”. We will be back to this later. \square

As an updated conclusion to all this, we have, modulo a few bugs, to be fixed:

CONCLUSION 11.26. *The Fourier transform converts the derivatives into multiplications by the variable, and convolutions into products, and vice versa.*

We will see applications of this later, after developing some more general theory. So, let us develop now more theory for the Fourier transform. We first have:

THEOREM 11.27. *Given $f \in L^1(\mathbb{R})$, its Fourier transform satisfies*

$$\lim_{\xi \rightarrow \pm\infty} \widehat{f}(\xi) = 0$$

called Riemann-Lebesgue property of \widehat{f} .

PROOF. This is something quite technical, as follows:

(1) Given a function $f : \mathbb{R} \rightarrow \mathbb{C}$ and a number $y \in \mathbb{R}$, let us set:

$$f_y(x) = f(x - y)$$

Our claim is then is that if $f \in L^p(\mathbb{R})$, then the following function is uniformly continuous, with respect to the usual p -norm on the right:

$$\mathbb{R} \rightarrow L^p(\mathbb{R}) \quad , \quad y \rightarrow f_y$$

(2) In order to prove this, fix $\varepsilon > 0$. Since $f \in L^p(\mathbb{R})$, we can find a function of type $g : [-K, K] \rightarrow \mathbb{C}$ which is continuous, such that:

$$\|f - g\|_p < \varepsilon$$

Now since g is uniformly continuous, we can find $\delta \in (0, K)$ such that:

$$|s - t| < \delta \implies |g(s) - g(t)| < (3K)^{-1/p} \varepsilon$$

But this shows that we have the following estimate:

$$\begin{aligned} \|g_s - g_t\|_p &= \left(\int_{\mathbb{R}} |g(x-s) - g(x-t)|^p dx \right)^{1/p} \\ &< [(3K)^{-1} \varepsilon^p (2k + \delta)]^{1/p} \\ &< \varepsilon \end{aligned}$$

By using now the formula $\|f\|_p = \|f_s\|_p$, which is clear, we obtain:

$$\begin{aligned} \|f_s - f_t\|_p &\leq \|f_s - g_s\|_p + \|g_s - g_t\|_p + \|g_t - f_t\|_p \\ &< \varepsilon + \varepsilon + \varepsilon \\ &= 3\varepsilon \end{aligned}$$

But this being true for any $|s - t| < \delta$, we have proved our claim.

(3) Let us prove now the Riemann-Lebesgue property of \widehat{f} , as formulated in the statement. By using $e^{\pi i} = -1$, and the change of variables $x \rightarrow x - \pi/\xi$, we have:

$$\begin{aligned} \widehat{f}(\xi) &= \int_{\mathbb{R}} e^{ix\xi} f(x) dx \\ &= - \int_{\mathbb{R}} e^{ix\xi} e^{\pi i} f(x) dx \\ &= - \int_{\mathbb{R}} e^{i\xi(x+\pi/\xi)} f(x) dx \\ &= - \int_{\mathbb{R}} e^{ix\xi} f\left(x - \frac{\pi}{\xi}\right) dx \end{aligned}$$

On the other hand, we have as well the following formula:

$$\widehat{f}(\xi) = \int_{\mathbb{R}} e^{ix\xi} f(x) dx$$

Thus by summing, we obtain the following formula:

$$2\widehat{f}(\xi) = \int_{\mathbb{R}} e^{ix\xi} \left(f(x) - f\left(x - \frac{\pi}{\xi}\right) \right) dx$$

But this gives the following estimate, with notations from (1):

$$2|\widehat{f}(\xi)| \leq \|f - f_{\pi/\xi}\|_1$$

Since by (1) this goes to 0 with $\xi \rightarrow \pm\infty$, this gives the result. \square

Quite remarkably, and as a main result now regarding Fourier transforms, a function $f : \mathbb{R} \rightarrow \mathbb{C}$ can be recovered from its Fourier transform $\widehat{f} : \mathbb{R} \rightarrow \mathbb{C}$, as follows:

THEOREM 11.28. *Assuming $f, \widehat{f} \in L^1(\mathbb{R})$, we have*

$$f(x) = \int_{\mathbb{R}} e^{-ix\xi} \widehat{f}(\xi) d\xi$$

almost everywhere, called Fourier inversion formula.

PROOF. This is something quite tricky, due to the fact that a direct attempt by double integration fails. Consider the following function, depending on a parameter $\lambda > 0$:

$$\varphi_\lambda(x) = \int_{\mathbb{R}} e^{-ix\xi - \lambda|\xi|} d\xi$$

We have then the following computation:

$$\begin{aligned} (f * \varphi_\lambda)(x) &= \int_{\mathbb{R}} f(x-y) \varphi_\lambda(y) dy \\ &= \int_{\mathbb{R}} \int_{\mathbb{R}} f(x-y) e^{-iy\xi - \lambda|\xi|} d\xi dy \\ &= \int_{\mathbb{R}} e^{-\lambda|\xi|} \left(\int_{\mathbb{R}} f(x-y) e^{-iy\xi} dy \right) d\xi \\ &= \int_{\mathbb{R}} e^{-\lambda|\xi|} e^{-ix\xi} \widehat{f}(\xi) d\xi \end{aligned}$$

By letting now $\lambda \rightarrow 0$, we obtain from this the following formula:

$$\lim_{\lambda \rightarrow 0} (f * \varphi_\lambda)(x) = \int_{\mathbb{R}} e^{-ix\xi} \widehat{f}(\xi) d\xi$$

On the other hand, by using Theorem 11.27 we obtain that, almost everywhere:

$$\lim_{\lambda \rightarrow 0} (f * \varphi_\lambda)(x) = f(x)$$

Thus, we are led to the conclusion in the statement. \square

There are many more things that can be said about Fourier transforms, a key result being the Plancherel formula, over $L^2(\mathbb{R})$. Also, we can talk about the Fourier transform over the Schwartz space \mathcal{S} . Also, we can talk about other types of Fourier transforms. As usual, if very interested in mathematical analysis, have a look at all this.

11d. Into the waves

We have now enough accumulated material for talking about waves. Hang on, tough mathematics and physics to come next. To start with, waves appear as follows:

THEOREM 11.29. *The wave equation in \mathbb{R}^N is*

$$\ddot{\varphi} = v^2 \Delta \varphi$$

where $v > 0$ is the propagation speed.

PROOF. The equation in the statement is of course what comes out of experiments. However, allowing us a bit of imagination, and trust in this imagination, we can mathematically “prove” this equation, by discretizing, as follows:

(1) Let us first consider the 1D case. In order to understand the propagation of waves, we will model \mathbb{R} as a network of balls, with springs between them, as follows:

$$\cdots \times \times \times \bullet \times \times \times \bullet \times \times \times \bullet \times \times \times \bullet \times \times \times \bullet \times \times \times \cdots$$

Now let us send an impulse, and see how the balls will be moving. For this purpose, we zoom on one ball. The situation here is as follows, l being the spring length:

$$\cdots \cdots \cdots \bullet_{\varphi(x-l)} \times \times \times \bullet_{\varphi(x)} \times \times \times \bullet_{\varphi(x+l)} \cdots \cdots \cdots$$

We have two forces acting at x . First is the Newton motion force, mass times acceleration, which is as follows, with m being the mass of each ball:

$$F_n = m \cdot \ddot{\varphi}(x)$$

And second is the Hooke force, displacement of the spring, times spring constant. Since we have two springs at x , this is as follows, k being the spring constant:

$$\begin{aligned} F_h &= F_h^r - F_h^l \\ &= k(\varphi(x+l) - \varphi(x)) - k(\varphi(x) - \varphi(x-l)) \\ &= k(\varphi(x+l) - 2\varphi(x) + \varphi(x-l)) \end{aligned}$$

We conclude that the equation of motion, in our model, is as follows:

$$m \cdot \ddot{\varphi}(x) = k(\varphi(x+l) - 2\varphi(x) + \varphi(x-l))$$

(2) Now let us take the limit of our model, as to reach to continuum. For this purpose we will assume that our system consists of $B \gg 0$ balls, having a total mass M , and spanning a total distance L . Thus, our previous infinitesimal parameters are as follows, with K being the spring constant of the total system, which is of course lower than k :

$$m = \frac{M}{B} \quad , \quad k = KB \quad , \quad l = \frac{L}{B}$$

With these changes, our equation of motion found in (1) reads:

$$\ddot{\varphi}(x) = \frac{KB^2}{M}(\varphi(x+l) - 2\varphi(x) + \varphi(x-l))$$

Now observe that this equation can be written, more conveniently, as follows:

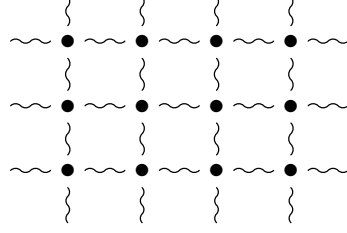
$$\ddot{\varphi}(x) = \frac{KL^2}{M} \cdot \frac{\varphi(x+l) - 2\varphi(x) + \varphi(x-l)}{l^2}$$

With $N \rightarrow \infty$, and therefore $l \rightarrow 0$, we obtain in this way:

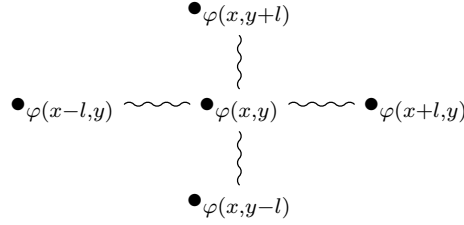
$$\ddot{\varphi}(x) = \frac{KL^2}{M} \cdot \frac{d^2\varphi}{dx^2}(x)$$

We are therefore led to the wave equation in the statement, which is $\ddot{\varphi} = v^2 \varphi''$ in our present $N = 1$ dimensional case, the propagation speed being $v = \sqrt{K/M} \cdot L$.

(3) In 2 dimensions now, the same argument carries on. Indeed, we can use here a lattice model as follows, with all the edges standing for small springs:



As before in one dimension, we send an impulse, and we zoom on one ball. The situation here is as follows, with l being the spring length:



We have two forces acting at (x, y) . First is the Newton motion force, mass times acceleration, which is as follows, with m being the mass of each ball:

$$F_n = m \cdot \ddot{\varphi}(x, y)$$

And second is the Hooke force, displacement of the spring, times spring constant. Since we have four springs at (x, y) , this is as follows, k being the spring constant:

$$\begin{aligned} F_h &= F_h^r - F_h^l + F_h^u - F_h^d \\ &= k(\varphi(x+l, y) - \varphi(x, y)) - k(\varphi(x, y) - \varphi(x-l, y)) \\ &+ k(\varphi(x, y+l) - \varphi(x, y)) - k(\varphi(x, y) - \varphi(x, y-l)) \\ &= k(\varphi(x+l, y) - 2\varphi(x, y) + \varphi(x-l, y)) \\ &+ k(\varphi(x, y+l) - 2\varphi(x, y) + \varphi(x, y-l)) \end{aligned}$$

We conclude that the equation of motion, in our model, is as follows:

$$\begin{aligned} m \cdot \ddot{\varphi}(x, y) &= k(\varphi(x+l, y) - 2\varphi(x, y) + \varphi(x-l, y)) \\ &+ k(\varphi(x, y+l) - 2\varphi(x, y) + \varphi(x, y-l)) \end{aligned}$$

(4) Now let us take the limit of our model, as to reach to continuum. For this purpose we will assume that our system consists of $B^2 \gg 0$ balls, having a total mass M , and

spanning a total area L^2 . Thus, our previous infinitesimal parameters are as follows, with K being the spring constant of the total system, taken to be equal to k :

$$m = \frac{M}{B^2} \quad , \quad k = K \quad , \quad l = \frac{L}{B}$$

With these changes, our equation of motion found in (3) reads:

$$\begin{aligned} \ddot{\varphi}(x, y) &= \frac{KB^2}{M} (\varphi(x+l, y) - 2\varphi(x, y) + \varphi(x-l, y)) \\ &+ \frac{KB^2}{M} (\varphi(x, y+l) - 2\varphi(x, y) + \varphi(x, y-l)) \end{aligned}$$

Now observe that this equation can be written, more conveniently, as follows:

$$\begin{aligned} \ddot{\varphi}(x, y) &= \frac{KL^2}{M} \times \frac{\varphi(x+l, y) - 2\varphi(x, y) + \varphi(x-l, y)}{l^2} \\ &+ \frac{KL^2}{M} \times \frac{\varphi(x, y+l) - 2\varphi(x, y) + \varphi(x, y-l)}{l^2} \end{aligned}$$

With $N \rightarrow \infty$, and therefore $l \rightarrow 0$, we obtain in this way:

$$\ddot{\varphi}(x, y) = \frac{KL^2}{M} \left(\frac{d^2\varphi}{dx^2} + \frac{d^2\varphi}{dy^2} \right) (x, y)$$

Thus, we are led in this way to the following wave equation in two dimensions, with $v = \sqrt{K/M} \cdot L$ being the propagation speed of our wave:

$$\ddot{\varphi}(x, y) = v^2 \left(\frac{d^2\varphi}{dx^2} + \frac{d^2\varphi}{dy^2} \right) (x, y)$$

But we recognize at right the Laplace operator, and we are done. As before in 1D, there is of course some discussion to be made here, arguing that our spring model in (3) is indeed the correct one. But do not worry, experiments confirm our findings.

(5) In 3 dimensions now, which is the case of the main interest, corresponding to our real-life world, the same argument carries over, and the wave equation is as follows:

$$\ddot{\varphi}(x, y, z) = v^2 \left(\frac{d^2\varphi}{dx^2} + \frac{d^2\varphi}{dy^2} + \frac{d^2\varphi}{dz^2} \right) (x, y, z)$$

(6) Finally, the same argument, namely a lattice model, carries on in arbitrary N dimensions, and the wave equation here is as follows:

$$\ddot{\varphi}(x_1, \dots, x_N) = v^2 \sum_{i=1}^N \frac{d^2\varphi}{dx_i^2} (x_1, \dots, x_N)$$

Thus, we are led to the conclusion in the statement. □

Getting now to mathematics, we will need a standard calculus result, as follows:

PROPOSITION 11.30. *The derivative of a function of type*

$$\varphi(x) = \int_{g(x)}^{h(x)} f(s)ds$$

is given by the formula $\varphi'(x) = f(h(x))h'(x) - f(g(x))g'(x)$.

PROOF. Consider a primitive of the function that we integrate, $F' = f$. We have:

$$\begin{aligned}\varphi(x) &= \int_{g(x)}^{h(x)} f(s)ds \\ &= \int_{g(x)}^{h(x)} F'(s)ds \\ &= F(h(x)) - F(g(x))\end{aligned}$$

By using now the chain rule for derivatives, we obtain from this:

$$\begin{aligned}\varphi'(x) &= F'(h(x))h'(x) - F'(g(x))g'(x) \\ &= f(h(x))h'(x) - f(g(x))g'(x)\end{aligned}$$

Thus, we are led to the formula in the statement. \square

Now back to the waves, the general result in 1D, due to d'Alembert, along with a little more, in relation with our lattice models above, is as follows:

THEOREM 11.31. *The solution of the 1D wave equation with initial value conditions $\varphi(x, 0) = f(x)$ and $\dot{\varphi}(x, 0) = g(x)$ is given by the d'Alembert formula, namely:*

$$\varphi(x, t) = \frac{f(x - vt) + f(x + vt)}{2} + \frac{1}{2v} \int_{x-vt}^{x+vt} g(s)ds$$

In the context of our lattice model discretizations, what happens is more or less that the above d'Alembert integral gets computed via Riemann sums.

PROOF. There are several things going on here, the idea being as follows:

(1) Let us first check that the d'Alembert solution is indeed a solution of the wave equation $\ddot{\varphi} = v^2\varphi''$. The first time derivative is computed as follows:

$$\dot{\varphi}(x, t) = \frac{-vf'(x - vt) + vf'(x + vt)}{2} + \frac{1}{2v}(vg(x + vt) + vg(x - vt))$$

The second time derivative is computed as follows:

$$\ddot{\varphi}(x, t) = \frac{v^2f''(x - vt) + v^2f''(x + vt)}{2} + \frac{vg'(x + vt) - vg'(x - vt)}{2}$$

Regarding now space derivatives, the first one is computed as follows:

$$\varphi'(x, t) = \frac{f'(x - vt) + f'(x + vt)}{2} + \frac{1}{2v}(g'(x + vt) - g'(x - vt))$$

As for the second space derivative, this is computed as follows:

$$\varphi''(x, t) = \frac{f''(x - vt) + f''(x + vt)}{2} + \frac{g''(x + vt) - g''(x - vt)}{2v}$$

Thus we have indeed $\ddot{\varphi} = v^2 \varphi''$. As for the initial conditions, $\varphi(x, 0) = f(x)$ is clear from our definition of φ , and $\dot{\varphi}(x, 0) = g(x)$ is clear from our above formula of $\dot{\varphi}$.

(2) Conversely now, we must show that our solution is unique, but instead of going here into abstract arguments, we will simply solve our equation, which among others will doublecheck out computations in (1). Let us make the following change of variables:

$$\xi = x - vt \quad , \quad \eta = x + vt$$

With this change of variables, which is quite tricky, mixing space and time variables, our wave equation $\ddot{\varphi} = v^2 \varphi''$ reformulates in a very simple way, as follows:

$$\frac{d^2 \varphi}{d\xi d\eta} = 0$$

But this latter equation tells us that our new ξ, η variables get separated, and we conclude from this that the solution must be of the following special form:

$$\varphi(x, t) = F(\xi) + G(\eta) = F(x - vt) + G(x + vt)$$

Now by taking into account the initial conditions $\varphi(x, 0) = f(x)$ and $\dot{\varphi}(x, 0) = g(x)$, and then integrating, we are led to the d'Alembert formula in the statement.

(3) In regards now with our discretization questions, by using a 1D lattice model with balls and springs as before, what happens to all the above is more or less that the above d'Alembert integral gets computed via Riemann sums, in our model, as stated. \square

At $N \geq 2$ and higher things are more complicated. In the case $N = 3$, which is the one that we are mainly interested in, the situation is as follows:

FACT 11.32. *The Laplace operator in spherical coordinates is given by the following formula, with our standard convention for the 3D spherical coordinates,*

$$\Delta = \frac{1}{r^2} \cdot \frac{d}{dr} \left(r^2 \cdot \frac{d}{dr} \right) + \frac{1}{r^2 \sin s} \cdot \frac{d}{ds} \left(\sin s \cdot \frac{d}{ds} \right) + \frac{1}{r^2 \sin^2 s} \cdot \frac{d^2}{dt^2}$$

and this allows to reformulate the 3D wave equation in spherical coordinates, and say a few things about it, say in relation with waves which propagate from a source.

And we will end our study here. Many interesting things that we have learned, in this chapter, and if interested in all this, a lot of exciting literature is waiting for you.

11e. Exercises

We had a quite broad chapter here, and as exercises on this, we have:

EXERCISE 11.33. *Learn more about the theory of distributions, on \mathbb{R} .*

EXERCISE 11.34. *Learn as well about distributions in higher dimensions.*

EXERCISE 11.35. *Review if needed the full proof of the Cauchy formula.*

EXERCISE 11.36. *Learn more about harmonic functions, in 2 or more dimensions.*

EXERCISE 11.37. *Learn the Plancherel formula, over $L^2(\mathbb{R})$, and its applications.*

EXERCISE 11.38. *Learn about the Fourier transform over the Schwartz space \mathcal{S} .*

EXERCISE 11.39. *Learn about the other possible types of Fourier transforms.*

EXERCISE 11.40. *Read some physics, on the wave equation in 2D, and 3D.*

As bonus exercise, now that we know about waves, learn about the heat equation too.

CHAPTER 12

Linear operators

12a. Operator theory

We discuss in this chapter the diagonalization of the linear operators $T : H \rightarrow H$ over a Hilbert space H , usually taken to be separable. Our motivation comes from both mathematics, because what we will be doing here will be a natural infinite dimensional extension of basic linear algebra, and physics, more specifically quantum mechanics.

Let us begin our study with a number of general results regarding the operators that we are targeting. Regarding the self-adjoint operators, we have the following result:

THEOREM 12.1. *For an operator $T \in B(H)$, the following conditions are equivalent, and if they are satisfied, we call T self-adjoint:*

- (1) $T = T^*$.
- (2) $\langle Tx, x \rangle \in \mathbb{R}$.

In finite dimensions, we recover in this way the usual self-adjointness notion.

PROOF. There are several assertions here, the idea being as follows:

- (1) \implies (2) This implication is clear, because we have:

$$\begin{aligned} \overline{\langle Tx, x \rangle} &= \langle x, Tx \rangle \\ &= \langle T^*x, x \rangle \\ &= \langle Tx, x \rangle \end{aligned}$$

(2) \implies (1) In order to prove this, observe that the beginning of the above computation shows that, when assuming $\langle Tx, x \rangle \in \mathbb{R}$, the following happens:

$$\langle Tx, x \rangle = \langle T^*x, x \rangle$$

Thus, in terms of the operator $S = T - T^*$, we have:

$$\langle Sx, x \rangle = 0$$

In order to finish, we use a polarization trick. We have the following formula:

$$\langle S(x+y), x+y \rangle = \langle Sx, x \rangle + \langle Sy, y \rangle + \langle Sx, y \rangle + \langle Sy, x \rangle$$

Since the first 3 terms vanish, the sum of the 2 last terms vanishes too. But, by using $S^* = -S$, coming from $S = T - T^*$, we can process this latter vanishing as follows:

$$\begin{aligned} \langle Sx, y \rangle &= -\langle Sy, x \rangle \\ &= \langle y, Sx \rangle \\ &= \overline{\langle Sx, y \rangle} \end{aligned}$$

Thus we must have $\langle Sx, y \rangle \in \mathbb{R}$, and with $y \rightarrow iy$ we obtain $\langle Sx, y \rangle \in i\mathbb{R}$ too, and so $\langle Sx, y \rangle = 0$. Thus $S = 0$, which gives $T = T^*$, as desired.

(3) Finally, in what regards finite dimensions, here the condition $T = T^*$ corresponds to the usual self-adjointness condition $M = M^*$ at the level of the associated matrices. \square

Regarding the normal operators, which are more general, we have:

THEOREM 12.2. *For an operator $T \in B(H)$, the following conditions are equivalent, and if they are satisfied, we call T normal:*

- (1) $TT^* = T^*T$.
- (2) $\|Tx\| = \|T^*x\|$.

In finite dimensions, we recover in this way the usual normality notion.

PROOF. There are several assertions here, the idea being as follows:

(1) \implies (2) This is clear, due to the following computation:

$$\begin{aligned} \|Tx\|^2 &= \langle Tx, Tx \rangle \\ &= \langle T^*Tx, x \rangle \\ &= \langle TT^*x, x \rangle \\ &= \langle T^*x, T^*x \rangle \\ &= \|T^*x\|^2 \end{aligned}$$

(2) \implies (1) This is clear as well, because the above computation shows that, when assuming $\|Tx\| = \|T^*x\|$, the following happens:

$$\langle TT^*x, x \rangle = \langle T^*Tx, x \rangle$$

Thus, in terms of the operator $S = TT^* - T^*T$, we have:

$$\langle Sx, x \rangle = 0$$

In order to finish, we use a polarization trick. We have the following formula:

$$\langle S(x+y), x+y \rangle = \langle Sx, x \rangle + \langle Sy, y \rangle + \langle Sx, y \rangle + \langle Sy, x \rangle$$

Since the first 3 terms vanish, the sum of the 2 last terms vanishes too. But, by using $S = S^*$, coming from $S = TT^* - T^*T$, we can process this latter vanishing as follows:

$$\begin{aligned} \langle Sx, y \rangle &= -\langle Sy, x \rangle \\ &= -\langle y, Sx \rangle \\ &= -\overline{\langle Sx, y \rangle} \end{aligned}$$

Thus we must have $\langle Sx, y \rangle \in i\mathbb{R}$, and with $y \rightarrow iy$ we obtain $\langle Sx, y \rangle \in \mathbb{R}$ too, and so $\langle Sx, y \rangle = 0$. Thus $S = 0$, which gives $TT^* = T^*T$, as desired.

(3) Finally, in what regards finite dimensions, or more generally the case where our Hilbert space comes with a basis, $H = l^2(I)$, here the condition $TT^* = T^*T$ corresponds to the usual normality condition $MM^* = M^*M$ at the level of the associated matrices. \square

Getting now to diagonalization work for the linear operators, as a first observation, we can talk about eigenvalues and eigenvectors of such operators, as follows:

DEFINITION 12.3. *Given an operator $T \in B(H)$, assuming that we have*

$$Tx = \lambda x$$

we say that $x \in H$ is an eigenvector of T , with eigenvalue $\lambda \in \mathbb{C}$.

We know many things about eigenvalues and eigenvectors, in the finite dimensional case. However, most of these will not extend to the infinite dimensional case, or at least not extend in a straightforward way, due to a number of reasons:

- (1) Most of basic linear algebra is based on the fact that $Tx = \lambda x$ is equivalent to $(T - \lambda)x = 0$, so that λ is an eigenvalue when $T - \lambda$ is not invertible. In the infinite dimensional setting $T - \lambda$ might be injective and not surjective, or vice versa, or invertible with $(T - \lambda)^{-1}$ not bounded, and so on.
- (2) Also, in linear algebra $T - \lambda$ is not invertible when $\det(T - \lambda) = 0$, and with this leading to most of the advanced results about eigenvalues and eigenvectors. In infinite dimensions, however, it is impossible to construct a determinant function $\det : B(H) \rightarrow \mathbb{C}$, and this even for the diagonal operators on $l^2(\mathbb{N})$.

Summarizing, we are in trouble. Forgetting about (2), which obviously leads nowhere, let us focus on the difficulties in (1). In order to cut short the discussion there, regarding the various properties of $T - \lambda$, we can just say that $T - \lambda$ is either invertible with bounded inverse, the “good case”, or not. We are led in this way to the following definition:

DEFINITION 12.4. *The spectrum of an operator $T \in B(H)$ is the set*

$$\sigma(T) = \left\{ \lambda \in \mathbb{C} \mid T - \lambda \notin B(H)^{-1} \right\}$$

where $B(H)^{-1} \subset B(H)$ is the set of invertible operators.

As a basic example, in the finite dimensional case, $H = \mathbb{C}^N$, the spectrum of a usual matrix $A \in M_N(\mathbb{C})$ is the collection of its eigenvalues, taken without multiplicities. We will see many other examples. In general, the spectrum has the following properties:

PROPOSITION 12.5. *The spectrum of $T \in B(H)$ contains the eigenvalue set*

$$\varepsilon(T) = \left\{ \lambda \in \mathbb{C} \mid \ker(T - \lambda) \neq \{0\} \right\}$$

and $\varepsilon(T) \subset \sigma(T)$ is an equality in finite dimensions, but not in infinite dimensions.

PROOF. We have several assertions here, the idea being as follows:

(1) First of all, the eigenvalue set is indeed the one in the statement, because $Tx = \lambda x$ tells us precisely that $T - \lambda$ must be not injective. The fact that we have $\varepsilon(T) \subset \sigma(T)$ is clear as well, because if $T - \lambda$ is not injective, it is not bijective.

(2) In finite dimensions we have $\varepsilon(T) = \sigma(T)$, because $T - \lambda$ is injective if and only if it is bijective, with the boundedness of the inverse being automatic.

(3) In infinite dimensions we can assume $H = l^2(\mathbb{N})$, and the shift operator $S(e_i) = e_{i+1}$ is injective but not surjective. Thus $0 \in \sigma(T) - \varepsilon(T)$. \square

Philosophically, the best way of thinking at this is as follows: the numbers $\lambda \notin \sigma(T)$ are good, because we can invert $T - \lambda$, the numbers $\lambda \in \sigma(T) - \varepsilon(T)$ are bad, because so they are, and the eigenvalues $\lambda \in \varepsilon(T)$ are evil. Welcome to operator theory.

Let us develop now some general theory. As a first result, we would like to prove that the spectra are non-empty. This is something tricky, and we will need:

PROPOSITION 12.6. *The following happen:*

- (1) $\|T\| < 1 \implies (1 - T)^{-1} = 1 + T + T^2 + \dots$
- (2) *The set $B(H)^{-1}$ is open.*
- (3) *The map $T \rightarrow T^{-1}$ is differentiable.*

PROOF. All these assertions are elementary, as follows:

(1) This follows as in the scalar case, the computation being as follows, provided that everything converges under the norm, which amounts in saying that $\|T\| < 1$:

$$\begin{aligned} (1 - T)(1 + T + T^2 + \dots) &= 1 - T + T - T^2 + T^2 - T^3 + \dots \\ &= 1 \end{aligned}$$

(2) Assuming $T \in B(H)^{-1}$, let us pick $S \in B(H)$ such that:

$$\|T - S\| < \frac{1}{\|T^{-1}\|}$$

With this choice, we have then the following estimate:

$$\begin{aligned} \|1 - T^{-1}S\| &= \|T^{-1}(T - S)\| \\ &\leq \|T^{-1}\| \cdot \|T - S\| \\ &< 1 \end{aligned}$$

Thus we have $T^{-1}S \in B(H)^{-1}$, and so $S \in B(H)^{-1}$, as desired.

(3) In the scalar case, the derivative of $f(t) = t^{-1}$ is $f'(t) = -t^{-2}$. In the present normed space setting the derivative is no longer a number, but rather a linear transformation, which can be found by developing $f(T) = T^{-1}$ at order 1, as follows:

$$\begin{aligned} (T + S)^{-1} &= ((1 + ST^{-1})T)^{-1} \\ &= T^{-1}(1 + ST^{-1})^{-1} \\ &= T^{-1}(1 - ST^{-1} + (ST^{-1})^2 - \dots) \\ &\simeq T^{-1}(1 - ST^{-1}) \\ &= T^{-1} - T^{-1}ST^{-1} \end{aligned}$$

Thus $f(T) = T^{-1}$ is indeed differentiable, with derivative $f'(T)S = -T^{-1}ST^{-1}$. \square

We can now formulate our first theorem about spectra, as follows:

THEOREM 12.7. *The spectrum of a bounded operator $T \in B(H)$ is:*

- (1) *Compact.*
- (2) *Contained in the disc $D_0(\|T\|)$.*
- (3) *Non-empty.*

PROOF. This can be proved by using Proposition 12.6, as follows:

(1) In view of (2) below, it is enough to prove that $\sigma(T)$ is closed. But this follows from the following computation, with $|\varepsilon|$ being small:

$$\begin{aligned} \lambda \notin \sigma(T) &\implies T - \lambda \in B(H)^{-1} \\ &\implies T - \lambda - \varepsilon \in B(H)^{-1} \\ &\implies \lambda + \varepsilon \notin \sigma(T) \end{aligned}$$

(2) This follows from the following computation:

$$\begin{aligned} \lambda > \|T\| &\implies \left\| \frac{T}{\lambda} \right\| < 1 \\ &\implies 1 - \frac{T}{\lambda} \in B(H)^{-1} \\ &\implies \lambda - T \in B(H)^{-1} \\ &\implies \lambda \notin \sigma(T) \end{aligned}$$

(3) Assume by contradiction $\sigma(T) = \emptyset$. Given a linear form $f \in B(H)^*$, consider the following map, which is well-defined, due to our assumption $\sigma(T) = \emptyset$:

$$\varphi : \mathbb{C} \rightarrow \mathbb{C} \quad , \quad \lambda \rightarrow f((T - \lambda)^{-1})$$

By using the fact that $T \rightarrow T^{-1}$ is differentiable, that we know from Proposition 12.6, we conclude that this map is differentiable, and so holomorphic. Also, we have:

$$\begin{aligned} \lambda \rightarrow \infty &\implies T - \lambda \rightarrow \infty \\ &\implies (T - \lambda)^{-1} \rightarrow 0 \\ &\implies f((T - \lambda)^{-1}) \rightarrow 0 \end{aligned}$$

Thus by the Liouville theorem we obtain $\varphi = 0$. But, in view of the definition of φ , this gives $(T - \lambda)^{-1} = 0$, which is a contradiction, as desired. \square

Here is now a second basic result regarding the spectra, inspired from what happens in finite dimensions, for the usual complex matrices, and which shows that things do not necessarily extend without troubles to the infinite dimensional setting:

THEOREM 12.8. *We have the following formula, valid for any operators S, T :*

$$\sigma(ST) \cup \{0\} = \sigma(TS) \cup \{0\}$$

In finite dimensions we have $\sigma(ST) = \sigma(TS)$, but this fails in infinite dimensions.

PROOF. There are several assertions here, the idea being as follows:

(1) This is something that we know in finite dimensions, coming from the fact that the characteristic polynomials of the associated matrices A, B coincide:

$$P_{AB} = P_{BA}$$

Thus we obtain $\sigma(ST) = \sigma(TS)$ in this case, as claimed. Observe that this improves twice the general formula in the statement, first because we have no issues at 0, and second because what we obtain is actually an equality of sets with multiplicities.

(2) In general now, let us first prove the main assertion, stating that $\sigma(ST), \sigma(TS)$ coincide outside 0. We first prove that we have the following implication:

$$1 \notin \sigma(ST) \implies 1 \notin \sigma(TS)$$

Assume indeed that $1 - ST$ is invertible, with inverse denoted R :

$$R = (1 - ST)^{-1}$$

We have then the following formulae, relating our variables R, S, T :

$$RST = STR = R - 1$$

By using $RST = R - 1$, we have the following computation:

$$\begin{aligned}(1 + TRS)(1 - TS) &= 1 + TRS - TS - TRSTS \\ &= 1 + TRS - TS - TRS + TS \\ &= 1\end{aligned}$$

A similar computation, using $STR = R - 1$, shows that we have:

$$(1 - TS)(1 + TRS) = 1$$

Thus $1 - TS$ is invertible, with inverse $1 + TRS$, which proves our claim. Now by multiplying by scalars, we deduce from this that for any $\lambda \in \mathbb{C} - \{0\}$ we have:

$$\lambda \notin \sigma(ST) \implies \lambda \notin \sigma(TS)$$

But this leads to the conclusion in the statement.

(3) Regarding now the counterexample to the formula $\sigma(ST) = \sigma(TS)$, in general, let us take S to be the shift on $H = L^2(\mathbb{N})$, given by the following formula:

$$S(e_i) = e_{i+1}$$

As for T , we can take it to be the adjoint of S , which is the following operator:

$$S^*(e_i) = \begin{cases} e_{i-1} & \text{if } i > 0 \\ 0 & \text{if } i = 0 \end{cases}$$

Let us compose now these two operators. In one sense, we have:

$$S^*S = 1 \implies 0 \notin \sigma(SS^*)$$

In the other sense, however, the situation is different, as follows:

$$SS^* = Proj(e_0^\perp) \implies 0 \in \sigma(SS^*)$$

Thus, the spectra do not match on 0, and we have our counterexample, as desired. \square

12b. Spectral radius

Let us develop now some systematic theory for the computation of the spectra, based on what we know about the eigenvalues of the usual complex matrices. As a first result, which is well-known for the usual matrices, and extends well, we have:

THEOREM 12.9. *We have the “polynomial functional calculus” formula*

$$\sigma(P(T)) = P(\sigma(T))$$

valid for any polynomial $P \in \mathbb{C}[X]$, and any operator $T \in B(H)$.

PROOF. We pick a scalar $\lambda \in \mathbb{C}$, and we decompose the polynomial $P - \lambda$:

$$P(X) - \lambda = c(X - r_1) \dots (X - r_n)$$

We have then the following equivalences:

$$\begin{aligned} \lambda \notin \sigma(P(T)) &\iff P(T) - \lambda \in B(H)^{-1} \\ &\iff c(T - r_1) \dots (T - r_n) \in B(H)^{-1} \\ &\iff T - r_1, \dots, T - r_n \in B(H)^{-1} \\ &\iff r_1, \dots, r_n \notin \sigma(T) \\ &\iff \lambda \notin P(\sigma(T)) \end{aligned}$$

Thus, we are led to the formula in the statement. \square

The above result is something very useful, and generalizing it will be our next task. As a first ingredient here, assuming that $A \in M_N(\mathbb{C})$ is invertible, we have:

$$\sigma(A^{-1}) = \sigma(A)^{-1}$$

It is possible to extend this formula to the arbitrary operators, and we will do this in a moment. Before starting, however, we have to find a class of functions generalizing both the polynomials $P \in \mathbb{C}[X]$ and the inverse function $x \rightarrow x^{-1}$. The answer to this question is provided by the rational functions, which are as follows:

DEFINITION 12.10. *A rational function $f \in \mathbb{C}(X)$ is a quotient of polynomials:*

$$f = \frac{P}{Q}$$

Assuming that P, Q are prime to each other, we can regard f as a usual function,

$$f : \mathbb{C} - X \rightarrow \mathbb{C}$$

with X being the set of zeros of Q , also called poles of f .

Now that we have our class of functions, the next step consists in applying them to operators. Here we cannot expect $f(T)$ to make sense for any f and any T , for instance because T^{-1} is defined only when T is invertible. We are led in this way to:

DEFINITION 12.11. *Given an operator $T \in B(H)$, and a rational function $f = P/Q$ having poles outside $\sigma(T)$, we can construct the following operator,*

$$f(T) = P(T)Q(T)^{-1}$$

that we can denote as a usual fraction, as follows,

$$f(T) = \frac{P(T)}{Q(T)}$$

due to the fact that $P(T), Q(T)$ commute, so that the order is irrelevant.

To be more precise, $f(T)$ is indeed well-defined, and the fraction notation is justified too. In more formal terms, we can say that we have a morphism of complex algebras as follows, with $\mathbb{C}(X)^T$ standing for the rational functions having poles outside $\sigma(T)$:

$$\mathbb{C}(X)^T \rightarrow B(H) \quad , \quad f \rightarrow f(T)$$

Summarizing, we have now a good class of functions, generalizing both the polynomials and the inverse map $x \rightarrow x^{-1}$. We can now extend Theorem 12.9, as follows:

THEOREM 12.12. *We have the “rational functional calculus” formula*

$$\sigma(f(T)) = f(\sigma(T))$$

valid for any rational function $f \in \mathbb{C}(X)$ having poles outside $\sigma(T)$.

PROOF. We pick a scalar $\lambda \in \mathbb{C}$, we write $f = P/Q$, and we set:

$$F = P - \lambda Q$$

By using now Theorem 12.9, for this polynomial, we obtain:

$$\begin{aligned} \lambda \in \sigma(f(T)) &\iff F(T) \notin B(H)^{-1} \\ &\iff 0 \in \sigma(F(T)) \\ &\iff 0 \in F(\sigma(T)) \\ &\iff \exists \mu \in \sigma(T), F(\mu) = 0 \\ &\iff \lambda \in f(\sigma(T)) \end{aligned}$$

Thus, we are led to the formula in the statement. \square

As an application of the above methods, we can investigate certain special classes of operators, such as the self-adjoint ones, and the unitary ones. Let us start with:

PROPOSITION 12.13. *The following happen:*

- (1) *We have $\sigma(T^*) = \overline{\sigma(T)}$, for any $T \in B(H)$.*
- (2) *If $T = T^*$ then $X = \sigma(T)$ satisfies $X = \overline{X}$.*
- (3) *If $U^* = U^{-1}$ then $X = \sigma(U)$ satisfies $X^{-1} = \overline{X}$.*

PROOF. We have several assertions here, the idea being as follows:

- (1) The spectrum of the adjoint operator T^* can be computed as follows:

$$\begin{aligned} \sigma(T^*) &= \left\{ \lambda \in \mathbb{C} \mid T^* - \lambda \notin B(H)^{-1} \right\} \\ &= \left\{ \lambda \in \mathbb{C} \mid T - \bar{\lambda} \notin B(H)^{-1} \right\} \\ &= \overline{\sigma(T)} \end{aligned}$$

- (2) This is clear indeed from (1).

(3) For a unitary operator, $U^* = U^{-1}$, Theorem 12.12 and (1) give:

$$\sigma(U)^{-1} = \sigma(U^{-1}) = \sigma(U^*) = \overline{\sigma(U)}$$

Thus, we are led to the conclusion in the statement. \square

In analogy with what happens for the usual matrices, we would like to improve now (2,3) above, with results stating that the spectrum $X = \sigma(T)$ satisfies $X \subset \mathbb{R}$ for self-adjoints, and $X \subset \mathbb{T}$ for unitaries. This will be tricky. Let us start with:

THEOREM 12.14. *The spectrum of a unitary operator*

$$U^* = U^{-1}$$

is on the unit circle, $\sigma(U) \subset \mathbb{T}$.

PROOF. Assuming $U^* = U^{-1}$, we have the following norm computation:

$$\|U\| = \sqrt{\|UU^*\|} = \sqrt{1} = 1$$

Now if we denote by D the unit disk, we obtain from this:

$$\sigma(U) \subset D$$

On the other hand, once again by using $U^* = U^{-1}$, we have as well:

$$\|U^{-1}\| = \|U^*\| = \|U\| = 1$$

Thus, as before with D being the unit disk in the complex plane, we have:

$$\sigma(U^{-1}) \subset D$$

Now by using Theorem 12.12, we obtain $\sigma(U) \subset D \cap D^{-1} = \mathbb{T}$, as desired. \square

We have as well a similar result for self-adjoints, as follows:

THEOREM 12.15. *The spectrum of a self-adjoint operator*

$$T = T^*$$

consists of real numbers, $\sigma(T) \subset \mathbb{R}$.

PROOF. The idea is that we can deduce the result from Theorem 12.14, by using the following remarkable rational function, depending on a parameter $r \in \mathbb{R}$:

$$f(z) = \frac{z + ir}{z - ir}$$

Indeed, for $r \gg 0$ the operator $f(T)$ is well-defined, and we have:

$$\left(\frac{T + ir}{T - ir}\right)^* = \frac{T - ir}{T + ir} = \left(\frac{T + ir}{T - ir}\right)^{-1}$$

Thus $f(T)$ is unitary, and by using Theorem 12.14 we obtain:

$$\begin{aligned}\sigma(T) &\subset f^{-1}(f(\sigma(T))) \\ &= f^{-1}(\sigma(f(T))) \\ &\subset f^{-1}(\mathbb{T}) \\ &= \mathbb{R}\end{aligned}$$

Thus, we are led to the conclusion in the statement. \square

Moving on, one key thing that we know about matrices, which is clear for the diagonalizable matrices, and then in general follows by density, is the following formula:

$$\sigma(e^A) = e^{\sigma(A)}$$

We would like to have such formulae for the general operators $T \in B(H)$, but this is something quite technical. Consider the rational calculus morphism from Definition 12.11, which is as follows, with the exponent standing for “having poles outside $\sigma(T)$ ”:

$$\mathbb{C}(X)^T \rightarrow B(H) \quad , \quad f \rightarrow f(T)$$

As mentioned before, the rational functions are holomorphic outside their poles, and this raises the question of extending this morphism, as follows:

$$Hol(\sigma(T)) \rightarrow B(H) \quad , \quad f \rightarrow f(T)$$

Normally this can be done in several steps. Let us start with:

PROPOSITION 12.16. *We can exponentiate any operator $T \in B(H)$, by setting:*

$$e^T = \sum_{k=0}^{\infty} \frac{T^k}{k!}$$

Similarly, we can define $f(T)$, for any holomorphic function $f : \mathbb{C} \rightarrow \mathbb{C}$.

PROOF. We must prove that the series defining e^T converges, and this follows from:

$$\|e^T\| \leq \sum_{k=0}^{\infty} \frac{\|T\|^k}{k!} = e^{\|T\|}$$

The case of the arbitrary holomorphic functions $f : \mathbb{C} \rightarrow \mathbb{C}$ is similar. \square

In general, the holomorphic functions are not entire, and the above method won't cover the rational functions $f \in \mathbb{C}(X)^T$ that we want to generalize. Thus, we must use something else. And the answer here comes from the Cauchy formula:

$$f(t) = \frac{1}{2\pi i} \int_{\gamma} \frac{f(z)}{z-t} dz$$

Indeed, given a rational function $f \in \mathbb{C}(X)^T$, the operator $f(T) \in B(H)$, constructed in Definition 12.9, can be recaptured in an analytic way, as follows:

$$f(T) = \frac{1}{2\pi i} \int_{\gamma} \frac{f(z)}{z - T} dz$$

Now given an arbitrary function $f \in \text{Hol}(\sigma(T))$, we can define $f(T) \in B(H)$ by the exactly same formula, and we obtain in this way the desired correspondence:

$$\text{Hol}(\sigma(T)) \rightarrow B(H) \quad , \quad f \rightarrow f(T)$$

This was for the plan. In practice now, all this needs a bit of care, with many verifications needed, and with the technical remark that a winding number must be added to the above Cauchy formulae, for things to be correct. Let us start with:

DEFINITION 12.17. *If γ is a loop in \mathbb{C} the number of times γ goes around a point $z \in \mathbb{C} - \gamma$ is computed by the following integral, called winding number:*

$$\text{Ind}(\gamma, z) = \frac{1}{2\pi i} \int_{\gamma} \frac{d\xi}{\xi - z}$$

We say that γ turns around z if $\text{Ind}(\gamma, z) = 1$, and that it does not turn if $\text{Ind}(\gamma, z) = 0$. Otherwise, we say that γ turns around z many times, or in the bad sense, or both.

Let $f : U \rightarrow \mathbb{C}$ be an holomorphic function defined on an open subset of \mathbb{C} , and γ be a loop in U . If $\text{Ind}(\gamma, z) \neq 0$ for $z \in \mathbb{C} - U$ then $f(z)$ is given by the Cauchy formula:

$$\text{Ind}(\gamma, z)f(z) = \frac{1}{2\pi i} \int_{\gamma} \frac{f(\xi)}{\xi - z} d\xi$$

Also, if $\text{Ind}(\gamma, z) = 0$ for $z \in \mathbb{C} - U$ then the integral of f on γ is zero:

$$\int_{\gamma} f(\xi) d\xi = 0$$

It is convenient to use formal combinations of loops, called cycles:

$$\Sigma = n_1\gamma_1 + \dots + n_r\gamma_r$$

The winding number for Σ is by definition the corresponding linear combination of winding numbers of its loop components, and the Cauchy formula holds for arbitrary cycles. Now by getting back to the linear operators, we can formulate:

DEFINITION 12.18. *Let $T \in B(H)$ and let $f : U \rightarrow \mathbb{C}$ be an holomorphic function defined on an open set containing $\sigma(T)$. Define an element $f(T)$ by the formula*

$$f(T) = \frac{1}{2\pi i} \int_{\Sigma} \frac{f(\xi)}{\xi - T} d\xi$$

where Σ is a cycle in $U - \sigma(T)$ which turns around $\sigma(T)$ and doesn't turn around $\mathbb{C} - U$.

The formula makes sense because Σ is in $U - \sigma(T)$. Also, $f(T)$ is independent of the choice of Σ . Indeed, let Σ_1 and Σ_2 be two cycles. Their difference $\Sigma_1 - \Sigma_2$ is a cycle which doesn't turn around $\sigma(a)$, neither around $\mathbb{C} - U$. The function $z \rightarrow f(z)/(z - T)$ being holomorphic $U - \sigma(T) \rightarrow B(H)$, its integral on $\Sigma_1 - \Sigma_2$ must be zero:

$$\int_{\Sigma_1 - \Sigma_2} \frac{f(\xi)}{\xi - T} d\xi = 0$$

Thus $f(T)$ is the same with respect to Σ_1 and to Σ_2 , and so Definition 12.18 is fully justified. Now with this definition in hand, we have the following result:

THEOREM 12.19. *Given $T \in B(H)$, we have a morphism of algebras as follows, where $Hol(\sigma(T))$ is the algebra of functions which are holomorphic around $\sigma(T)$,*

$$Hol(\sigma(T)) \rightarrow B(H) \quad , \quad f \rightarrow f(T)$$

which extends the previous rational functional calculus $f \rightarrow f(T)$. We have:

$$\sigma(f(T)) = f(\sigma(T))$$

Moreover, if $\sigma(T)$ is contained in an open set U and $f_n, f : U \rightarrow \mathbb{C}$ are holomorphic functions such that $f_n \rightarrow f$ uniformly on compact subsets of U then $f_n(T) \rightarrow f(T)$.

PROOF. There are several things to be proved here, as follows:

(1) Consider indeed the algebra $Hol(\sigma(T))$, with the convention that two functions are identified if they coincide on an open set containing $\sigma(T)$. We have then a construction $f \rightarrow f(T)$ as in the statement, provided by Definition 12.18.

(2) Let us prove now that our construction extends the one for rational functions. Since $1, z$ generate $\mathbb{C}(X)$, it is enough to show that $f(z) = 1$ implies $f(T) = 1$, and that $f(z) = z$ implies $f(T) = T$. For this purpose, we prove that $f(z) = z^n$ implies $f(T) = T^n$ for any n . But this follows by integrating over a circle γ of big radius, as follows:

$$\begin{aligned} f(T) &= \frac{1}{2\pi i} \int_{\gamma} \frac{\xi^n}{\xi - T} d\xi \\ &= \frac{1}{2\pi i} \int_{\gamma} \xi^{n-1} \left(1 - \frac{T}{\xi}\right)^{-1} d\xi \\ &= \frac{1}{2\pi i} \int_{\gamma} \xi^{n-1} \left(\sum_{k=0}^{\infty} \xi^{-k} T^k\right) d\xi \\ &= \sum_{k=0}^{\infty} \left(\frac{1}{2\pi i} \int_{\gamma} \xi^{n-k-1} d\xi\right) T^k \\ &= T^n \end{aligned}$$

(3) Regarding $\sigma(f(T)) = f(\sigma(T))$, it is enough to prove that this equality holds on the point 0, and we can do this by double inclusion, as follows:

“ \supset ”. Assume that $f(\sigma(T))$ contains 0, and let $z_0 \in \sigma(T)$ be such that $f(z_0) = 0$. Consider the function $g(z) = f(z)/(z - z_0)$. We have $g(T)(T - z_0) = f(T)$ by using the morphism property. Since $T - z_0$ is not invertible, $f(T)$ is not invertible either.

“ \subset ”. Assume now that $f(\sigma(T))$ does not contain 0. With the holomorphic function $g(z) = 1/f(z)$ we get $g(T) = f(T)^{-1}$, so $f(T)$ is invertible, and we are done.

(4) Finally, regarding the last assertion, this is clear from definitions. And with the remark that this can be applied to holomorphic functions written as series:

$$f(z) = \sum_{n=0}^{\infty} a_n(z - z_0)^n$$

Indeed, if this is the expansion of f around z_0 , with convergence radius r , and if $\sigma(T)$ is contained in the disc centered at z_0 of radius r , then $f(T)$ is given by:

$$f(T) = \sum_{n=0}^{\infty} a_n(T - z_0)^n$$

Summarizing, we have proved the result, and fully extended Theorem 12.12. \square

In order to formulate now our next result, we will need the following notion:

DEFINITION 12.20. *Given an operator $T \in B(H)$, its spectral radius*

$$\rho(T) \in [0, \|T\|]$$

is the radius of the smallest disk centered at 0 containing $\sigma(T)$.

Now with this notion in hand, we have the following key result, improving our key theoretical result so far about spectra, namely $\sigma(T) \neq \emptyset$, from Theorem 12.7:

THEOREM 12.21. *The spectral radius of an operator $T \in B(H)$ is given by*

$$\rho(T) = \lim_{n \rightarrow \infty} \|T^n\|^{1/n}$$

and in this formula, we can replace the limit by an inf.

PROOF. We have several things to be proved, the idea being as follows:

(1) Our first claim is that the numbers $u_n = \|T^n\|^{1/n}$ satisfy:

$$(n + m)u_{n+m} \leq nu_n + mu_m$$

Indeed, we have the following estimate, using the Young inequality $ab \leq a^p/p + b^q/q$, with exponents $p = (n+m)/n$ and $q = (n+m)/m$:

$$\begin{aligned}
 u_{n+m} &= \|T^{n+m}\|^{1/(n+m)} \\
 &\leq \|T^n\|^{1/(n+m)} \|T^m\|^{1/(n+m)} \\
 &\leq \|T^n\|^{1/n} \cdot \frac{n}{n+m} + \|T^m\|^{1/m} \cdot \frac{m}{n+m} \\
 &= \frac{nu_n + mu_m}{n+m}
 \end{aligned}$$

(2) Our second claim is that the second assertion holds, namely:

$$\lim_{n \rightarrow \infty} \|T^n\|^{1/n} = \inf_n \|T^n\|^{1/n}$$

For this purpose, we just need the inequality found in (1). Indeed, fix $m \geq 1$, let $n \geq 1$, and write $n = lm + r$ with $0 \leq r \leq m - 1$. By using twice $u_{ab} \leq u_b$, we get:

$$\begin{aligned}
 u_n &\leq \frac{1}{n}(lmu_{lm} + ru_r) \\
 &\leq \frac{1}{n}(lmu_m + ru_1) \\
 &\leq u_m + \frac{r}{n}u_1
 \end{aligned}$$

It follows that we have $\limsup_n u_n \leq u_m$, which proves our claim.

(3) Summarizing, we are left with proving the main formula, which is as follows, and with the remark that we already know that the sequence on the right converges:

$$\rho(T) = \lim_{n \rightarrow \infty} \|T^n\|^{1/n}$$

In one sense, we can use the polynomial calculus formula $\sigma(T^n) = \sigma(T)^n$. Indeed, this gives the following estimate, valid for any n , as desired:

$$\begin{aligned}
 \rho(T) &= \sup_{\lambda \in \sigma(T)} |\lambda| \\
 &= \sup_{\rho \in \sigma(T)^n} |\rho|^{1/n} \\
 &= \sup_{\rho \in \sigma(T^n)} |\rho|^{1/n} \\
 &= \rho(T^n)^{1/n} \\
 &\leq \|T^n\|^{1/n}
 \end{aligned}$$

(4) For the reverse inequality, we fix a number $\rho > \rho(T)$, and we want to prove that we have $\rho \geq \lim_{n \rightarrow \infty} \|T^n\|^{1/n}$. By using the Cauchy formula, we have:

$$\begin{aligned} \frac{1}{2\pi i} \int_{|z|=\rho} \frac{z^n}{z-T} dz &= \frac{1}{2\pi i} \int_{|z|=\rho} \sum_{k=0}^{\infty} z^{n-k-1} T^k dz \\ &= \sum_{k=0}^{\infty} \frac{1}{2\pi i} \left(\int_{|z|=\rho} z^{n-k-1} dz \right) T^k \\ &= \sum_{k=0}^{\infty} \delta_{n,k+1} T^k \\ &= T^{n-1} \end{aligned}$$

By applying the norm we obtain from this formula:

$$\begin{aligned} \|T^{n-1}\| &\leq \frac{1}{2\pi} \int_{|z|=\rho} \left\| \frac{z^n}{z-T} \right\| dz \\ &\leq \rho^n \cdot \sup_{|z|=\rho} \left\| \frac{1}{z-T} \right\| \end{aligned}$$

Since the sup does not depend on n , by taking n -th roots, we obtain in the limit:

$$\rho \geq \lim_{n \rightarrow \infty} \|T^n\|^{1/n}$$

Now recall that ρ was by definition an arbitrary number satisfying $\rho > \rho(T)$. Thus, we have obtained the following estimate, valid for any $T \in B(H)$:

$$\rho(T) \geq \lim_{n \rightarrow \infty} \|T^n\|^{1/n}$$

Thus, we are led to the conclusion in the statement. □

In the case of the normal elements, we have the following finer result:

THEOREM 12.22. *The spectral radius of a normal element,*

$$TT^* = T^*T$$

is equal to its norm.

PROOF. We can proceed in two steps, as follows:

Step 1. In the case $T = T^*$ we have $\|T^n\| = \|T\|^n$ for any exponent of the form $n = 2^k$, by using the formula $\|TT^*\| = \|T\|^2$, and by taking n -th roots we get:

$$\rho(T) \geq \|T\|$$

Thus, we are done with the self-adjoint case, with the result $\rho(T) = \|T\|$.

Step 2. In the general normal case $TT^* = T^*T$ we have $T^n(T^n)^* = (TT^*)^n$, and by using this, along with the result from Step 1, applied to TT^* , we obtain:

$$\begin{aligned}
 \rho(T) &= \lim_{n \rightarrow \infty} \|T^n\|^{1/n} \\
 &= \sqrt{\lim_{n \rightarrow \infty} \|T^n(T^n)^*\|^{1/n}} \\
 &= \sqrt{\lim_{n \rightarrow \infty} \|(TT^*)^n\|^{1/n}} \\
 &= \sqrt{\rho(TT^*)} \\
 &= \sqrt{\|T\|^2} \\
 &= \|T\|
 \end{aligned}$$

Thus, we are led to the conclusion in the statement. \square

12c. Normal operators

By using Theorem 12.22 we can say a number of non-trivial things about the self-adjoint and normal operators, commonly known as “spectral theorems”. As a first result here, we can improve the polynomial functional calculus formula, as follows:

THEOREM 12.23. *Given $T \in B(H)$ normal, we have a morphism of algebras*

$$\mathbb{C}[X] \rightarrow B(H) \quad , \quad P \rightarrow P(T)$$

having the properties $\|P(T)\| = \|P_{|\sigma(T)}\|$, and $\sigma(P(T)) = P(\sigma(T))$.

PROOF. This is an improvement of Theorem 12.9 in the normal case, with the extra assertion being the norm estimate. But the element $P(T)$ being normal, we can apply to it the spectral radius formula for normal elements, and we obtain:

$$\begin{aligned}
 \|P(T)\| &= \rho(P(T)) \\
 &= \sup_{\lambda \in \sigma(P(T))} |\lambda| \\
 &= \sup_{\lambda \in P(\sigma(T))} |\lambda| \\
 &= \|P_{|\sigma(T)}\|
 \end{aligned}$$

Thus, we are led to the conclusions in the statement. \square

We can improve as well the rational calculus formula, and the holomorphic calculus formula, in the same way. Importantly now, at a more advanced level, we have:

THEOREM 12.24. *Given $T \in B(H)$ normal, we have a morphism of algebras*

$$C(\sigma(T)) \rightarrow B(H) \quad , \quad f \rightarrow f(T)$$

which is isometric, $\|f(T)\| = \|f\|$, and has the property $\sigma(f(T)) = f(\sigma(T))$.

PROOF. The idea here is to “complete” the morphism in Theorem 12.23, namely:

$$\mathbb{C}[X] \rightarrow B(H) \quad , \quad P \rightarrow P(T)$$

Indeed, we know from Theorem 12.23 that this morphism is continuous, and is in fact isometric, when regarding the polynomials $P \in \mathbb{C}[X]$ as functions on $\sigma(T)$:

$$\|P(T)\| = \|P|_{\sigma(T)}\|$$

Thus, by Stone-Weierstrass, we have a unique isometric extension, as follows:

$$C(\sigma(T)) \rightarrow B(H) \quad , \quad f \rightarrow f(T)$$

It remains to prove $\sigma(f(T)) = f(\sigma(T))$, and we can do this by double inclusion:

“ \subset ” Given a continuous function $f \in C(\sigma(T))$, we must prove that we have:

$$\lambda \notin f(\sigma(T)) \implies \lambda \notin \sigma(f(T))$$

For this purpose, consider the following function, which is well-defined:

$$\frac{1}{f - \lambda} \in C(\sigma(T))$$

We can therefore apply this function to T , and we obtain:

$$\left(\frac{1}{f - \lambda} \right) T = \frac{1}{f(T) - \lambda}$$

In particular $f(T) - \lambda$ is invertible, so $\lambda \notin \sigma(f(T))$, as desired.

“ \supset ” Given a continuous function $f \in C(\sigma(T))$, we must prove that we have:

$$\lambda \in f(\sigma(T)) \implies \lambda \in \sigma(f(T))$$

But this is the same as proving that we have:

$$\mu \in \sigma(T) \implies f(\mu) \in \sigma(f(T))$$

For this purpose, we approximate our function by polynomials, $P_n \rightarrow f$, and we examine the following convergence, which follows from $P_n \rightarrow f$:

$$P_n(T) - P_n(\mu) \rightarrow f(T) - f(\mu)$$

We know from polynomial functional calculus that we have:

$$P_n(\mu) \in P_n(\sigma(T)) = \sigma(P_n(T))$$

Thus, the operators $P_n(T) - P_n(\mu)$ are not invertible. On the other hand, we know that the set formed by the invertible operators is open, so its complement is closed. Thus the limit $f(T) - f(\mu)$ is not invertible either, and so $f(\mu) \in \sigma(f(T))$, as desired. \square

As an important comment, Theorem 12.24 is not exactly in final form, because it misses an important point, namely that our correspondence maps:

$$\bar{z} \rightarrow T^*$$

However, this is something non-trivial, and we will be back to this later. Observe however that Theorem 12.24 is fully powerful for the self-adjoint operators, $T = T^*$, where the spectrum is real, so where $z = \bar{z}$ on the spectrum. We will be back to this.

As a second result now, along the same lines, we can further extend Theorem 12.24 into a measurable functional calculus theorem, as follows:

THEOREM 12.25. *Given $T \in B(H)$ normal, we have a morphism of algebras as follows, with L^∞ standing for abstract measurable functions, or Borel functions,*

$$L^\infty(\sigma(T)) \rightarrow B(H) \quad , \quad f \rightarrow f(T)$$

which is isometric, $\|f(T)\| = \|f\|$, and has the property $\sigma(f(T)) = f(\sigma(T))$.

PROOF. As before, the idea will be that of “completing” what we have. To be more precise, we can use the Riesz theorem and a polarization trick, as follows:

- (1) Given a vector $x \in H$, consider the following functional:

$$C(\sigma(T)) \rightarrow \mathbb{C} \quad , \quad g \rightarrow \langle g(T)x, x \rangle$$

By the Riesz theorem, this functional must be the integration with respect to a certain measure μ on the space $\sigma(T)$. Thus, we have a formula as follows:

$$\langle g(T)x, x \rangle = \int_{\sigma(T)} g(z) d\mu(z)$$

Now given an arbitrary Borel function $f \in L^\infty(\sigma(T))$, as in the statement, we can define a number $\langle f(T)x, x \rangle \in \mathbb{C}$, by using exactly the same formula, namely:

$$\langle f(T)x, x \rangle = \int_{\sigma(T)} f(z) d\mu(z)$$

Thus, we have managed to define numbers $\langle f(T)x, x \rangle \in \mathbb{C}$, for all vectors $x \in H$, and in addition we can recover these numbers as follows, with $g_n \in C(\sigma(T))$:

$$\langle f(T)x, x \rangle = \lim_{g_n \rightarrow f} \langle g_n(T)x, x \rangle$$

- (2) In order to define now numbers $\langle f(T)x, y \rangle \in \mathbb{C}$, for all vectors $x, y \in H$, we can use a polarization trick. Indeed, for any operator $S \in B(H)$ we have:

$$\langle S(x+y), x+y \rangle = \langle Sx, x \rangle + \langle Sy, y \rangle + \langle Sx, y \rangle + \langle Sy, x \rangle$$

By replacing $y \rightarrow iy$, we have as well the following formula:

$$\langle S(x+iy), x+iy \rangle = \langle Sx, x \rangle + \langle Sy, y \rangle - i \langle Sx, y \rangle + i \langle Sy, x \rangle$$

By multiplying this formula by i , and summing with the first one, we obtain:

$$\begin{aligned} \langle S(x+y), x+y \rangle + i \langle S(x+iy), x+iy \rangle &= (1+i)[\langle Sx, x \rangle + \langle Sy, y \rangle] \\ &+ 2 \langle Sx, y \rangle \end{aligned}$$

(3) But with this, we can now finish. Indeed, by combining (1,2), given a Borel function $f \in L^\infty(\sigma(T))$, we can define numbers $\langle f(T)x, y \rangle \in \mathbb{C}$ for any $x, y \in H$, and it is routine to check, by using approximation by continuous functions $g_n \rightarrow f$ as in (1), that we obtain in this way an operator $f(T) \in B(H)$, having all the desired properties. \square

As a comment here, the above result and its proof provide us with more than a Borel functional calculus, because what we got is a certain measure on the spectrum $\sigma(T)$, along with a functional calculus for the L^∞ functions with respect to this measure. We will be back to this later, and for the moment we will only need Theorem 12.25 as formulated, with $L^\infty(\sigma(T))$ standing, a bit abusively, for the Borel functions on $\sigma(T)$.

12d. Diagonalization

We can now diagonalize the normal operators. We will do this in 3 steps, first for the self-adjoint operators, then for the families of commuting self-adjoint operators, and finally for the general normal operators, by using a trick of the following type:

$$T = \operatorname{Re}(T) + i\operatorname{Im}(T)$$

The diagonalization in infinite dimensions is more tricky than in finite dimensions, and instead of writing a formula of type $T = UDU^*$, with $U, D \in B(H)$ being respectively unitary and diagonal, we will express our operator as $T = U^*MU$, with $U : H \rightarrow K$ being a certain unitary, and $M \in B(K)$ being a certain diagonal operator. This is how the spectral theorem is best formulated, in view of applications.

In practice, the explicit construction of U, M , which will be actually part of the proof, is also needed. For the self-adjoint operators, the statement and proof are as follows:

THEOREM 12.26. *Any self-adjoint operator $T \in B(H)$ can be diagonalized,*

$$T = U^*M_fU$$

with $U : H \rightarrow L^2(X)$ being a unitary operator from H to a certain L^2 space associated to T , with $f : X \rightarrow \mathbb{R}$ being a certain function, once again associated to T , and with

$$M_f(g) = fg$$

being the usual multiplication operator by f , on the Hilbert space $L^2(X)$.

PROOF. The construction of U, f can be done in several steps, as follows:

(1) We first prove the result in the special case where our operator T has a cyclic vector $x \in H$, with this meaning that the following holds:

$$\overline{\text{span} \left(T^k x \mid n \in \mathbb{N} \right)} = H$$

For this purpose, let us go back to the proof of Theorem 12.25. We will use the following formula from there, with μ being the measure on $X = \sigma(T)$ associated to x :

$$\langle g(T)x, x \rangle = \int_{\sigma(T)} g(z) d\mu(z)$$

Our claim is that we can define a unitary $U : H \rightarrow L^2(X)$, first on the dense part spanned by the vectors $T^k x$, by the following formula, and then by continuity:

$$U[g(T)x] = g$$

Indeed, the following computation shows that U is well-defined, and isometric:

$$\begin{aligned} \|g(T)x\|^2 &= \langle g(T)x, g(T)x \rangle \\ &= \langle g(T)^* g(T)x, x \rangle \\ &= \langle |g|^2(T)x, x \rangle \\ &= \int_{\sigma(T)} |g(z)|^2 d\mu(z) \\ &= \|g\|_2^2 \end{aligned}$$

We can then extend U by continuity into a unitary $U : H \rightarrow L^2(X)$, as claimed. Now observe that we have the following formula:

$$\begin{aligned} UTU^*g &= U[Tg(T)x] \\ &= U[(zg)(T)x] \\ &= zg \end{aligned}$$

Thus our result is proved in the present case, with U as above, and with $f(z) = z$.

(2) We discuss now the general case. Our first claim is that H has a decomposition as follows, with each H_i being invariant under T , and admitting a cyclic vector x_i :

$$H = \bigoplus_i H_i$$

Indeed, this is something elementary, the construction being by recurrence in finite dimensions, in the obvious way, and by using the Zorn lemma in general. Now with this decomposition in hand, we can make a direct sum of the diagonalizations obtained in (1), for each of the restrictions $T|_{H_i}$, and we obtain the formula in the statement. \square

We have the following technical generalization of the above result:

THEOREM 12.27. *Any family of commuting self-adjoint operators $T_i \in B(H)$ can be jointly diagonalized,*

$$T_i = U^* M_{f_i} U$$

with $U : H \rightarrow L^2(X)$ being a unitary operator from H to a certain L^2 space associated to $\{T_i\}$, with $f_i : X \rightarrow \mathbb{R}$ being certain functions, once again associated to T_i , and with

$$M_{f_i}(g) = f_i g$$

being the usual multiplication operator by f_i , on the Hilbert space $L^2(X)$.

PROOF. This is similar to the proof of Theorem 12.26, by suitably modifying the measurable calculus formula, and the measure μ itself, as to have this formula working for all the operators T_i . With this modification done, everything extends. \square

In order to discuss now the case of the arbitrary normal operators, we will need:

PROPOSITION 12.28. *Any operator $T \in B(H)$ can be written as*

$$T = \operatorname{Re}(T) + i\operatorname{Im}(T)$$

with $\operatorname{Re}(T), \operatorname{Im}(T) \in B(H)$ being self-adjoint, and this decomposition is unique.

PROOF. This is something elementary, the idea being as follows:

(1) We can use indeed the same formulae for the real and imaginary part as in the complex number case, the decomposition formula being as follows:

$$T = \frac{T + T^*}{2} + i \cdot \frac{T - T^*}{2i}$$

To be more precise, both the operators on the right are self-adjoint, and the summing formula holds indeed, and so we have our decomposition result, as desired.

(2) Regarding now the uniqueness, by linearity it is enough to show that $R + iS = 0$ with R, S both self-adjoint implies $R = S = 0$. But this follows by applying the adjoint to $R + iS = 0$, which gives $R - iS = 0$, and so $R = S = 0$, as desired. \square

We can now discuss the case of arbitrary normal operators, as follows:

THEOREM 12.29. *Any normal operator $T \in B(H)$ can be diagonalized,*

$$T = U^* M_f U$$

with $U : H \rightarrow L^2(X)$ being a unitary operator from H to a certain L^2 space associated to T , with $f : X \rightarrow \mathbb{C}$ being a certain function, once again associated to T , and with

$$M_f(g) = fg$$

being the usual multiplication operator by f , on the Hilbert space $L^2(X)$.

PROOF. This is our main diagonalization theorem, the idea being as follows:

(1) Consider the decomposition of T into its real and imaginary parts, as constructed in the proof of Proposition 12.28, namely:

$$T = \frac{T + T^*}{2} + i \cdot \frac{T - T^*}{2i}$$

We know that the real and imaginary parts are self-adjoint operators. Now since T was assumed to be normal, $TT^* = T^*T$, these real and imaginary parts commute:

$$\left[\frac{T + T^*}{2}, \frac{T - T^*}{2i} \right] = 0$$

Thus Theorem 12.27 applies to these real and imaginary parts, and gives the result. \square

This was for our series of diagonalization theorems. There is of course one more result here, regarding the families of commuting normal operators, as follows:

THEOREM 12.30. *Any family of commuting normal operators $T_i \in B(H)$ can be jointly diagonalized,*

$$T_i = U^* M_{f_i} U$$

with $U : H \rightarrow L^2(X)$ being a unitary operator from H to a certain L^2 space associated to $\{T_i\}$, with $f_i : X \rightarrow \mathbb{C}$ being certain functions, once again associated to T_i , and with

$$M_{f_i}(g) = f_i g$$

being the usual multiplication operator by f_i , on the Hilbert space $L^2(X)$.

PROOF. This is similar to the proof of Theorem 12.27 and Theorem 12.29, by combining the arguments there. To be more precise, this follows as Theorem 12.27, by using the decomposition trick from the proof of Theorem 12.29. \square

With the above diagonalization results in hand, we can now “fix” the continuous and measurable functional calculus theorems, with a key complement, as follows:

THEOREM 12.31. *Given a normal operator $T \in B(H)$, the following hold, for both the functional calculus and the measurable calculus morphisms:*

- (1) *These morphisms are $*$ -morphisms.*
- (2) *The function \bar{z} gets mapped to T^* .*
- (3) *The functions $\operatorname{Re}(z)$, $\operatorname{Im}(z)$ get mapped to $\operatorname{Re}(T)$, $\operatorname{Im}(T)$.*
- (4) *The function $|z|^2$ gets mapped to $TT^* = T^*T$.*
- (5) *If f is real, then $f(T)$ is self-adjoint.*

PROOF. These assertions are more or less equivalent, with (1) being the main one, which obviously implies everything else. But this assertion (1) follows from the diagonalization result for normal operators, from Theorem 12.29. \square

12e. Exercises

This was a quite fundamental chapter, and as exercises on this, we have:

EXERCISE 12.32. *Develop the theory of positive operators, satisfying $\langle Tx, x \rangle \geq 0$.*

EXERCISE 12.33. *Discuss as well the strict positivity, $\langle Tx, x \rangle > 0$ for $x \neq 0$.*

EXERCISE 12.34. *Check the various details of the holomorphic calculus theorem.*

EXERCISE 12.35. *Check as well the details of the measurable calculus theorem.*

EXERCISE 12.36. *Learn the various alternative proofs for the spectral theorem.*

EXERCISE 12.37. *Work out a polar decomposition result, for the linear operators.*

EXERCISE 12.38. *Learn about the compact operators, and their various properties.*

EXERCISE 12.39. *Learn also about unbounded operators, and quantum mechanics.*

As bonus exercise, and no surprise here, learn some systematic operator theory.

Part IV

Haar integration

*Down the road I look and there runs Mary
Hair of gold and lips like cherries
It's good to touch
The green, green grass of home*

CHAPTER 13

Random walks

13a. Random walks

We have learned the basics of measure theory and probability, and time now to see if this knowledge can be of any help, in relation with concrete questions. The first question that we would like to discuss, which is something very basic, is as follows:

QUESTION 13.1. *Given a graph X , with a distinguished vertex $*$:*

- (1) *What is the number L_k of length k loops on X , based at $*$?*
- (2) *Equivalently, what is the measure μ having L_k as moments?*

To be more precise, we are mainly interested in the first question, counting loops on graphs, with this being notoriously related to many applied mathematics questions, of discrete type. As for the second question, this is a technical, useful probabilistic reformulation of the first question, that we will usually prefer, in what follows.

Actually, in relation with this, the fact that a real measure μ as above exists indeed is not exactly obvious. But comes from the following result, which is something quite elementary, and which can be very helpful for explicit computations:

THEOREM 13.2. *Given a graph X , with adjacency matrix $d \in M_N(0, 1)$, we have:*

$$L_k = (d^k)_{**}$$

When writing $d = UDU^t$ with $U \in O_N$ and $D = \text{diag}(\lambda_1, \dots, \lambda_N)$ with $\lambda_i \in \mathbb{R}$, we have

$$L_k = \sum_i U_{*i}^2 \lambda_i^k$$

and the real probability measure μ having these numbers as moments is given by

$$\mu = \sum_i U_{*i}^2 \delta_{\lambda_i}$$

with the delta symbols standing as usual for Dirac masses.

PROOF. There are several things going on here, the idea being as follows:

(1) According to the usual rule of matrix multiplication, the formula for the powers of the adjacency matrix $d \in M_N(0, 1)$ is as follows:

$$\begin{aligned}
 (d^k)_{i_0 i_k} &= \sum_{i_1, \dots, i_{k-1}} d_{i_0 i_1} d_{i_1 i_2} \dots d_{i_{k-1} i_k} \\
 &= \sum_{i_1, \dots, i_{k-1}} \delta_{i_0 - i_1} \delta_{i_1 - i_2} \dots \delta_{i_{k-1} - i_k} \\
 &= \sum_{i_1, \dots, i_{k-1}} \delta_{i_0 - i_1 - \dots - i_{k-1} - i_k} \\
 &= \# \left\{ i_0 - i_1 - \dots - i_{k-1} - i_k \right\}
 \end{aligned}$$

In particular, with $i_0 = i_k = *$, we obtain the following formula, as claimed:

$$(d^k)_{**} = \# \left\{ * - i_1 - \dots - i_{k-1} - * \right\} = L_k$$

(2) Now since $d \in M_N(0, 1)$ is symmetric, this matrix is diagonalizable, with the diagonalization being as follows, with $U \in O_N$, and $D = \text{diag}(\lambda_1, \dots, \lambda_N)$ with $\lambda_i \in \mathbb{R}$:

$$d = U D U^t$$

By using this formula, we obtain the second formula in the statement:

$$\begin{aligned}
 L_k &= (d^k)_{**} \\
 &= (U D^k U^t)_{**} \\
 &= \sum_i U_{*i} \lambda_i^k (U^t)_{i*} \\
 &= \sum_i U_{*i}^2 \lambda_i^k
 \end{aligned}$$

(3) Finally, the last assertion is clear from this, because the moments of the measure in the statement, $\mu = \sum_i U_{*i}^2 \delta_{\lambda_i}$, are the following numbers:

$$\begin{aligned}
 M_k &= \int_{\mathbb{R}} x^k d\mu(x) \\
 &= \sum_i U_{*i}^2 \lambda_i^k \\
 &= L_k
 \end{aligned}$$

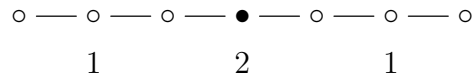
Observe also that μ is indeed of mass 1, because all rows of $U \in O_N$ must be of norm 1, and so $\sum_i U_{*i}^2 = 1$. Thus, we are led to the conclusions in the statement. \square

At the level of examples, the finite graphs can be quite complicated, and we will discuss them later. So, let us look instead at the infinite graphs, and try for instance to count

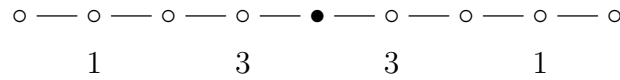
the length k paths on \mathbb{Z} , based at 0. At $k = 1$ we have 2 such paths, ending at -1 and 1 , and the count results can be pictured as follows, in a self-explanatory way:



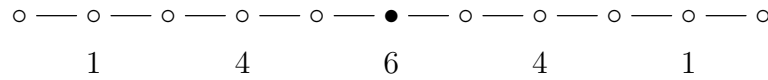
At $k = 2$ now, we have 4 paths, one of which ends at -2 , two of which end at 0 , and one of which ends at 2 . The results can be pictured as follows:



At $k = 3$ now, we have 8 paths, the distribution of the endpoints being as follows:



As for $k = 4$, here we have 16 paths, the distribution of the endpoints being as follows:



And good news, we can see in the above the Pascal triangle. Thus, getting back now to Question 13.1, we can answer it for the graph \mathbb{Z} , the result being as follows:

THEOREM 13.3. *The paths on \mathbb{Z} are counted by the binomial coefficients. In particular, the $2k$ -paths based at 0 are counted by the central binomial coefficients,*

$$L_{2k} = \binom{2k}{k}$$

and μ is the centered measure having these numbers as even moments.

PROOF. This basically follows from the above discussion, as follows:

(1) In what regards the count, we certainly have the Pascal triangle, as discovered above, and the rest is just a matter of finishing. There are many possible ways here, a straightforward one being that of arguing that the number C_k^l of length k loops $0 \rightarrow l$ is subject, due to the binary choice at the end, to the following recurrence relation:

$$C_k^l = C_{k-1}^{l-1} + C_{k-1}^{l+1}$$

But this is exactly the recurrence for the Pascal triangle, so done with the count.

(2) As for the second assertion, the first part, regarding L_{2k} , is clear from this, and the second part is more of an empty statement, with μ still remaining to be computed. \square

13b. Catalan numbers

As a second illustration, let us try to count the loops of \mathbb{N} , based at 0. This is something less obvious, and at the experimental level, the result is as follows:

PROPOSITION 13.4. *The Catalan numbers C_k , counting the loops on \mathbb{N} based at 0,*

$$C_k = \#\left\{0 - i_1 - \dots - i_{2k-1} - 0\right\}$$

are numerically 1, 2, 5, 14, 42, 132, 429, 1430, 4862, 16796, 58786, ...

PROOF. To start with, we have indeed $C_1 = 1$, the only loop here being $0 - 1 - 0$. Then we have $C_2 = 2$, due to two possible loops, namely:

$$0 - 1 - 0 - 1 - 0$$

$$0 - 1 - 2 - 1 - 0$$

Then we have $C_3 = 5$, the possible loops here being as follows:

$$0 - 1 - 0 - 1 - 0 - 1 - 0$$

$$0 - 1 - 0 - 1 - 2 - 1 - 0$$

$$0 - 1 - 2 - 1 - 0 - 1 - 0$$

$$0 - 1 - 2 - 1 - 2 - 1 - 0$$

$$0 - 1 - 2 - 3 - 2 - 1 - 0$$

In general, the same method works, with $C_4 = 14$ being left to you, as an exercise, and with C_5 and higher to me, and I will be back with the solution, in due time. \square

Obviously, computing the numbers C_k is no easy task, and finding the formula of C_k , out of the data that we have, does not look as an easy task either. So, we will do what combinatorists do, let me teach you. The first step is to relax, then to look around, not with the aim of computing your numbers C_k , but rather with the aim of finding other objects counted by the same numbers C_k . With a bit of luck, among these objects some will be easier to count than the others, and this will eventually compute C_k .

This was for the strategy. In practice now, we first have the following result:

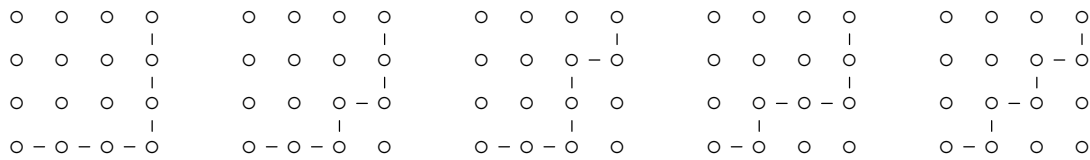
THEOREM 13.5. *The Catalan numbers C_k count:*

- (1) *The length $2k$ loops on \mathbb{N} , based at 0.*
- (2) *The noncrossing pairings of $1, \dots, 2k$.*
- (3) *The noncrossing partitions of $1, \dots, k$.*
- (4) *The length $2k$ Dyck paths in the plane.*

PROOF. All this is standard combinatorics, the idea being as follows:

(1) To start with, in what regards the various objects involved, the length $2k$ loops on \mathbb{N} are the length $2k$ loops on \mathbb{N} that we know, and the same goes for the noncrossing pairings of $1, \dots, 2k$, and for the noncrossing partitions of $1, \dots, k$, the idea here being that you must be able to draw the pairing or partition in a noncrossing way.

(2) Regarding now the length $2k$ Dyck paths in the plane, these are by definition the paths from $(0, 0)$ to (k, k) , marching North-East over the integer lattice $\mathbb{Z}^2 \subset \mathbb{R}^2$, by staying inside the square $[0, k] \times [0, k]$, and staying as well under the diagonal of this square. As an example, here are the 5 possible Dyck paths at $n = 3$:



(3) Thus, we have definitions for all the objects involved, and in each case, if you start counting them, as we did in Proposition 13.4 with the loops on \mathbb{N} , you always end up with the same sequence of numbers, namely those found in Proposition 13.4:

$$1, 2, 5, 14, 42, 132, 429, 1430, 4862, 16796, 58786, \dots$$

(4) In order to prove now that (1-4) produce indeed the same numbers, many things can be said. The idea is that, leaving aside mathematical brevity, and more specifically abstract reasonings of type $a = b, b = c \implies a = c$, what we have to do, in order to fully understand what is going on, is to establish $\binom{4}{2} = 6$ equalities, via bijective proofs.

(5) But this can be done, indeed. As an example here, the noncrossing pairings of $1, \dots, 2k$ from (2) are in bijection with the noncrossing partitions of $1, \dots, k$ from (3), via fattening the pairings and shrinking the partitions. We will leave the details here as an instructive exercise, and exercise as well, to add (1) and (4) to the picture.

(6) However, matter of having our theorem formally proved, I mean by me professor and not by you student, here is a less elegant argument, which is however very quick, and does the job. The point is that, in each of the cases (1-4) under consideration, the numbers C_k that we get are easily seen to be subject to the following recurrence:

$$C_{k+1} = \sum_{a+b=k} C_a C_b$$

The initial data being the same, namely $C_1 = 1$ and $C_2 = 2$, in each of the cases (1-4) under consideration, we get indeed the same numbers. \square

Now we can pass to the second step, namely selecting in the above list the objects that we find the most convenient to count, and count them. This leads to:

THEOREM 13.6. *The Catalan numbers are given by the formula*

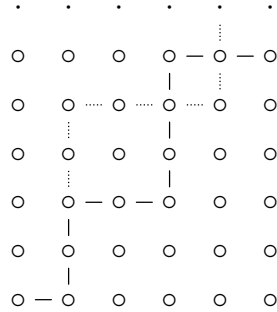
$$C_k = \frac{1}{k+1} \binom{2k}{k}$$

with this being best seen by counting the length $2k$ Dyck paths in the plane.

PROOF. This is something quite tricky, the idea being as follows:

(1) Let us count indeed the Dyck paths in the plane. For this purpose, we use a trick. Indeed, if we ignore the assumption that our path must stay under the diagonal of the square, we have $\binom{2k}{k}$ such paths. And among these, we have the “good” ones, those that we want to count, and then the “bad” ones, those that we want to ignore.

(2) So, let us count the bad paths, those crossing the diagonal of the square, and reaching the higher diagonal next to it, the one joining $(0, 1)$ and $(k, k+1)$. In order to count these, the trick is to “flip” their bad part over that higher diagonal, as follows:



(3) Now observe that, as it is obvious on the above picture, due to the flipping, the flipped bad path will no longer end in (k, k) , but rather in $(k-1, k+1)$. Moreover, more is true, in the sense that, by thinking a bit, we see that the flipped bad paths are precisely those ending in $(k-1, k+1)$. Thus, we can count these flipped bad paths, and so the bad paths, and so the good paths too, and so good news, we are done.

(4) To finish now, by putting everything together, we have:

$$\begin{aligned} C_k &= \binom{2k}{k} - \binom{2k}{k-1} \\ &= \binom{2k}{k} - \frac{k}{k+1} \binom{2k}{k} \\ &= \frac{1}{k+1} \binom{2k}{k} \end{aligned}$$

Thus, we are led to the formula in the statement. \square

We have as well another approach to all this, computation of the Catalan numbers, this time based on rock-solid standard calculus, as follows:

THEOREM 13.7. *The Catalan numbers have the following properties:*

- (1) *They satisfy $C_{k+1} = \sum_{a+b=k} C_a C_b$.*
- (2) *The series $f(z) = \sum_{k \geq 0} C_k z^k$ satisfies $zf^2 - f + 1 = 0$.*
- (3) *This series is given by $f(z) = \frac{1 - \sqrt{1 - 4z}}{2z}$.*
- (4) *We have the formula $C_k = \frac{1}{k+1} \binom{2k}{k}$.*

PROOF. This is best viewed by using noncrossing pairings, as follows:

(1) Let us count the noncrossing pairings of $\{1, \dots, 2k+2\}$. Such a pairing appears by pairing 1 to an odd number, $2a+1$, and then inserting a noncrossing pairing of $\{2, \dots, 2a\}$, and a noncrossing pairing of $\{2a+2, \dots, 2k+2\}$. Thus we have, as claimed:

$$C_{k+1} = \sum_{a+b=k} C_a C_b$$

(2) Consider now the generating series of the Catalan numbers, $f(z) = \sum_{k \geq 0} C_k z^k$. In terms of this generating series, the above recurrence gives, as desired:

$$\begin{aligned} zf^2 &= \sum_{a,b \geq 0} C_a C_b z^{a+b+1} \\ &= \sum_{k \geq 1} \sum_{a+b=k-1} C_a C_b z^k \\ &= \sum_{k \geq 1} C_k z^k \\ &= f - 1 \end{aligned}$$

(3) By solving the equation $zf^2 - f + 1 = 0$ found above, and choosing the solution which is bounded at $z = 0$, we obtain the following formula, as claimed:

$$f(z) = \frac{1 - \sqrt{1 - 4z}}{2z}$$

(4) In order to compute this function, we use the generalized binomial formula, which is as follows, with $p \in \mathbb{R}$ being an arbitrary exponent, and with $|t| < 1$:

$$(1+t)^p = \sum_{k=0}^{\infty} \binom{p}{k} t^k$$

To be more precise, this formula, which generalizes the usual binomial formula, holds indeed due to the Taylor formula, with the binomial coefficients being given by:

$$\binom{p}{k} = \frac{p(p-1) \dots (p-k+1)}{k!}$$

(5) For the exponent $p = 1/2$, the generalized binomial coefficients are:

$$\begin{aligned}
 \binom{1/2}{k} &= \frac{1/2(-1/2)(-3/2)\dots(3/2-k)}{k!} \\
 &= (-1)^{k-1} \frac{1 \cdot 3 \cdot 5 \dots (2k-3)}{2^k k!} \\
 &= (-1)^{k-1} \frac{(2k-2)!}{2^{k-1}(k-1)! 2^k k!} \\
 &= \frac{(-1)^{k-1}}{2^{2k-1}} \cdot \frac{1}{k} \binom{2k-2}{k-1} \\
 &= -2 \left(\frac{-1}{4}\right)^k \cdot \frac{1}{k} \binom{2k-2}{k-1}
 \end{aligned}$$

(6) Thus the generalized binomial formula at exponent $p = 1/2$ reads:

$$\sqrt{1+t} = 1 - 2 \sum_{k=1}^{\infty} \frac{1}{k} \binom{2k-2}{k-1} \left(\frac{-t}{4}\right)^k$$

With $t = -4z$ we obtain from this the following formula:

$$\sqrt{1-4z} = 1 - 2 \sum_{k=1}^{\infty} \frac{1}{k} \binom{2k-2}{k-1} z^k$$

(7) Now back to our series f , we obtain the following formula for it:

$$\begin{aligned}
 f(z) &= \frac{1 - \sqrt{1-4z}}{2z} \\
 &= \sum_{k=1}^{\infty} \frac{1}{k} \binom{2k-2}{k-1} z^{k-1} \\
 &= \sum_{k=0}^{\infty} \frac{1}{k+1} \binom{2k}{k} z^k
 \end{aligned}$$

(8) Thus the Catalan numbers are given by the formula the statement, namely:

$$C_k = \frac{1}{k+1} \binom{2k}{k}$$

So done, and note in passing that I kept my promise, from the proof of Proposition 13.4. Indeed, with the above final formula, the numerics are easily worked out. \square

Many other things can be said about the Catalan numbers, as a continuation of the above, and about the central binomial coefficients too. We will be back to this.

13c. Stieltjes inversion

According to the above, the problem is now, how to recover a probability measure out of its moments. And the answer here, which is something non-trivial, is as follows:

THEOREM 13.8. *The density of a real probability measure μ can be recaptured from the sequence of moments $\{M_k\}_{k \geq 0}$ via the Stieltjes inversion formula*

$$d\mu(x) = \lim_{t \searrow 0} -\frac{1}{\pi} \operatorname{Im} (G(x + it)) \cdot dx$$

where the function on the right, given in terms of moments by

$$G(\xi) = \xi^{-1} + M_1 \xi^{-2} + M_2 \xi^{-3} + \dots$$

is the Cauchy transform of the measure μ .

PROOF. The Cauchy transform of our measure μ is given by:

$$\begin{aligned} G(\xi) &= \xi^{-1} \sum_{k=0}^{\infty} M_k \xi^{-k} \\ &= \int_{\mathbb{R}} \frac{\xi^{-1}}{1 - \xi^{-1}y} d\mu(y) \\ &= \int_{\mathbb{R}} \frac{1}{\xi - y} d\mu(y) \end{aligned}$$

Now with $\xi = x + it$, we obtain the following formula:

$$\begin{aligned} \operatorname{Im}(G(x + it)) &= \int_{\mathbb{R}} \operatorname{Im} \left(\frac{1}{x - y + it} \right) d\mu(y) \\ &= \int_{\mathbb{R}} \frac{1}{2i} \left(\frac{1}{x - y + it} - \frac{1}{x - y - it} \right) d\mu(y) \\ &= - \int_{\mathbb{R}} \frac{t}{(x - y)^2 + t^2} d\mu(y) \end{aligned}$$

By integrating over $[a, b]$ we obtain, with the change of variables $x = y + tz$:

$$\begin{aligned} \int_a^b \operatorname{Im}(G(x + it)) dx &= - \int_{\mathbb{R}} \int_a^b \frac{t}{(x - y)^2 + t^2} dx d\mu(y) \\ &= - \int_{\mathbb{R}} \int_{(a-y)/t}^{(b-y)/t} \frac{t}{(tz)^2 + t^2} t dz d\mu(y) \\ &= - \int_{\mathbb{R}} \int_{(a-y)/t}^{(b-y)/t} \frac{1}{1 + z^2} dz d\mu(y) \\ &= - \int_{\mathbb{R}} \left(\arctan \frac{b-y}{t} - \arctan \frac{a-y}{t} \right) d\mu(y) \end{aligned}$$

Now observe that with $t \searrow 0$ we have:

$$\lim_{t \searrow 0} \left(\arctan \frac{b-y}{t} - \arctan \frac{a-y}{t} \right) = \begin{cases} \frac{\pi}{2} - \frac{\pi}{2} = 0 & (y < a) \\ \frac{\pi}{2} - 0 = \frac{\pi}{2} & (y = a) \\ \frac{\pi}{2} - (-\frac{\pi}{2}) = \pi & (a < y < b) \\ 0 - (-\frac{\pi}{2}) = \frac{\pi}{2} & (y = b) \\ -\frac{\pi}{2} - (-\frac{\pi}{2}) = 0 & (y > b) \end{cases}$$

We therefore obtain the following formula:

$$\lim_{t \searrow 0} \int_a^b \operatorname{Im}(G(x+it)) dx = -\pi \left(\mu(a, b) + \frac{\mu(a) + \mu(b)}{2} \right)$$

Thus, we are led to the conclusion in the statement. \square

Before getting further, let us mention that the above result does not fully solve the moment problem, because we still have the question of understanding when a sequence of numbers M_1, M_2, M_3, \dots can be the moments of a measure μ . We have here:

THEOREM 13.9. *A sequence of numbers $M_0, M_1, M_2, M_3, \dots \in \mathbb{R}$, with $M_0 = 1$, is the series of moments of a real probability measure μ precisely when:*

$$|M_0| \geq 0 \quad , \quad \begin{vmatrix} M_0 & M_1 \\ M_1 & M_2 \end{vmatrix} \geq 0 \quad , \quad \begin{vmatrix} M_0 & M_1 & M_2 \\ M_1 & M_2 & M_3 \\ M_2 & M_3 & M_4 \end{vmatrix} \geq 0 \quad , \quad \dots$$

That is, the associated Hankel determinants must be all positive.

PROOF. This is something a bit more advanced, the idea being as follows:

(1) As a first observation, the positivity conditions in the statement tell us that the following associated linear forms must be positive:

$$\sum_{i,j=1}^n c_i \bar{c}_j M_{i+j} \geq 0$$

(2) But this is something very classical, in one sense the result being elementary, coming from the following computation, which shows that we have positivity indeed:

$$\begin{aligned} \int_{\mathbb{R}} \left| \sum_{i=1}^n c_i x^i \right|^2 d\mu(x) &= \int_{\mathbb{R}} \sum_{i,j=1}^n c_i \bar{c}_j x^{i+j} d\mu(x) \\ &= \sum_{i,j=1}^n c_i \bar{c}_j M_{i+j} \end{aligned}$$

(3) As for the other sense, here the result comes once again from the above formula, this time via some standard functional analysis. \square

As a basic application of the Stieltjes formula, let us solve the moment problem for the Catalan numbers C_k , and for the central binomial coefficients D_k . We first have:

THEOREM 13.10. *The real measure having as even moments the Catalan numbers, $C_k = \frac{1}{k+1} \binom{2k}{k}$, and having all odd moments 0 is the measure*

$$\gamma_1 = \frac{1}{2\pi} \sqrt{4 - x^2} dx$$

called Wigner semicircle law on $[-2, 2]$.

PROOF. In order to apply the inversion formula, our starting point will be the formula from Theorem 13.7 for the generating series of the Catalan numbers, namely:

$$\sum_{k=0}^{\infty} C_k z^k = \frac{1 - \sqrt{1 - 4z}}{2z}$$

By using this formula with $z = \xi^{-2}$, we obtain the following formula:

$$\begin{aligned} G(\xi) &= \xi^{-1} \sum_{k=0}^{\infty} C_k \xi^{-2k} \\ &= \xi^{-1} \cdot \frac{1 - \sqrt{1 - 4\xi^{-2}}}{2\xi^{-2}} \\ &= \frac{\xi}{2} \left(1 - \sqrt{1 - 4\xi^{-2}} \right) \\ &= \frac{\xi}{2} - \frac{1}{2} \sqrt{\xi^2 - 4} \end{aligned}$$

Now let us apply Theorem 13.8. The study here goes as follows:

(1) According to the general philosophy of the Stieltjes formula, the first term, namely $\xi/2$, which is “trivial”, will not contribute to the density.

(2) As for the second term, which is something non-trivial, this will contribute to the density, the rule here being that the square root $\sqrt{\xi^2 - 4}$ will be replaced by the “dual” square root $\sqrt{4 - x^2} dx$, and that we have to multiply everything by $-1/\pi$.

(3) As a conclusion, by Stieltjes inversion we obtain the following density:

$$d\mu(x) = -\frac{1}{\pi} \cdot -\frac{1}{2} \sqrt{4 - x^2} dx = \frac{1}{2\pi} \sqrt{4 - x^2} dx$$

Thus, we have obtained the measure in the statement, and we are done. \square

We have the following version of the above result:

THEOREM 13.11. *The real measure having as sequence of moments the Catalan numbers, $C_k = \frac{1}{k+1} \binom{2k}{k}$, is the measure*

$$\pi_1 = \frac{1}{2\pi} \sqrt{4x^{-1} - 1} dx$$

called Marchenko-Pastur law on $[0, 4]$.

PROOF. As before, we use the standard formula for the generating series of the Catalan numbers. With $z = \xi^{-1}$ in that formula, we obtain the following formula:

$$\begin{aligned} G(\xi) &= \xi^{-1} \sum_{k=0}^{\infty} C_k \xi^{-k} \\ &= \xi^{-1} \cdot \frac{1 - \sqrt{1 - 4\xi^{-1}}}{2\xi^{-1}} \\ &= \frac{1}{2} \left(1 - \sqrt{1 - 4\xi^{-1}} \right) \\ &= \frac{1}{2} - \frac{1}{2} \sqrt{1 - 4\xi^{-1}} \end{aligned}$$

With this in hand, let us apply now the Stieltjes inversion formula, from Theorem 13.8. We obtain, a bit as before in Theorem 13.10, the following density:

$$d\mu(x) = -\frac{1}{\pi} \cdot -\frac{1}{2} \sqrt{4x^{-1} - 1} dx = \frac{1}{2\pi} \sqrt{4x^{-1} - 1} dx$$

Thus, we are led to the conclusion in the statement. □

Regarding now the central binomial coefficients, we have here:

THEOREM 13.12. *The real probability measure having as moments the central binomial coefficients, $D_k = \binom{2k}{k}$, is the measure*

$$\alpha_1 = \frac{1}{\pi \sqrt{x(4-x)}} dx$$

called arcsine law on $[0, 4]$.

PROOF. We have the following computation, using some standard formulae:

$$\begin{aligned}
 G(\xi) &= \xi^{-1} \sum_{k=0}^{\infty} D_k \xi^{-k} \\
 &= \frac{1}{\xi} \sum_{k=0}^{\infty} D_k \left(-\frac{t}{4}\right)^k \\
 &= \frac{1}{\xi} \cdot \frac{1}{\sqrt{1-4/\xi}} \\
 &= \frac{1}{\sqrt{\xi(\xi-4)}}
 \end{aligned}$$

But this gives the density in the statement, via Theorem 13.8. \square

Finally, we have the following version of the above result:

THEOREM 13.13. *The real probability measure having as moments the middle binomial coefficients, $E_k = \binom{k}{[k/2]}$, is the following law on $[-2, 2]$,*

$$\sigma_1 = \frac{1}{2\pi} \sqrt{\frac{2+x}{2-x}} dx$$

called modified arcsine law on $[-2, 2]$.

PROOF. In terms of the central binomial coefficients D_k , we have:

$$E_{2k} = \binom{2k}{k} = \frac{(2k)!}{k!k!} = D_k$$

$$E_{2k-1} = \binom{2k-1}{k} = \frac{(2k-1)!}{k!(k-1)!} = \frac{D_k}{2}$$

Standard calculus based on the Taylor formula for $(1+t)^{-1/2}$ gives:

$$\frac{1}{2x} \left(\sqrt{\frac{1+2x}{1-2x}} - 1 \right) = \sum_{k=0}^{\infty} E_k x^k$$

With $x = \xi^{-1}$ we obtain the following formula for the Cauchy transform:

$$\begin{aligned} G(\xi) &= \xi^{-1} \sum_{k=0}^{\infty} E_k \xi^{-k} \\ &= \frac{1}{\xi} \left(\sqrt{\frac{1 + 2/\xi}{1 - 2/\xi}} - 1 \right) \\ &= \frac{1}{\xi} \left(\sqrt{\frac{\xi + 2}{\xi - 2}} - 1 \right) \end{aligned}$$

By Stieltjes inversion we obtain the density in the statement. \square

All this is very nice, and we are obviously building here, as this book goes by, some solid knowledge in classical probability. We will be back to all this later.

13d. Finite graphs

With the above done, we can come back now to walks on finite graphs, that we know from the above to be related to the eigenvalues of the adjacency matrix $d \in M_N(0, 1)$. But here, we are led to the following philosophical question, to start with:

QUESTION 13.14. *What are the most important finite graphs, that we should do our computations for?*

Not an easy question, you have to agree with me, with the answer to this obviously depending on your previous experience with mathematics, or physics, or chemistry, or computer science, or other branch of science that you are interested in, and also, on the specific problems that you are the most in love with, in that part of science.

So, we have to be subjective here. And with me writing this book, and doing some sort of complicated quantum physics, as daytime job, I will choose the ADE graphs. It is beyond our scope here to explain where these ADE graphs exactly come from, and what they are good for, but as a piece of advertisement for them, we have:

FACT 13.15. *The ADE graphs classify the following:*

- (1) *Basic Lie groups and algebras.*
- (2) *Subgroups of SU_2 and of SO_3 .*
- (3) *Singularities of algebraic manifolds.*
- (4) *Basic invariants of knots and links.*
- (5) *Subfactors and planar algebras of small index.*
- (6) *Subgroups of the quantum permutation group S_4^+ .*
- (7) *Basic quantum field theories, and other physics beasts.*

Which sounds exciting, doesn't it. So, have a look at this, and with the comment that some heavy learning work is needed, in order to understand how all this works. And with the extra comment that, in view of (7), tough physics, no one really understands how all this works. A nice introduction to all this is the book of Jones [56].

Getting to work now, we first need to know what the ADE graphs are. The A graphs, which are the simplest, are as follows, with the distinguished vertex being denoted \bullet , and with A_n having $n \geq 2$ vertices, and \tilde{A}_{2n} having $2n \geq 2$ vertices:

$$\begin{array}{ll} A_n = \bullet - \circ - \circ - \circ \cdots \circ - \circ - \circ & A_\infty = \bullet - \circ - \circ - \circ - \circ \cdots \\ \tilde{A}_{2n} = \begin{array}{c} \circ - \circ - \circ \cdots \circ - \circ - \circ \\ | \qquad \qquad \qquad | \\ \bullet - \circ - \circ - \circ - \circ - \circ - \circ \end{array} & \tilde{A}_\infty = \begin{array}{c} \circ - \circ - \circ - \circ \cdots \\ | \\ \bullet - \circ - \circ - \circ - \circ \cdots \end{array} \end{array}$$

You might probably say, why not stopping here, and doing our unfinished business for the segment and the circle, with whatever new ideas that we might have. Good point, but in answer, these ideas will apply as well, with minimal changes, to the D graphs, which are as follows, with D_n having $n \geq 3$ vertices, and \tilde{D}_n having $n + 1 \geq 5$ vertices:

$$\begin{array}{ll} D_n = \bullet - \circ - \circ - \circ \cdots \circ - \begin{array}{c} \circ \\ | \\ \circ \end{array} - \circ & \\ \tilde{D}_n = \bullet - \begin{array}{c} \circ \\ | \\ \circ \end{array} - \circ - \circ \cdots \circ - \begin{array}{c} \circ \\ | \\ \circ \end{array} - \circ & \\ D_\infty = \bullet - \begin{array}{c} \circ \\ | \\ \circ \end{array} - \circ - \circ - \circ \cdots & \end{array}$$

As a comment here, the labeling conventions for the AD graphs, while very standard, can be a bit confusing. The first graph in each series is by definition as follows:

$$\begin{array}{llll} A_2 = \bullet - \circ & \tilde{A}_2 = \begin{array}{c} \circ \\ || \\ \bullet \end{array} & D_3 = \begin{array}{c} \circ \\ | \\ \bullet - \circ \end{array} & \tilde{D}_4 = \begin{array}{c} \circ \quad \circ \\ \backslash \quad / \\ \bullet - \circ - \circ \end{array} \end{array}$$

Finally, there are also a number of exceptional ADE graphs. First we have:

$$\begin{aligned}
 E_6 &= \bullet - \circ - \overset{\circ}{\underset{|}{\circ}} - \circ - \circ \\
 E_7 &= \bullet - \circ - \circ - \overset{\circ}{\underset{|}{\circ}} - \circ - \circ \\
 E_8 &= \bullet - \circ - \circ - \circ - \circ - \overset{\circ}{\underset{|}{\circ}} - \circ
 \end{aligned}$$

Then, we have extended versions of the above exceptional graphs, as follows:

$$\begin{aligned}
 \tilde{E}_6 &= \bullet - \circ - \overset{\circ}{\underset{\circ}{\underset{|}{\circ}}} - \circ - \circ \\
 \tilde{E}_7 &= \bullet - \circ - \circ - \overset{\circ}{\underset{|}{\circ}} - \circ - \circ - \circ \\
 \tilde{E}_8 &= \bullet - \circ - \circ - \circ - \circ - \circ - \overset{\circ}{\underset{|}{\circ}} - \circ
 \end{aligned}$$

And good news, according to the general theory, that is all. Getting now to work, we have plenty of graphs here, and the problem that we would like to solve is:

PROBLEM 13.16. *How to count loops on the ADE graphs?*

In answer, we know that A_∞ and \tilde{A}_∞ are respectively the graphs that we previously called \mathbb{N} and \mathbb{Z} . So, based on our previous work for these graphs, where the combinatorics naturally led us into generating series, let us formulate the following definition:

DEFINITION 13.17. *The Poincaré series of a rooted bipartite graph X is*

$$f(z) = \sum_{k=0}^{\infty} L_{2k} z^k$$

where L_{2k} is the number of $2k$ -loops based at the root.

To be more precise, observe that all the above ADE graphs are indeed bipartite. Now the point is that, for a bipartite graph, the loops based at any point must have even length. Thus, in order to study the loops on the ADE graphs, based at the root, we just

have to count the above numbers L_{2k} . And then, considering the generating series $f(z)$ of these numbers, and calling this Poincaré series, is something very standard.

Before getting into computations, let us introduce as well:

DEFINITION 13.18. *The positive spectral measure μ of a rooted bipartite graph X is the real probability measure having the numbers L_{2k} as moments:*

$$\int_{\mathbb{R}} x^k d\mu(x) = L_{2k}$$

Equivalently, we must have the Stieltjes transform formula

$$f(z) = \int_{\mathbb{R}} \frac{1}{1 - xz} d\mu(x)$$

where f is the Poincaré series of X .

Here the existence of μ , and the fact that this is indeed a positive measure, meaning a measure supported on $[0, \infty)$, comes from the following simple fact:

THEOREM 13.19. *The positive spectral measure of a rooted bipartite graph X is given by the following formula, with d being the adjacency matrix of the graph,*

$$\mu = \text{law}(d^2)$$

and with the probabilistic computation being with respect to the expectation

$$A \rightarrow \langle A \rangle$$

with $\langle A \rangle$ being the $(, *)$ -entry of a matrix A , where $*$ is the root.*

PROOF. With the above conventions, we have the following computation:

$$\begin{aligned} f(z) &= \sum_{k=0}^{\infty} L_{2k} z^k \\ &= \sum_{k=0}^{\infty} \langle d^{2k} \rangle z^k \\ &= \left\langle \frac{1}{1 - d^2 z} \right\rangle \end{aligned}$$

But this shows that we have $\mu = \text{law}(d^2)$, as desired. \square

The above result shows that computing μ might be actually a simpler problem than computing f , and in practice, this is indeed the case. So, in what follows we will rather forget about loops and Definition 13.17, and use Definition 13.18 instead, with our computations to follow being based on the concrete interpretation from Theorem 13.19.

However, even with this probabilistic trick in our bag, things are not exactly trivial. Let us introduce as well the following notion:

DEFINITION 13.20. *The circular measure ε of a rooted bipartite graph X is given by*

$$d\varepsilon(q) = d\mu((q + q^{-1})^2)$$

where μ is the associated positive spectral measure.

To be more precise, we know from Theorem 13.19 that the positive measure μ is the spectral measure of a certain positive matrix, $d^2 \geq 0$, and it follows from this, and from basic spectral theory, that this measure is supported by the positive reals:

$$\text{supp}(\mu) \subset \mathbb{R}_+$$

But then, with this observation in hand, we can define indeed the circular measure ε as above, as being the pullback of μ via the following map:

$$\mathbb{R} \cup \mathbb{T} \rightarrow \mathbb{R}_+ \quad , \quad q \rightarrow (q + q^{-1})^2$$

As a basic example for this, to start with, assume that μ is a discrete measure, supported by n positive numbers $x_1 < \dots < x_n$, with corresponding densities p_1, \dots, p_n :

$$\mu = \sum_{i=1}^n p_i \delta_{x_i}$$

For each $i \in \{1, \dots, n\}$ the equation $(q + q^{-1})^2 = x_i$ has then four solutions, that we can denote $q_i, q_i^{-1}, -q_i, -q_i^{-1}$. And with this notation, we have:

$$\varepsilon = \frac{1}{4} \sum_{i=1}^n p_i \left(\delta_{q_i} + \delta_{q_i^{-1}} + \delta_{-q_i} + \delta_{-q_i^{-1}} \right)$$

In general, the basic properties of ε can be summarized as follows:

THEOREM 13.21. *The circular measure has the following properties:*

- (1) ε has equal density at $q, q^{-1}, -q, -q^{-1}$.
- (2) The odd moments of ε are 0.
- (3) The even moments of ε are half-integers.
- (4) When X has norm ≤ 2 , ε is supported by the unit circle.
- (5) When X is finite, ε is discrete.
- (6) If K is a solution of $d = K + K^{-1}$, then $\varepsilon = \text{law}(K)$.

PROOF. These results can be deduced from definitions, the idea being that (1-5) are trivial, and that (6) follows from the formula of μ from Theorem 13.19. \square

Getting now to computations, we first have the following result:

THEOREM 13.22. *The circular measure of the basic index 4 graph, namely*

$$\tilde{A}_{2n} = \begin{array}{cccccccc} \circ & - & \circ & - & \circ & \cdots & \circ & - & \circ & - & \circ \\ | & & & & & & & & & & | \\ \bullet & - & \circ & - & \circ & - & \circ & - & \circ & - & \circ \end{array}$$

is the uniform measure on the $2n$ -roots of unity.

PROOF. Let us identify the vertices of $X = \tilde{A}_{2n}$ with the group $\{w^k\}$ formed by the $2n$ -th roots of unity in the complex plane, where $w = e^{\pi i/n}$. The adjacency matrix of X acts then on the functions $f \in C(X)$ in the following way:

$$df(w^s) = f(w^{s-1}) + f(w^{s+1})$$

But this shows that we have $d = K + K^{-1}$, where K is given by:

$$Kf(w^s) = f(w^{s+1})$$

Thus we can use Theorem 13.19 and Theorem 13.21 (6), and we get:

$$\varepsilon = law(K)$$

But this is the uniform measure on the $2n$ -roots of unity, as claimed. \square

All this is nice, but before going ahead with more computations, let us have a look into subfactor theory, and explain what is behind this. Following Jones [56], we have:

DEFINITION 13.23. *The theta series of a rooted bipartite graph X is*

$$\Theta(q) = q + \frac{1-q}{1+q} f\left(\frac{q}{(1+q)^2}\right)$$

where f is the Poincaré series.

The theta series can be written as $\Theta(q) = \sum a_r q^r$, and it follows from the above formula, via some simple manipulations, that its coefficients are integers:

$$a_r \in \mathbb{Z}$$

When X is the principal graph of a subfactor of index $N > 4$, it is known that the numbers a_r are certain multiplicities associated to the planar algebra inclusion $TL_N \subset P$. In particular, the coefficients of the theta series are in this case positive integers:

$$a_r \in \mathbb{N}$$

In relation now with the circular measure, the result here, which is quite similar to the Stieltjes transform formula from Definition 13.18, is as follows:

THEOREM 13.24. *We have the Stieltjes transform type formula*

$$2 \int \frac{1}{1 - qu^2} d\varepsilon(u) = 1 + T(q)(1 - q)$$

where the T series of a rooted bipartite graph X is by definition given by

$$T(q) = \frac{\Theta(q) - q}{1 - q}$$

with Θ being the associated theta series.

PROOF. This follows indeed by applying the change of variables $q \rightarrow (q + q^{-1})^2$ to the fact that f is the Stieltjes transform of μ . \square

In order to discuss all this more systematically, let us introduce as well:

DEFINITION 13.25. *The series of the form*

$$\xi(n_1, \dots, n_s : m_1, \dots, m_t) = \frac{(1 - q^{n_1}) \dots (1 - q^{n_s})}{(1 - q^{m_1}) \dots (1 - q^{m_t})}$$

with $n_i, m_i \in \mathbb{N}$ are called cyclotomic.

It is technically convenient to allow as well $1 + q^n$ factors, to be designated by n^+ symbols in the above writing. For instance we have, by definition:

$$\xi(2^+ : 3) = \xi(4 : 2, 3)$$

Also, it is convenient in what follows to use the following notations:

$$\xi' = \frac{\xi}{1 - q} \quad , \quad \xi'' = \frac{\xi}{1 - q^2}$$

The Poincaré series of the ADE graphs are given by quite complicated formulae. However, the corresponding T series are all cyclotomic, as follows:

THEOREM 13.26. *The T series of the ADE graphs are as follows:*

- (1) For A_{n-1} we have $T = \xi(n - 1 : n)$.
- (2) For D_{n+1} we have $T = \xi(n - 1^+ : n^+)$.
- (3) For \tilde{A}_{2n} we have $T = \xi'(n^+ : n)$.
- (4) For \tilde{D}_{n+2} we have $T = \xi''(n + 1^+ : n)$.
- (5) For E_6 we have $T = \xi(8 : 3, 6^+)$.
- (6) For E_7 we have $T = \xi(12 : 4, 9^+)$.
- (7) For E_8 we have $T = \xi(5^+, 9^+ : 15^+)$.
- (8) For \tilde{E}_6 we have $T = \xi(6^+ : 3, 4)$.
- (9) For \tilde{E}_7 we have $T = \xi(9^+ : 4, 6)$.
- (10) For \tilde{E}_8 we have $T = \xi(15^+ : 6, 10)$.

PROOF. These formulae can be obtained by counting loops, and then by making the following change of variables, and factorizing the resulting series:

$$z^{-1/2} = q^{1/2} + q^{-1/2}$$

An alternative proof can be obtained by using planar algebra methods. \square

Our purpose now will be that of converting the above technical results, regarding the T series, into some final results, regarding the corresponding circular measures ε . In order to formulate our results, we will need some more theory. First, we have:

DEFINITION 13.27. *A cyclotomic measure is a probability measure ε on the unit circle, having the following properties:*

- (1) ε is supported by the $2n$ -roots of unity, for some $n \in \mathbb{N}$.
- (2) ε has equal density at $q, q^{-1}, -q, -q^{-1}$.

As a first observation, it follows from Theorem 13.21 and from Theorem 13.26 that the circular measures of the finite ADE graphs are supported by certain roots of unity, hence are cyclotomic. We will be back to this in a moment, with details, and computations.

At the general level now, let us introduce as well the following notion:

DEFINITION 13.28. *The T series of a cyclotomic measure ε is given by*

$$1 + T(q)(1 - q) = 2 \int \frac{1}{1 - qu^2} d\varepsilon(u)$$

with ε being as usual the circular spectral measure.

Observe that this formula is nothing but the one in Theorem 13.24, written now in the other sense. In other words, if the cyclotomic measure ε happens to be the circular measure of a rooted bipartite graph, then the T series as defined above coincides with the T series as defined before. This is useful for explicit computations.

Good news, with this technology in hand, and with a computation already done, in Theorem 13.22, we are now ready to discuss the circular measures of all ADE graphs.

The idea will be that these measures are all cyclotomic, of level ≤ 3 , and can be expressed in terms of the basic polynomial densities of degree ≤ 6 , namely:

$$\alpha = Re(1 - q^2)$$

$$\beta = Re(1 - q^4)$$

$$\gamma = Re(1 - q^6)$$

To be more precise, we have the following final result on the subject, with α, β, γ being as above, with d_n being the uniform measure on the $2n$ -th roots of unity, and with $d'_n = 2d_{2n} - d_n$ being the uniform measure on the odd $4n$ -roots of unity:

THEOREM 13.29. *The circular measures of the ADE graphs are given by:*

- (1) $A_{n-1} \rightarrow \alpha_n$.
- (2) $\tilde{A}_{2n} \rightarrow d_n$.
- (3) $D_{n+1} \rightarrow \alpha'_n$.
- (4) $\tilde{D}_{n+2} \rightarrow (d_n + d'_1)/2$.
- (5) $E_6 \rightarrow \alpha_{12} + (d_{12} - d_6 - d_4 + d_3)/2$.
- (6) $E_7 \rightarrow \beta'_9 + (d'_1 - d'_3)/2$.
- (7) $E_8 \rightarrow \alpha'_{15} + \gamma'_{15} - (d'_5 + d'_3)/2$.
- (8) $\tilde{E}_{n+3} \rightarrow (d_n + d_3 + d_2 - d_1)/2$.

PROOF. This is something which can be proved in three steps, as follows:

- (1) For the simplest graph, namely the circle \tilde{A}_{2n} , we already have the result, from Theorem 13.22, with the proof there being something elementary.
- (2) For the other non-exceptional graphs, that is, of type A and D, the same method works, namely direct loop counting, with some matrix tricks.
- (3) In general, this follows from the T series formulae in Theorem 13.26, via some manipulations based on the general conversion formulae given above. \square

We refer to [9] and the subfactor literature for more on all this. Also, let us point out that all this leads to a more conceptual understanding of what we did before, for the graphs \mathbb{N} and \mathbb{Z} . Indeed, even for these very basic graphs, using the unit circle and circular measures as above leads to a better understanding of the combinatorics.

13e. Exercises

This was a rather elementary chapter, and as exercises on this, we have:

EXERCISE 13.30. *Work out loop count results for various products of graphs.*

EXERCISE 13.31. *Learn more about Catalan numbers, the more, the better.*

EXERCISE 13.32. *Clarify what happens to the atoms, in the Stieltjes formula.*

EXERCISE 13.33. *Learn more about Hankel determinants, and the moment problem.*

EXERCISE 13.34. *Learn about the Wigner law, and its various remarkable properties.*

EXERCISE 13.35. *Learn about the Marchenko-Pastur law, and its various properties.*

EXERCISE 13.36. *Learn as well about the arcsine law, and its various properties.*

EXERCISE 13.37. *Learn about various ADE classifications, in math and physics.*

As bonus exercise, learn about subfactors, say from the book of Jones [56].

CHAPTER 14

Haar integration

14a. Compact groups

We have seen so far the foundations and basic results of measure theory and probability. Before stepping into more complicated things, such as random matrices and free probability, we would like to clarify one important question which appeared several times, namely the computation of integrals over the compact groups of unitary matrices $G \subset U_N$, and its probabilistic consequences. The precise question that we have in mind is:

QUESTION 14.1. *Given a compact group $G \subset U_N$, how to compute the integrals*

$$I_{ij}^e = \int_G g_{i_1 j_1}^{e_1} \cdots g_{i_k j_k}^{e_k} dg$$

and then, how to use their formula in order to compute the laws of variables of type

$$f_P = P\left(\{g_{ij}\}_{i,j=1,\dots,N}\right)$$

depending on a polynomial P ? What about the $N \rightarrow \infty$ asymptotics of such laws?

All this is quite subtle, and as a basic illustration for this, we have a fundamental result from chapter 3, stating that for $G = S_N$ the law of the variable $\chi = \sum_i g_{ii}$ can be explicitly computed, and becomes Poisson (1) with $N \rightarrow \infty$. This is something truly remarkable, and it is this kind of result that we would like to systematically have.

We will discuss this in this whole chapter, and later on too. This might seem of course quite long, but believe me, it is worth the effort, because it is quite hard to do any type of advanced probability theory without knowing the answer to Question 14.1. But probably enough advertisement, let us get to work. Following Weyl, we first have:

DEFINITION 14.2. *A unitary representation of a compact group G is a continuous group morphism into a unitary group*

$$v : G \rightarrow U_N \quad , \quad g \rightarrow v_g$$

which can be faithful or not. The character of such a representation is the function

$$\chi : G \rightarrow \mathbb{C} \quad , \quad g \rightarrow \text{Tr}(v_g)$$

where Tr is the usual, unnormalized trace of the $N \times N$ matrices.

At the level of examples, most of the compact groups that we met so far, finite or continuous, naturally appear as closed subgroups $G \subset U_N$. In this case, the embedding $G \subset U_N$ is of course a representation, called fundamental representation. In general now, let us first discuss the various operations on the representations. We have here:

PROPOSITION 14.3. *The representations of a compact group G are subject to:*

- (1) *Making sums. Given representations v, w , of dimensions N, M , their sum is the $N + M$ -dimensional representation $v + w = \text{diag}(v, w)$.*
- (2) *Making products. Given representations v, w , of dimensions N, M , their product is the NM -dimensional representation $(v \otimes w)_{ia,jb} = v_{ij}w_{ab}$.*
- (3) *Taking conjugates. Given a N -dimensional representation v , its conjugate is the N -dimensional representation $(\bar{v})_{ij} = \bar{v}_{ij}$.*
- (4) *Spinning by unitaries. Given a N -dimensional representation v , and a unitary $U \in U_N$, we can spin v by this unitary, $v \rightarrow UvU^*$.*

PROOF. The fact that the operations in the statement are indeed well-defined, among morphisms from G to unitary groups, is indeed clear from definitions. \square

In relation now with characters, we have the following result:

PROPOSITION 14.4. *We have the following formulae, regarding characters*

$$\chi_{v+w} = \chi_v + \chi_w \quad , \quad \chi_{v \otimes w} = \chi_v \chi_w \quad , \quad \chi_{\bar{v}} = \bar{\chi}_v \quad , \quad \chi_{UvU^*} = \chi_v$$

in relation with the basic operations for the representations.

PROOF. All these assertions are elementary, by using the following well-known trace formulae, valid for any square matrices V, W , and any unitary U :

$$\begin{aligned} \text{Tr}(\text{diag}(V, W)) &= \text{Tr}(V) + \text{Tr}(W) \quad , \quad \text{Tr}(V \otimes W) = \text{Tr}(V)\text{Tr}(W) \\ \text{Tr}(\bar{V}) &= \overline{\text{Tr}(V)} \quad , \quad \text{Tr}(UVU^*) = \text{Tr}(V) \end{aligned}$$

Thus, we are led to the formulae in the statement. \square

Assume now that we are given a closed subgroup $G \subset U_N$. By using the above operations, we can construct a whole family of representations of G , as follows:

DEFINITION 14.5. *Given a closed subgroup $G \subset U_N$, its Peter-Weyl representations are the various tensor products between the fundamental representation and its conjugate:*

$$v : G \subset U_N \quad , \quad \bar{v} : G \subset U_N$$

We denote these tensor products $v^{\otimes k}$, with $k = \circ \bullet \bullet \circ \dots$ being a colored integer, with the colored tensor powers being defined according to the rules

$$v^{\otimes \circ} = v \quad , \quad v^{\otimes \bullet} = \bar{v} \quad , \quad v^{\otimes kl} = v^{\otimes k} \otimes v^{\otimes l}$$

and with the convention that $v^{\otimes \emptyset}$ is the trivial representation $1 : G \rightarrow U_1$.

Here are a few examples of such representations, namely those coming from the colored integers of length 2, which will often appear in what follows:

$$\begin{aligned} v^{\otimes \circ \circ} &= v \otimes v \quad , \quad v^{\otimes \circ \bullet} = v \otimes \bar{v} \\ v^{\otimes \bullet \circ} &= \bar{v} \otimes v \quad , \quad v^{\otimes \bullet \bullet} = \bar{v} \otimes \bar{v} \end{aligned}$$

In relation now with characters, we have the following result:

PROPOSITION 14.6. *The characters of the Peter-Weyl representations are given by*

$$\chi_{v^{\otimes k}} = (\chi_v)^k$$

with the colored powers being given by $\chi^\circ = \chi$, $\chi^\bullet = \bar{\chi}$ and multiplicativity.

PROOF. This follows indeed from the additivity, multiplicativity and conjugation formulae from Proposition 14.4, via the conventions in Definition 14.5. \square

Getting back now to our motivations, we can see the interest in the above constructions. Indeed, the joint moments of the main character $\chi = \chi_v$ and its adjoint $\bar{\chi} = \chi_{\bar{v}}$ are the expectations of the characters of various Peter-Weyl representations:

$$\int_G \chi^k = \int_G \chi_{v^{\otimes k}}$$

In order to advance, we must develop some general theory. Let us start with:

DEFINITION 14.7. *Given a compact group G , and two of its representations,*

$$v : G \rightarrow U_N \quad , \quad w : G \rightarrow U_M$$

we define the space of intertwiners between these representations as being

$$\text{Hom}(v, w) = \left\{ T \in M_{M \times N}(\mathbb{C}) \mid T v_g = w_g T, \forall g \in G \right\}$$

and we use the following conventions:

- (1) *We use the notations $\text{Fix}(v) = \text{Hom}(1, v)$, and $\text{End}(v) = \text{Hom}(v, v)$.*
- (2) *We write $v \sim w$ when $\text{Hom}(v, w)$ contains an invertible element.*
- (3) *We say that v is irreducible, and write $v \in \text{Irr}(G)$, when $\text{End}(v) = \mathbb{C}1$.*

The terminology here is standard, with Fix , Hom , End standing for fixed points, homomorphisms and endomorphisms. We will see later that irreducible means indecomposable, in a suitable sense. Here are now a few basic results, regarding these spaces:

PROPOSITION 14.8. *The spaces of intertwiners have the following properties:*

- (1) $T \in \text{Hom}(v, w), S \in \text{Hom}(w, z) \implies ST \in \text{Hom}(v, z)$.
- (2) $S \in \text{Hom}(v, w), T \in \text{Hom}(z, t) \implies S \otimes T \in \text{Hom}(v \otimes z, w \otimes t)$.
- (3) $T \in \text{Hom}(v, w) \implies T^* \in \text{Hom}(w, v)$.

In abstract terms, we say that the Hom spaces form a tensor $$ -category.*

PROOF. All the formulae in the statement are indeed clear from definitions, via elementary computations. As for the last assertion, this is something coming from (1,2,3). We will be back to tensor categories later on, with more details on this latter fact. \square

As a main consequence of the above result, we have:

PROPOSITION 14.9. *Given a representation $v : G \rightarrow U_N$, the linear space*

$$\text{End}(v) \subset M_N(\mathbb{C})$$

is a $$ -algebra, with respect to the usual involution of the matrices.*

PROOF. By definition, $\text{End}(v)$ is a linear subspace of $M_N(\mathbb{C})$. We know from Proposition 14.8 (1) that this subspace $\text{End}(v)$ is a subalgebra of $M_N(\mathbb{C})$, and then we know as well from Proposition 14.8 (3) that this subalgebra is stable under the involution $*$. Thus, what we have here is a $*$ -subalgebra of $M_N(\mathbb{C})$, as claimed. \square

In order to exploit the above fact, we will need a basic result from linear algebra, stating that any $*$ -algebra $A \subset M_N(\mathbb{C})$ decomposes as a direct sum, as follows:

$$A \simeq M_{N_1}(\mathbb{C}) \oplus \dots \oplus M_{N_k}(\mathbb{C})$$

Indeed, let us write the unit $1 \in A$ as $1 = p_1 + \dots + p_k$, with $p_i \in A$ being central minimal projections. Then each of the spaces $A_i = p_i A p_i$ is a subalgebra of A , and we have a decomposition $A = A_1 \oplus \dots \oplus A_k$. But since each central projection $p_i \in A$ was chosen minimal, we have $A_i \simeq M_{N_i}(\mathbb{C})$, with $N_i = \text{rank}(p_i)$, as desired.

We can now formulate our first Peter-Weyl type theorem, as follows:

THEOREM 14.10 (Peter-Weyl 1). *Let $v : G \rightarrow U_N$ be a representation, consider the algebra $A = \text{End}(v)$, and write its unit $1 = p_1 + \dots + p_k$ as above. We have then*

$$v = v_1 + \dots + v_k$$

with each v_i being an irreducible representation, obtained by restricting v to $\text{Im}(p_i)$.

PROOF. This basically follows from Proposition 14.9, as follows:

(1) We first associate to our representation $v : G \rightarrow U_N$ the corresponding action map on \mathbb{C}^N . If a linear subspace $W \subset \mathbb{C}^N$ is invariant, the restriction of the action map to W is an action map too, which must come from a subrepresentation $w \subset v$.

(2) Consider now a projection $p \in \text{End}(v)$. From $pv = vp$ we obtain that the linear space $W = \text{Im}(p)$ is invariant under v , and so this space must come from a subrepresentation $w \subset v$. It is routine to check that the operation $p \rightarrow w$ maps subprojections to subrepresentations, and minimal projections to irreducible representations.

(3) With these preliminaries in hand, let us decompose the algebra $\text{End}(v)$ as above, by using the decomposition $1 = p_1 + \dots + p_k$ into central minimal projections. If we

denote by $v_i \subset v$ the subrepresentation coming from the vector space $V_i = \text{Im}(p_i)$, then we obtain in this way a decomposition $v = v_1 + \dots + v_k$, as in the statement. \square

Here is now our second Peter-Weyl theorem, complementing Theorem 14.10:

THEOREM 14.11 (Peter-Weyl 2). *Given a closed subgroup $G \subset_v U_N$, any of its irreducible smooth representations*

$$w : G \rightarrow U_M$$

appears inside a tensor product of the fundamental representation v and its adjoint \bar{v} .

PROOF. Given a representation $w : G \rightarrow U_M$, we define the space of coefficients $C_w \subset C(G)$ of this representation as being the following linear space:

$$C_w = \text{span} \left[g \rightarrow w(g)_{ij} \right]$$

With this notion in hand, the result can be deduced as follows:

(1) The construction $w \rightarrow C_w$ is functorial, in the sense that it maps subrepresentations into linear subspaces. This is indeed something which is routine to check.

(2) A closed subgroup $G \subset_v U_N$ is a Lie group, and a representation $w : G \rightarrow U_M$ is smooth when we have an inclusion $C_w \subset \langle C_v \rangle$. This is indeed well-known.

(3) By definition of the Peter-Weyl representations, as arbitrary tensor products between the fundamental representation v and its conjugate \bar{v} , we have:

$$\langle C_v \rangle = \sum_k C_{v^{\otimes k}}$$

(4) Now by putting together the above observations (2,3) we conclude that we must have an inclusion as follows, for certain exponents k_1, \dots, k_p :

$$C_w \subset C_{v^{\otimes k_1} \oplus \dots \oplus v^{\otimes k_p}}$$

(5) By using now (1), we deduce that we have an inclusion $w \subset v^{\otimes k_1} \oplus \dots \oplus v^{\otimes k_p}$, and by applying Theorem 14.10, this leads to the conclusion in the statement. \square

14b. Haar integration

In order to further advance with Peter-Weyl theory, we need to talk about integration over G . In the finite group case the situation is trivial, as follows:

PROPOSITION 14.12. *Any finite group G has a unique probability measure which is invariant under left and right translations,*

$$\mu(E) = \mu(gE) = \mu(Eg)$$

and this is the normalized counting measure on G , given by $\mu(E) = |E|/|G|$.

PROOF. This is indeed something trivial, which follows from definitions. \square

In the general, continuous case, let us begin with the following key result:

PROPOSITION 14.13. *Given a unital positive linear form $\psi : C(G) \rightarrow \mathbb{C}$, the limit*

$$\int_{\varphi} f = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n \psi^{*k}(f)$$

exists, and for a coefficient of a representation $f = (\tau \otimes id)w$ we have

$$\int_{\varphi} f = \tau(P)$$

where P is the orthogonal projection onto the 1-eigenspace of $(id \otimes \psi)w$.

PROOF. By linearity it is enough to prove the first assertion for functions of the following type, where w is a Peter-Weyl representation, and τ is a linear form:

$$f = (\tau \otimes id)w$$

Thus we are led into the second assertion, and more precisely we can have the whole result proved if we can establish the following formula, with $f = (\tau \otimes id)w$:

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n \psi^{*k}(f) = \tau(P)$$

In order to prove this latter formula, observe that we have:

$$\psi^{*k}(f) = (\tau \otimes \psi^{*k})w = \tau((id \otimes \psi^{*k})w)$$

Let us set $M = (id \otimes \psi)w$. In terms of this matrix, we have:

$$((id \otimes \psi^{*k})w)_{i_0 i_{k+1}} = \sum_{i_1 \dots i_k} M_{i_0 i_1} \dots M_{i_k i_{k+1}} = (M^k)_{i_0 i_{k+1}}$$

Thus we have the following formula, valid for any $k \in \mathbb{N}$:

$$(id \otimes \psi^{*k})w = M^k$$

It follows that our Cesàro limit is given by the following formula:

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n \psi^{*k}(f) = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n \tau(M^k) = \tau \left(\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n M^k \right)$$

Now since w is unitary we have $\|w\| = 1$, and so $\|M\| \leq 1$. Thus the last Cesàro limit converges, and equals the orthogonal projection onto the 1-eigenspace of M :

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n M^k = P$$

Thus our initial Cesàro limit converges as well, to $\tau(P)$, as desired. \square

When the linear form $\psi \in C(G)^*$ is faithful, we have the following finer result:

PROPOSITION 14.14. *Given a faithful unital linear form $\psi \in C(G)^*$, the limit*

$$\int_{\psi} f = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n \psi^{*k}(f)$$

exists, and is independent of ψ , given on coefficients of representations by

$$\left(id \otimes \int_{\psi} \right) w = P$$

where P is the orthogonal projection onto the space $Fix(w) = \{\xi \in \mathbb{C}^n \mid w\xi = \xi\}$.

PROOF. In view of Proposition 14.13, it remains to prove that when ψ is faithful, the 1-eigenspace of the matrix $M = (id \otimes \psi)w$ equals the space $Fix(w)$.

“ \supset ” This is clear, and for any ψ , because we have the following implication:

$$w\xi = \xi \implies M\xi = \xi$$

“ \subset ” Here we must prove that, when ψ is faithful, we have:

$$M\xi = \xi \implies w\xi = \xi$$

For this purpose, assume that we have $M\xi = \xi$, and consider the following function:

$$f = \sum_i \left(\sum_j w_{ij} \xi_j - \xi_i \right) \left(\sum_k w_{ik} \xi_k - \xi_i \right)^*$$

We must prove that we have $f = 0$. Since v is unitary, we have:

$$\begin{aligned} f &= \sum_{ijk} w_{ij} w_{ik}^* \xi_j \bar{\xi}_k - \frac{1}{N} w_{ij} \xi_j \bar{\xi}_i - \frac{1}{N} w_{ik}^* \xi_i \bar{\xi}_k + \frac{1}{N^2} \xi_i \bar{\xi}_i \\ &= \sum_j |\xi_j|^2 - \sum_{ij} w_{ij} \xi_j \bar{\xi}_i - \sum_{ik} w_{ik}^* \xi_i \bar{\xi}_k + \sum_i |\xi_i|^2 \\ &= ||\xi||^2 - \langle w\xi, \xi \rangle - \overline{\langle w\xi, \xi \rangle} + ||\xi||^2 \\ &= 2(||\xi||^2 - \operatorname{Re}(\langle w\xi, \xi \rangle)) \end{aligned}$$

By using now our assumption $M\xi = \xi$, we obtain from this:

$$\begin{aligned} \psi(f) &= 2\psi(||\xi||^2 - \operatorname{Re}(\langle w\xi, \xi \rangle)) \\ &= 2(||\xi||^2 - \operatorname{Re}(\langle M\xi, \xi \rangle)) \\ &= 2(||\xi||^2 - ||\xi||^2) \\ &= 0 \end{aligned}$$

Now since ψ is faithful, this gives $f = 0$, and so $w\xi = \xi$, as claimed. \square

We can now formulate a main result, as follows:

THEOREM 14.15. *Any compact group G has a unique Haar integration, which can be constructed by starting with any faithful positive unital form $\psi \in C(G)^*$, and setting:*

$$\int_G = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n \psi^{*k}$$

Moreover, for any representation w we have the formula

$$\left(id \otimes \int_G \right) w = P$$

where P is the orthogonal projection onto $Fix(w) = \{\xi \in \mathbb{C}^n \mid w\xi = \xi\}$.

PROOF. Let us first go back to the general context of Proposition 14.13. Since convolving one more time with ψ will not change the Cesàro limit appearing there, the functional $\int_\psi \in C(G)^*$ constructed there has the following invariance property:

$$\int_\psi * \psi = \psi * \int_\psi = \int_\psi$$

In the case where ψ is assumed to be faithful, as in Proposition 14.14, our claim is that we have the following formula, valid this time for any $\varphi \in C(G)^*$:

$$\int_\psi * \varphi = \varphi * \int_\psi = \varphi(1) \int_\psi$$

Indeed, it is enough to prove this formula on a coefficient of a corepresentation:

$$f = (\tau \otimes id)w$$

In order to do so, consider the following two matrices:

$$P = \left(id \otimes \int_\psi \right) w \quad , \quad Q = (id \otimes \varphi)w$$

We have then the following formulae, which all follow from definitions:

$$\left(\int_\psi * \varphi \right) f = \tau(PQ) \quad , \quad \left(\varphi * \int_\psi \right) f = \tau(QP) \quad , \quad \varphi(1) \int_\psi f = \varphi(1) \tau(P)$$

Thus, in order to prove our claim, it is enough to establish the following formula:

$$PQ = QP = \psi(1)P$$

But this follows from the fact, from Proposition 14.14, that $P = (id \otimes \int_\psi)w$ is the orthogonal projection onto $Fix(w)$. Thus, we proved our claim. Now observe that, with $\Delta f(g \otimes h) = f(gh)$, this formula that we proved can be written as follows:

$$\varphi \left(\int_\psi \otimes id \right) \Delta = \varphi \left(id \otimes \int_\psi \right) \Delta = \varphi \int_\psi (.)1$$

This formula being true for any $\varphi \in C(G)^*$, we can simply delete φ , and we conclude that $\int_G = \int_\psi$ has the required left and right invariance property, namely:

$$\left(\int_G \otimes id \right) \Delta = \left(id \otimes \int_G \right) \Delta = \int_G (\cdot) 1$$

Finally, the uniqueness is clear as well, because if we have two invariant integrals \int_G, \int'_G , then their convolution equals on one hand \int_G , and on the other hand, \int'_G . \square

Summarizing, we know how to integrate over G . Before getting into probabilistic applications, let us develop however more Peter-Weyl theory. We will need:

PROPOSITION 14.16. *We have a Frobenius type isomorphism*

$$Hom(v, w) \simeq Fix(v \otimes \bar{w})$$

valid for any two representations v, w .

PROOF. According to definitions, we have the following equivalences:

$$\begin{aligned} T \in Hom(v, w) &\iff Tv = wT \\ &\iff \sum_i T_{ai} v_{ij} = \sum_b w_{ab} T_{bj}, \forall a, j \end{aligned}$$

On the other hand, we have as well the following equivalences:

$$\begin{aligned} T \in Fix(v \otimes \bar{w}) &\iff (v \otimes \bar{w})T = \xi \\ &\iff \sum_{bi} v_{ji} \bar{w}_{ab} T_{bi} = T_{aj} \forall a, j \end{aligned}$$

With these formulae in hand, both inclusions follow from the unitarity of v, w . \square

We can now formulate a third Peter-Weyl theorem, as follows:

THEOREM 14.17 (Peter-Weyl 3). *The dense subalgebra $\mathcal{C}(G) \subset C(G)$ generated by the coefficients of the fundamental representation decomposes as a direct sum*

$$\mathcal{C}(G) = \bigoplus_{w \in Irr(G)} M_{\dim(w)}(\mathbb{C})$$

with the summands being pairwise orthogonal with respect to the scalar product

$$\langle f, g \rangle = \int_G f \bar{g}$$

where \int_G is the Haar integration over G .

PROOF. By combining the previous two Peter-Weyl results, Theorems 14.10 and 14.11, we deduce that we have a linear space decomposition as follows:

$$\mathcal{C}(G) = \sum_{w \in Irr(G)} C_w = \sum_{w \in Irr(G)} M_{\dim(w)}(\mathbb{C})$$

Thus, in order to conclude, it is enough to prove that for any two irreducible representations $v, w \in \text{Irr}(G)$, the corresponding spaces of coefficients are orthogonal:

$$v \not\sim w \implies C_v \perp C_w$$

But this follows from Theorem 14.15, via Proposition 14.16. Let us set indeed:

$$P_{ia,jb} = \int_G v_{ij} \bar{w}_{ab}$$

Then P is the orthogonal projection onto the following vector space:

$$\text{Fix}(v \otimes \bar{w}) \simeq \text{Hom}(v, w) = \{0\}$$

Thus we have $P = 0$, and this gives the result. \square

Finally, we have the following result, completing the Peter-Weyl theory:

THEOREM 14.18 (Peter-Weyl 4). *The characters of irreducible representations belong to the algebra*

$$\mathcal{C}(G)_{\text{central}} = \left\{ f \in \mathcal{C}(G) \mid f(gh) = f(hg), \forall g, h \in G \right\}$$

called algebra of central functions on G , and form an orthonormal basis of it.

PROOF. Observe first that $\mathcal{C}(G)_{\text{central}}$ is indeed an algebra, which contains all the characters. Conversely, consider a function $f \in \mathcal{C}(G)$, written as follows:

$$f = \sum_{w \in \text{Irr}(G)} f_w$$

The condition $f \in \mathcal{C}(G)_{\text{central}}$ states then that for any $w \in \text{Irr}(G)$, we must have:

$$f_w \in \mathcal{C}(G)_{\text{central}}$$

But this means that f_w must be a scalar multiple of χ_w , so the characters form a basis of $\mathcal{C}(G)_{\text{central}}$, as stated. Also, the fact that we have an orthogonal basis follows from Theorem 14.17. As for the fact that the characters have norm 1, this follows from:

$$\int_G \chi_w \bar{\chi}_w = \sum_{ij} \int_G w_{ii} \bar{w}_{jj} = \sum_i \frac{1}{M} = 1$$

Here we have used the fact, coming from Theorem 14.15 and Proposition 14.16, that the integrals $\int_G w_{ij} \bar{w}_{kl}$ form the orthogonal projection onto the following vector space:

$$\text{Fix}(w \otimes \bar{w}) \simeq \text{End}(w) = \mathbb{C}1$$

Thus, the proof of our theorem is now complete. \square

14c. Diagrams, easiness

In view of the above results, no matter on what we want to do with our group, we must compute the spaces $Fix(v^{\otimes k})$. It is technically convenient to slightly enlarge the class of spaces to be computed, by talking about Tannakian categories, as follows:

DEFINITION 14.19. *The Tannakian category associated to a closed subgroup $G \subset_v U_N$ is the collection $C_G = (C_G(k, l))$ of vector spaces*

$$C_G(k, l) = Hom(v^{\otimes k}, v^{\otimes l})$$

where the representations $v^{\otimes k}$ with $k = \circ \bullet \bullet \circ \dots$ colored integer, defined by

$$v^{\otimes \emptyset} = 1 \quad , \quad v^{\otimes \circ} = v \quad , \quad v^{\otimes \bullet} = \bar{v}$$

and multiplicativity, $v^{\otimes kl} = v^{\otimes k} \otimes v^{\otimes l}$, are the Peter-Weyl representations.

Let us make a summary of what we have so far, regarding these spaces $C_G(k, l)$. In order to formulate our result, let us start with the following definition:

DEFINITION 14.20. *Let H be a finite dimensional Hilbert space. A tensor category over H is a collection $C = (C(k, l))$ of linear spaces*

$$C(k, l) \subset \mathcal{L}(H^{\otimes k}, H^{\otimes l})$$

satisfying the following conditions:

- (1) $S, T \in C$ implies $S \otimes T \in C$.
- (2) If $S, T \in C$ are composable, then $ST \in C$.
- (3) $T \in C$ implies $T^* \in C$.
- (4) $C(k, k)$ contains the identity operator.
- (5) $C(\emptyset, k)$ with $k = \circ \bullet, \bullet \circ$ contain the operator $R : 1 \rightarrow \sum_i e_i \otimes e_i$.
- (6) $C(kl, lk)$ with $k, l = \circ, \bullet$ contain the flip operator $\Sigma : a \otimes b \rightarrow b \otimes a$.

Here the tensor power Hilbert spaces $H^{\otimes k}$, with $k = \circ \bullet \bullet \circ \dots$ being a colored integer, are defined by the following formulae, and multiplicativity:

$$H^{\otimes \emptyset} = \mathbb{C} \quad , \quad H^{\otimes \circ} = H \quad , \quad H^{\otimes \bullet} = \bar{H} \simeq H$$

With these conventions, we have the following result, summarizing our knowledge on the subject, coming from the results established in the above:

THEOREM 14.21. *For a closed subgroup $G \subset_v U_N$, the associated Tannakian category*

$$C_G(k, l) = Hom(v^{\otimes k}, v^{\otimes l})$$

is a tensor category over the Hilbert space $H = \mathbb{C}^N$.

PROOF. We know that the fundamental representation v acts on the Hilbert space $H = \mathbb{C}^N$, and that its conjugate \bar{v} acts on the Hilbert space $\bar{H} = \mathbb{C}^N$. Now by multiplicativity we conclude that any Peter-Weyl representation $v^{\otimes k}$ acts on the Hilbert space $H^{\otimes k}$, and so that we have embeddings as in Definition 14.20, as follows:

$$C_G(k, l) \subset \mathcal{L}(H^{\otimes k}, H^{\otimes l})$$

Regarding now the fact that the axioms (1-6) in Definition 14.20 are indeed satisfied, this is something that we basically already know. To be more precise, (1-4) are clear, and (5) follows from the fact that each element $g \in G$ is a unitary, which gives:

$$R \in \text{Hom}(1, g \otimes \bar{g}) \quad , \quad R \in \text{Hom}(1, \bar{g} \otimes g)$$

As for (6), this is something trivial, coming from the fact that the matrix coefficients $g \rightarrow g_{ij}$ and their complex conjugates $g \rightarrow \bar{g}_{ij}$ commute with each other. \square

Our purpose now will be that of showing that any closed subgroup $G \subset U_N$ is uniquely determined by its Tannakian category $C_G = (C_G(k, l))$. This result, known as Tannakian duality, is something quite deep, and extremely useful. Indeed, the idea is that what we would have here is a “linearization” of G , allowing us to do combinatorics, and to ultimately reach to concrete and powerful results, regarding G itself. We first have:

THEOREM 14.22. *Given a tensor category $C = (C(k, l))$ over a finite dimensional Hilbert space $H \simeq \mathbb{C}^N$, the following construction,*

$$G_C = \left\{ g \in U_N \mid Tg^{\otimes k} = g^{\otimes l}T, \forall k, l, \forall T \in C(k, l) \right\}$$

produces a closed subgroup $G_C \subset U_N$.

PROOF. This is something elementary, with the fact that the closed subset $G_C \subset U_N$ constructed in the statement is indeed stable under the multiplication, unit and inversion operation for the unitary matrices $g \in U_N$ being clear from definitions. \square

We can now formulate the Tannakian duality result, as follows:

THEOREM 14.23. *The above Tannakian constructions*

$$G \rightarrow C_G \quad , \quad C \rightarrow G_C$$

are bijective, and inverse to each other.

PROOF. This is something quite technical, obtained by doing some abstract algebra, and for details here, we refer to the Tannakian duality literature. The whole subject is actually, in modern times, for the most part of quantum algebra, and you can consult here various quantum group papers and books, such as [9], for details on the above. \square

In order to reach now to more concrete things, following Brauer, we have:

DEFINITION 14.24. Let $P(k, l)$ be the set of partitions between an upper colored integer k , and a lower colored integer l . A collection of subsets

$$D = \bigsqcup_{k, l} D(k, l)$$

with $D(k, l) \subset P(k, l)$ is called a category of partitions when it has the following properties:

- (1) Stability under the horizontal concatenation, $(\pi, \sigma) \rightarrow [\pi\sigma]$.
- (2) Stability under vertical concatenation $(\pi, \sigma) \rightarrow \begin{bmatrix} \sigma \\ \pi \end{bmatrix}$, with matching middle symbols.
- (3) Stability under the upside-down turning $*$, with switching of colors, $\circ \leftrightarrow \bullet$.
- (4) Each set $P(k, k)$ contains the identity partition $|| \dots ||$.
- (5) The sets $P(\emptyset, \circ\bullet)$ and $P(\emptyset, \bullet\circ)$ both contain the semicircle \cap .
- (6) The sets $P(k, \bar{k})$ with $|k| = 2$ contain the crossing partition \bowtie .

There are many examples of such categories, as for instance the category of all pairings P_2 , or of all matching pairings \mathcal{P}_2 . We will be back to examples in a moment.

Let us formulate as well the following definition:

DEFINITION 14.25. Given a partition $\pi \in P(k, l)$ and an integer $N \in \mathbb{N}$, we can construct a linear map between tensor powers of \mathbb{C}^N ,

$$T_\pi : (\mathbb{C}^N)^{\otimes k} \rightarrow (\mathbb{C}^N)^{\otimes l}$$

by the following formula, with e_1, \dots, e_N being the standard basis of \mathbb{C}^N ,

$$T_\pi(e_{i_1} \otimes \dots \otimes e_{i_k}) = \sum_{j_1 \dots j_l} \delta_\pi \begin{pmatrix} i_1 & \dots & i_k \\ j_1 & \dots & j_l \end{pmatrix} e_{j_1} \otimes \dots \otimes e_{j_l}$$

and with the coefficients on the right being Kronecker type symbols,

$$\delta_\pi \begin{pmatrix} i_1 & \dots & i_k \\ j_1 & \dots & j_l \end{pmatrix} \in \{0, 1\}$$

whose values depend on whether the indices fit or not.

To be more precise, we put the indices of i, j on the legs of π , in the obvious way. In case all the blocks of π contain equal indices of i, j , we set $\delta_\pi \begin{pmatrix} i \\ j \end{pmatrix} = 1$. Otherwise, we set $\delta_\pi \begin{pmatrix} i \\ j \end{pmatrix} = 0$. The relation with the Tannakian categories comes from:

PROPOSITION 14.26. The assignment $\pi \rightarrow T_\pi$ is categorical, in the sense that

$$T_\pi \otimes T_\nu = T_{[\pi\nu]} \quad , \quad T_\pi T_\nu = N^{c(\pi, \nu)} T_{[\pi]_\nu}^* \quad , \quad T_\pi^* = T_{\pi^*}$$

where $c(\pi, \nu)$ are certain integers, coming from the erased components in the middle.

PROOF. This is something elementary, the computations being as follows:

(1) The concatenation axiom can be checked as follows:

$$\begin{aligned}
& (T_\pi \otimes T_\nu)(e_{i_1} \otimes \dots \otimes e_{i_p} \otimes e_{k_1} \otimes \dots \otimes e_{k_r}) \\
&= \sum_{j_1 \dots j_q} \sum_{l_1 \dots l_s} \delta_\pi \begin{pmatrix} i_1 & \dots & i_p \\ j_1 & \dots & j_q \end{pmatrix} \delta_\nu \begin{pmatrix} k_1 & \dots & k_r \\ l_1 & \dots & l_s \end{pmatrix} e_{j_1} \otimes \dots \otimes e_{j_q} \otimes e_{l_1} \otimes \dots \otimes e_{l_s} \\
&= \sum_{j_1 \dots j_q} \sum_{l_1 \dots l_s} \delta_{[\pi\nu]} \begin{pmatrix} i_1 & \dots & i_p & k_1 & \dots & k_r \\ j_1 & \dots & j_q & l_1 & \dots & l_s \end{pmatrix} e_{j_1} \otimes \dots \otimes e_{j_q} \otimes e_{l_1} \otimes \dots \otimes e_{l_s} \\
&= T_{[\pi\nu]}(e_{i_1} \otimes \dots \otimes e_{i_p} \otimes e_{k_1} \otimes \dots \otimes e_{k_r})
\end{aligned}$$

(2) The composition axiom can be checked as follows:

$$\begin{aligned}
& T_\pi T_\nu(e_{i_1} \otimes \dots \otimes e_{i_p}) \\
&= \sum_{j_1 \dots j_q} \delta_\nu \begin{pmatrix} i_1 & \dots & i_p \\ j_1 & \dots & j_q \end{pmatrix} \sum_{k_1 \dots k_r} \delta_\pi \begin{pmatrix} j_1 & \dots & j_q \\ k_1 & \dots & k_r \end{pmatrix} e_{k_1} \otimes \dots \otimes e_{k_r} \\
&= \sum_{k_1 \dots k_r} N^{c(\pi, \nu)} \delta_{[\pi]}^{\nu} \begin{pmatrix} i_1 & \dots & i_p \\ k_1 & \dots & k_r \end{pmatrix} e_{k_1} \otimes \dots \otimes e_{k_r} \\
&= N^{c(\pi, \nu)} T_{[\pi]}^{\nu}(e_{i_1} \otimes \dots \otimes e_{i_p})
\end{aligned}$$

(3) Finally, the involution axiom can be checked as follows:

$$\begin{aligned}
& T_\pi^*(e_{j_1} \otimes \dots \otimes e_{j_q}) \\
&= \sum_{i_1 \dots i_p} \langle T_\pi^*(e_{j_1} \otimes \dots \otimes e_{j_q}), e_{i_1} \otimes \dots \otimes e_{i_p} \rangle e_{i_1} \otimes \dots \otimes e_{i_p} \\
&= \sum_{i_1 \dots i_p} \delta_\pi \begin{pmatrix} i_1 & \dots & i_p \\ j_1 & \dots & j_q \end{pmatrix} e_{i_1} \otimes \dots \otimes e_{i_p} \\
&= T_{\pi^*}(e_{j_1} \otimes \dots \otimes e_{j_q})
\end{aligned}$$

Summarizing, our correspondence is indeed categorical. \square

In relation now with the groups, we have the following result:

THEOREM 14.27. *Each category of partitions $D = (D(k, l))$ produces a family of compact groups $G = (G_N)$, with $G_N \subset_v U_N$, via the formula*

$$Hom(v^{\otimes k}, v^{\otimes l}) = span \left(T_\pi \middle| \pi \in D(k, l) \right)$$

and the Tannakian duality correspondence.

PROOF. Given an integer $N \in \mathbb{N}$, consider the correspondence $\pi \rightarrow T_\pi$ constructed in Definition 14.25, and then the collection of linear spaces in the statement, namely:

$$C(k, l) = \text{span} \left(T_\pi \Big| \pi \in D(k, l) \right)$$

According to Proposition 14.26, and to our axioms for the categories of partitions, from Definition 14.24, this collection of spaces $C = (C(k, l))$ satisfies the axioms for the Tannakian categories, from Definition 14.20. Thus the Tannakian duality result, Theorem 14.23, applies, and provides us with a closed subgroup $G_N \subset_v U_N$ such that:

$$C(k, l) = \text{Hom}(v^{\otimes k}, v^{\otimes l})$$

Thus, we are led to the conclusion in the statement. \square

We can now formulate a key definition, as follows:

DEFINITION 14.28. *A closed subgroup $G \subset_v U_N$ is called easy when we have*

$$\text{Hom}(v^{\otimes k}, v^{\otimes l}) = \text{span} \left(T_\pi \Big| \pi \in D(k, l) \right)$$

for any colored integers k, l , for a certain category of partitions $D \subset P$.

The notion of easiness goes back to the results of Brauer regarding the orthogonal group O_N , and the unitary group U_N , which can be formulated as follows:

THEOREM 14.29. *We have the following results:*

- (1) U_N is easy, coming from the category of matching pairings \mathcal{P}_2 .
- (2) O_N is easy too, coming from the category of all pairings P_2 .

PROOF. This is something very standard, the idea being as follows:

(1) The group U_N being defined via the relations $v^* = v^{-1}$, $v^t = \bar{v}^{-1}$, the associated Tannakian category is $C = \text{span}(T_\pi | \pi \in D)$, with:

$$D = \langle \begin{array}{c} \cap \\ \bullet \bullet \end{array}, \begin{array}{c} \cap \\ \bullet \circ \end{array} \rangle = \mathcal{P}_2$$

(2) The group $O_N \subset U_N$ being defined by imposing the relations $v_{ij} = \bar{v}_{ij}$, the associated Tannakian category is $C = \text{span}(T_\pi | \pi \in D)$, with:

$$D = \langle \mathcal{P}_2, \begin{array}{c} \updownarrow \\ \bullet \bullet \end{array}, \begin{array}{c} \updownarrow \\ \bullet \circ \end{array} \rangle = P_2$$

Thus, we are led to the conclusion in the statement. \square

Beyond this, a first natural question is that of computing the easy group associated to the category P itself, and we have here the following Brauer type theorem:

THEOREM 14.30. *The symmetric group S_N , regarded as group of unitary matrices,*

$$S_N \subset O_N \subset U_N$$

via the permutation matrices, is easy, coming from the category of all partitions P .

PROOF. Consider the easy group $G \subset O_N$ coming from the category of all partitions P . Since P is generated by the one-block partition $Y \in P(2, 1)$, we have:

$$C(G) = C(O_N) / \left\langle T_Y \in \text{Hom}(v^{\otimes 2}, v) \right\rangle$$

The linear map associated to Y is given by the following formula:

$$T_Y(e_i \otimes e_j) = \delta_{ij} e_i$$

Thus, the relation defining the above group $G \subset O_N$ reformulates as follows:

$$T_Y \in \text{Hom}(v^{\otimes 2}, v) \iff v_{ij} v_{ik} = \delta_{jk} v_{ij}, \forall i, j, k$$

In other words, the elements v_{ij} must be projections, and these projections must be pairwise orthogonal on the rows of $v = (v_{ij})$. We conclude that $G \subset O_N$ is the subgroup of matrices $g \in O_N$ having the property $g_{ij} \in \{0, 1\}$. Thus we have $G = S_N$, as claimed. \square

In fact, we have the following general easiness result, regarding the series of complex reflection groups $H_N^s \subset U_N$, that we introduced in chapter 3:

THEOREM 14.31. *The group $H_N^s = \mathbb{Z}_s \wr S_N$ is easy, the corresponding category P^s consisting of the partitions satisfying $\# \circ = \# \bullet (s)$ in each block. In particular:*

- (1) S_N is easy, coming from the category P .
- (2) H_N is easy, coming from the category P_{even} .
- (3) K_N is easy, coming from the category $\mathcal{P}_{\text{even}}$.

PROOF. This is something that we already know at $s = 1$, from Theorem 14.30. In general, the proof is similar, based on Tannakian duality. To be more precise, in what regards the main assertion, the idea here is that the one-block partition $\pi \in P(s)$, which generates the category P^s , implements the relations producing the subgroup $H_N^s \subset U_N$. As for the last assertions, these follow from the following observations:

(1) At $s = 1$ we know that we have $H_N^1 = S_N$. Regarding now the corresponding category, here the condition $\# \circ = \# \bullet (1)$ is automatic, and so $P^1 = P$.

(2) At $s = 2$ we know that we have $H_N^2 = H_N$. Regarding now the corresponding category, here the condition $\# \circ = \# \bullet (2)$ reformulates as follows:

$$\# \circ + \# \bullet = 0(2)$$

Thus each block must have even size, and we obtain, as claimed, $P^2 = P_{\text{even}}$.

(3) At $s = \infty$ we know that we have $H_N^\infty = K_N$. Regarding now the corresponding category, here the condition $\# \circ = \# \bullet (\infty)$ reads:

$$\# \circ = \# \bullet$$

But this is the condition defining $\mathcal{P}_{\text{even}}$, and so $P^\infty = \mathcal{P}_{\text{even}}$, as claimed. \square

Let us go back now to probability questions, with the aim of applying the above abstract theory, to questions regarding characters. The situation here is as follows:

(1) Given a closed subgroup $G \subset_v U_N$, we know from Peter-Weyl that the moments of the main character count the fixed points of the representations $v^{\otimes k}$.

(2) On the other hand, assuming that our group $G \subset_v U_N$ is easy, coming from a category of partitions $D = (D(k, l))$, the space formed by these fixed points is spanned by the following vectors, indexed by partitions π belonging to the set $D(k) = D(0, k)$:

$$\xi_\pi = \sum_{i_1 \dots i_k} \delta_\pi(i_1 \dots i_k) e_{i_1} \otimes \dots \otimes e_{i_k}$$

(3) Thus, we are left with investigating linear independence questions for the vectors ξ_π , and once these questions solved, to compute the moments of χ .

In order to investigate linear independence questions for the vectors ξ_π , we will use the Gram matrix of these vectors. Let us begin with some standard definitions:

DEFINITION 14.32. *Let $P(k)$ be the set of partitions of $\{1, \dots, k\}$, and let $\pi, \nu \in P(k)$.*

- (1) *We write $\pi \leq \nu$ if each block of π is contained in a block of ν .*
- (2) *We let $\pi \vee \nu \in P(k)$ be the partition obtained by superposing π, ν .*

As an illustration here, at $k = 2$ we have $P(2) = \{||, \sqcup\}$, and the order is:

$$|| \leq \sqcup$$

At $k = 3$ we have $P(3) = \{|||, \sqcup|, \sqcap|, |\sqcup, \sqcap\sqcup\}$, and the order relation is as follows:

$$||| \leq \sqcup|, \sqcap|, |\sqcup \leq \sqcap\sqcup$$

Observe also that we have $\pi, \nu \leq \pi \vee \nu$. In fact, $\pi \vee \nu$ is the smallest partition with this property, called supremum of π, ν . Now back to the easy groups, we have:

PROPOSITION 14.33. *The Gram matrix $G_{kN}(\pi, \nu) = \langle \xi_\pi, \xi_\nu \rangle$ is given by*

$$G_{kN}(\pi, \nu) = N^{|\pi \vee \nu|}$$

where $|\cdot|$ is the number of blocks.

PROOF. According to our formula of the vectors ξ_π , we have:

$$\begin{aligned} \langle \xi_\pi, \xi_\nu \rangle &= \sum_{i_1 \dots i_k} \delta_\pi(i_1, \dots, i_k) \delta_\nu(i_1, \dots, i_k) \\ &= \sum_{i_1 \dots i_k} \delta_{\pi \vee \nu}(i_1, \dots, i_k) \\ &= N^{|\pi \vee \nu|} \end{aligned}$$

Thus, we have obtained the formula in the statement. □

In order to study the Gram matrix, and more specifically to compute its determinant, we will need several standard facts about the partitions. We first have:

DEFINITION 14.34. *The Möbius function of any lattice, and so of P , is given by*

$$\mu(\pi, \nu) = \begin{cases} 1 & \text{if } \pi = \nu \\ -\sum_{\pi \leq \tau < \nu} \mu(\pi, \tau) & \text{if } \pi < \nu \\ 0 & \text{if } \pi \not\leq \nu \end{cases}$$

with the construction being performed by recurrence.

As an illustration here, let us go back to the set of 2-point partitions, $P(2) = \{||, \sqcap\}$. Here we have by definition:

$$\mu(||, ||) = \mu(\sqcap, \sqcap) = 1$$

Also, we know that we have $|| < \sqcap$, with no intermediate partition in between, and so the above recurrence procedure gives the following formular:

$$\mu(||, \sqcap) = -\mu(||, ||) = -1$$

Finally, we have $\sqcap \not\leq ||$, which gives $\mu(\sqcap, ||) = 0$. Thus, as a conclusion, the Möbius matrix $M_{\pi\nu} = \mu(\pi, \nu)$ of the lattice $P(2) = \{||, \sqcap\}$ is as follows:

$$M = \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix}$$

The interest in the Möbius function comes from the Möbius inversion formula:

$$f(\nu) = \sum_{\pi \leq \nu} g(\pi) \implies g(\nu) = \sum_{\pi \leq \nu} \mu(\pi, \nu) f(\pi)$$

In linear algebra terms, the statement and proof of this formula are as follows:

THEOREM 14.35. *The inverse of the adjacency matrix of P , given by*

$$A_{\pi\nu} = \begin{cases} 1 & \text{if } \pi \leq \nu \\ 0 & \text{if } \pi \not\leq \nu \end{cases}$$

is the Möbius matrix of P , given by $M_{\pi\nu} = \mu(\pi, \nu)$.

PROOF. This is well-known, coming for instance from the fact that A is upper triangular. Thus, when inverting, we are led into the recurrence from Definition 14.34. \square

As an illustration here, for $P(2)$ the formula $M = A^{-1}$ appears as follows:

$$\begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}^{-1}$$

Now back to our Gram matrix considerations, we have the following result:

PROPOSITION 14.36. *The Gram matrix is given by $G_{kN} = AL$, where*

$$L(\pi, \nu) = \begin{cases} N(N-1) \dots (N - |\pi| + 1) & \text{if } \nu \leq \pi \\ 0 & \text{otherwise} \end{cases}$$

and where $A = M^{-1}$ is the adjacency matrix of $P(k)$.

PROOF. We have the following computation:

$$\begin{aligned} N^{|\pi \vee \nu|} &= \# \left\{ i_1, \dots, i_k \in \{1, \dots, N\} \mid \ker i \geq \pi \vee \nu \right\} \\ &= \sum_{\tau \geq \pi \vee \nu} \# \left\{ i_1, \dots, i_k \in \{1, \dots, N\} \mid \ker i = \tau \right\} \\ &= \sum_{\tau \geq \pi \vee \nu} N(N-1) \dots (N - |\tau| + 1) \end{aligned}$$

According to Proposition 14.33 and to the definition of A, L , this formula reads:

$$(G_{kN})_{\pi\nu} = \sum_{\tau \geq \pi} L_{\tau\nu} = \sum_{\tau} A_{\pi\tau} L_{\tau\nu} = (AL)_{\pi\nu}$$

Thus, we obtain the formula in the statement. \square

With the above result in hand, we can now investigate the linear independence properties of the vectors ξ_π . To be more precise, we have the following result:

THEOREM 14.37. *The determinant of the Gram matrix G_{kN} is given by*

$$\det(G_{kN}) = \prod_{\pi \in P(k)} \frac{N!}{(N - |\pi|)!}$$

and in particular, for $N \geq k$, the vectors $\{\xi_\pi \mid \pi \in P(k)\}$ are linearly independent.

PROOF. According to the formula in Proposition 14.36, we have:

$$\det(G_{kN}) = \det(A) \det(L)$$

Now if we order $P(k)$ as usual, with respect to the number of blocks, and then lexicographically, we see that A is upper triangular, and that L is lower triangular. Thus $\det(A)$ can be computed simply by making the product on the diagonal, and we obtain 1. As for $\det(L)$, this can be computed as well by making the product on the diagonal, and we obtain the number in the statement, with the technical remark that in the case $N < k$ the convention is that we obtain a vanishing determinant. \square

Now back to the laws of characters, we can formulate:

PROPOSITION 14.38. *For an easy group $G = (G_N)$, coming from a category of partitions $D = (D(k, l))$, the asymptotic moments of the main character are given by*

$$\lim_{N \rightarrow \infty} \int_{G_N} \chi^k = \#D(k)$$

where $D(k) = D(\emptyset, k)$, with the limiting sequence on the left consisting of certain integers, and being stationary at least starting from the k -th term.

PROOF. This follows indeed from the Peter-Weyl theory, by using the linear independence result for the vectors ξ_π coming from Theorem 14.37. \square

With these preliminaries in hand, we can now state and prove:

THEOREM 14.39. *In the $N \rightarrow \infty$ limit, the laws of the main character for the main easy groups, real and complex, and discrete and continuous, are as follows,*

$$\begin{array}{ccc} K_N & \longrightarrow & U_N \\ \uparrow & & \uparrow \\ H_N & \longrightarrow & O_N \end{array} \quad : \quad \begin{array}{ccc} B_1 & \longrightarrow & G_1 \\ \uparrow & & \uparrow \\ b_1 & \longrightarrow & g_1 \end{array}$$

with these laws, namely the real and complex Gaussian and Bessel laws, being the main limiting laws in real and complex, and discrete and continuous probability.

PROOF. This follows from the above results. To be more precise, we know that the above groups are all easy, the corresponding categories of partitions being as follows:

$$\begin{array}{ccc} \mathcal{P}_{\text{even}} & \longleftarrow & \mathcal{P}_2 \\ \downarrow & & \downarrow \\ P_{\text{even}} & \longleftarrow & P_2 \end{array}$$

Thus, we can use Proposition 14.38, as we are led into counting partitions, and then recovering the measures via their moments, and this leads to the result. \square

14d. Weingarten formula

Our aim now is to go beyond what we have, with results regarding the truncated characters. Let us start with a general formula coming from Peter-Weyl, namely:

THEOREM 14.40. *The Haar integration over a closed subgroup $G \subset_v U_N$ is given on the dense subalgebra of smooth functions by the Weingarten type formula*

$$\int_G g_{i_1 j_1}^{e_1} \cdots g_{i_k j_k}^{e_k} dg = \sum_{\pi, \nu \in D(k)} \delta_\pi(i) \delta_\nu(j) W_k(\pi, \nu)$$

valid for any colored integer $k = e_1 \dots e_k$ and any multi-indices i, j , where $D(k)$ is a linear basis of $\text{Fix}(v^{\otimes k})$, the associated generalized Kronecker symbols are given by

$$\delta_\pi(i) = \langle \pi, e_{i_1} \otimes \dots \otimes e_{i_k} \rangle$$

and $W_k = G_k^{-1}$ is the inverse of the Gram matrix, $G_k(\pi, \nu) = \langle \pi, \nu \rangle$.

PROOF. This is something very standard, coming from the fact that the above integrals form altogether the orthogonal projection P^k onto the following space:

$$\text{Fix}(v^{\otimes k}) = \text{span}(D(k))$$

Consider now the following linear map, with $D(k) = \{\xi_k\}$ being as in the statement:

$$E(x) = \sum_{\pi \in D(k)} \langle x, \xi_\pi \rangle \xi_\pi$$

By a standard linear algebra computation, it follows that we have $P = WE$, where W is the inverse of the restriction of E to the following space:

$$K = \text{span} \left(T_\pi \Big|_{\pi \in D(k)} \right)$$

But this restriction is the linear map given by the matrix G_k , and so W is the linear map given by the inverse matrix $W_k = G_k^{-1}$, and this gives the result. \square

In the easy case, we have the following more concrete result:

THEOREM 14.41. *For an easy group $G \subset U_N$, coming from a category of partitions $D = (D(k, l))$, we have the Weingarten formula*

$$\int_G g_{i_1 j_1}^{e_1} \cdots g_{i_k j_k}^{e_k} dg = \sum_{\pi, \nu \in D(k)} \delta_\pi(i) \delta_\nu(j) W_{kN}(\pi, \nu)$$

for any $k = e_1 \dots e_k$ and any i, j , where $D(k) = D(\emptyset, k)$, δ are usual Kronecker type symbols, checking whether the indices match, and $W_{kN} = G_{kN}^{-1}$, with

$$G_{kN}(\pi, \nu) = N^{|\pi \vee \nu|}$$

where $|\cdot|$ is the number of blocks.

PROOF. We use the abstract Weingarten formula, from Theorem 14.40. Indeed, the Kronecker type symbols there are then the usual ones, as shown by:

$$\begin{aligned}\delta_{\xi\pi}(i) &= \langle \xi_\pi, e_{i_1} \otimes \dots \otimes e_{i_k} \rangle \\ &= \left\langle \sum_j \delta_\pi(j_1, \dots, j_k) e_{j_1} \otimes \dots \otimes e_{j_k}, e_{i_1} \otimes \dots \otimes e_{i_k} \right\rangle \\ &= \delta_\pi(i_1, \dots, i_k)\end{aligned}$$

The Gram matrix being as well the correct one, we obtain the result. \square

Let us go back now to the general easy groups $G \subset U_N$, with the idea in mind of computing the laws of truncated characters. First, we have the following formula:

PROPOSITION 14.42. *The moments of truncated characters are given by the formula*

$$\int_G (g_{11} + \dots + g_{ss})^k dg = \text{Tr}(W_{kN} G_{ks})$$

where G_{kN} and $W_{kN} = G_{kN}^{-1}$ are the associated Gram and Weingarten matrices.

PROOF. We have indeed the following computation:

$$\begin{aligned}\int_G (g_{11} + \dots + g_{ss})^k dg &= \sum_{i_1=1}^s \dots \sum_{i_k=1}^s \int_G g_{i_1 i_1} \dots g_{i_k i_k} dg \\ &= \sum_{\pi, \nu \in D(k)} W_{kN}(\pi, \nu) \sum_{i_1=1}^s \dots \sum_{i_k=1}^s \delta_\pi(i) \delta_\nu(i) \\ &= \sum_{\pi, \nu \in D(k)} W_{kN}(\pi, \nu) G_{ks}(\nu, \pi) \\ &= \text{Tr}(W_{kN} G_{ks})\end{aligned}$$

Thus, we have reached to the formula in the statement. \square

In order to process now the above formula, and reach to concrete results, we must impose on our group a uniformity condition. Let us start with:

PROPOSITION 14.43. *For an easy group $G = (G_N)$, coming from a category of partitions $D \subset P$, the following conditions are equivalent:*

- (1) $G_{N-1} = G_N \cap U_{N-1}$, via the embedding $U_{N-1} \subset U_N$ given by $u \rightarrow \text{diag}(u, 1)$.
- (2) $G_{N-1} = G_N \cap U_{N-1}$, via the N possible diagonal embeddings $U_{N-1} \subset U_N$.
- (3) D is stable under the operation which consists in removing blocks.

If these conditions are satisfied, we say that $G = (G_N)$ is uniform.

PROOF. The equivalence (1) \iff (2) comes from the inclusion $S_N \subset G_N$, which makes everything S_N -invariant. Regarding (1) \iff (3), given a subgroup $K \subset_v U_{N-1}$, consider the matrix $u = \text{diag}(v, 1)$. Our claim is that for any $\pi \in P(k)$ we have:

$$\xi_\pi \in \text{Fix}(u^{\otimes k}) \iff \xi_{\pi'} \in \text{Fix}(u^{\otimes k'}), \forall \pi' \in P(k'), \pi' \subset \pi$$

In order to prove this claim, we must study the condition on the left. We have:

$$\begin{aligned} \xi_\pi \in \text{Fix}(u^{\otimes k}) &\iff (u^{\otimes k} \xi_\pi)_{i_1 \dots i_k} = (\xi_\pi)_{i_1 \dots i_k}, \forall i \\ &\iff \sum_j (u^{\otimes k})_{i_1 \dots i_k, j_1 \dots j_k} (\xi_\pi)_{j_1 \dots j_k} = (\xi_\pi)_{i_1 \dots i_k}, \forall i \\ &\iff \sum_j \delta_\pi(j_1, \dots, j_k) u_{i_1 j_1} \dots u_{i_k j_k} = \delta_\pi(i_1, \dots, i_k), \forall i \end{aligned}$$

Now let us recall that our representation has the special form $u = \text{diag}(v, 1)$. We conclude from this that for any index $a \in \{1, \dots, k\}$, we have:

$$i_a = N \implies j_a = N$$

With this observation in hand, if we denote by i', j' the multi-indices obtained from i, j obtained by erasing all the above $i_a = j_a = N$ values, and by $k' \leq k$ the common length of these new multi-indices, our condition becomes:

$$\sum_{j'} \delta_\pi(j_1, \dots, j_k) (u^{\otimes k'})_{i' j'} = \delta_\pi(i_1, \dots, i_k), \forall i$$

Here the index j is by definition obtained from the index j' by filling with N values. In order to finish now, we have two cases, depending on i , as follows:

Case 1. Assume that the index set $\{a | i_a = N\}$ corresponds to a certain subpartition $\pi' \subset \pi$. In this case, the N values will not matter, and our formula becomes:

$$\sum_{j'} \delta_\pi(j'_1, \dots, j'_{k'}) (u^{\otimes k'})_{i' j'} = \delta_\pi(i'_1, \dots, i'_{k'})$$

Case 2. Assume now the opposite, namely that the set $\{a | i_a = N\}$ does not correspond to a subpartition $\pi' \subset \pi$. In this case the indices mix, and our formula reads $0 = 0$. Thus we have $\xi_{\pi'} \in \text{Fix}(u^{\otimes k'})$ in both cases, for any subpartition $\pi' \subset \pi$, as desired. \square

Now back to the laws of truncated characters, we have the following result:

THEOREM 14.44. *For a uniform easy group $G = (G_N)$, we have the formula*

$$\lim_{N \rightarrow \infty} \int_{G_N} \chi_t^k = \sum_{\pi \in D(k)} t^{|\pi|}$$

with $D \subset P$ being the associated category of partitions.

PROOF. We use Proposition 14.42. With $s = [tN]$, the formula there becomes:

$$\int_{G_N} \chi_t^k = \text{Tr}(W_{kN} G_{k[tN]})$$

The point now is that in the uniform case the Gram matrix, and so the Weingarten matrix too, is asymptotically diagonal. Thus, we obtain the following estimate:

$$\begin{aligned} \int_{G_N} \chi_t^k &\simeq \sum_{\pi \in D(k)} W_{kN}(\pi, \pi) G_{k[tN]}(\pi, \pi) \\ &\simeq \sum_{\pi \in D(k)} N^{-|\pi|} (tN)^{|\pi|} \\ &= \sum_{\pi \in D(k)} t^{|\pi|} \end{aligned}$$

Thus, we are led to the formula in the statement. \square

We can now enlarge our collection of truncated character results, and we have:

THEOREM 14.45. *With $N \rightarrow \infty$, the laws of truncated characters are as follows:*

- (1) *For O_N we obtain the Gaussian law g_t .*
- (2) *For U_N we obtain the complex Gaussian law G_t .*
- (3) *For S_N we obtain the Poisson law p_t .*
- (4) *For H_N we obtain the Bessel law b_t .*
- (5) *For H_N^s we obtain the generalized Bessel law b_t^s .*
- (6) *For K_N we obtain the complex Bessel law B_t .*

PROOF. We already know these results at $t = 1$. In the general case, $t > 0$, these follow via some standard combinatorics, from the formula in Theorem 14.44. \square

14e. Exercises

This was a technical and exciting chapter, and as exercises on this, we have:

EXERCISE 14.46. *Learn the decomposition result for the $*$ -algebras $A \subset M_N(\mathbb{C})$.*

EXERCISE 14.47. *Learn more about central functions, on finite and compact groups.*

EXERCISE 14.48. *Find and read one of the available proofs of Tannakian duality.*

EXERCISE 14.49. *Learn as well about the theorems of Deligne, and Doplicher-Roberts.*

EXERCISE 14.50. *Work out all the details for the easiness property of H_N^s .*

EXERCISE 14.51. *Compute the Gram determinants for other easy groups.*

EXERCISE 14.52. *Work out all the details for our main character computations.*

EXERCISE 14.53. *Work out too the details for our truncated character computations.*

As bonus exercise, learn as well about Haar integration over locally compact groups.

CHAPTER 15

Random matrices

15a. Spectral measures

Welcome to the random matrices, which are first class mathematics and physics. In order to talk about such matrices and their spectral measures, we need to do some more linear algebra in infinite dimensions, as a continuation of our operator theory discussion from chapters 10 and 12. It is convenient to upgrade our formalism, as follows:

DEFINITION 15.1. *An abstract operator algebra, or C^* -algebra, is a complex algebra A having a norm $\|\cdot\|$ and an involution $*$, subject to the following conditions:*

- (1) *A is closed with respect to the norm.*
- (2) *We have $\|aa^*\| = \|a\|^2$, for any $a \in A$.*

In other words, what we did here is to axiomatize the abstract properties of the operator algebras $A \subset B(H)$, coming from the various general results about linear operators from chapter 10, without any reference to the ambient Hilbert space H .

As basic examples here, we have the usual matrix algebras $M_N(\mathbb{C})$, with the norm and the involution being the usual matrix norm and involution, given by:

$$\|A\| = \sup_{\|x\|=1} \|Ax\| \quad , \quad (A^*)_{ij} = \overline{A_{ji}}$$

Some other basic examples are the algebras $L^\infty(X)$ of essentially bounded functions $f : X \rightarrow \mathbb{C}$ on a measured space X , with the usual norm and involution, namely:

$$\|f\| = \sup_{x \in X} |f(x)| \quad , \quad f^*(x) = \overline{f(x)}$$

We can put these two basic classes of examples together, as follows:

PROPOSITION 15.2. *The random matrix algebras $A = M_N(L^\infty(X))$ are C^* -algebras, with their usual norm and involution, given by:*

$$\|Z\| = \sup_{x \in X} \|Z_x\| \quad , \quad (Z^*)_{ij} = \overline{Z_{ij}}$$

These algebras generalize both the algebras $M_N(\mathbb{C})$, and the algebras $L^\infty(X)$.

PROOF. The fact that the C^* -algebra axioms are satisfied is clear from definitions. As for the last assertion, this follows by taking $X = \{.\}$ and $N = 1$, respectively. \square

We can in fact say more about the above algebras, as follows:

THEOREM 15.3. *Any algebra of type $L^\infty(X)$ is an operator algebra, as follows:*

$$L^\infty(X) \subset B(L^2(X)) \quad , \quad f \rightarrow (g \rightarrow fg)$$

More generally, any random matrix algebra is an operator algebra, as follows,

$$M_N(L^\infty(X)) \subset B(\mathbb{C}^N \otimes L^2(X))$$

with the embedding being the above one, tensored with the identity.

PROOF. We have two assertions to be proved, the idea being as follows:

(1) Given $f \in L^\infty(X)$, consider the following operator, acting on $H = L^2(X)$:

$$T_f(g) = fg$$

Observe that T_f is indeed well-defined, and bounded as well, because:

$$\|fg\|_2 = \sqrt{\int_X |f(x)|^2 |g(x)|^2 d\mu(x)} \leq \|f\|_\infty \|g\|_2$$

The application $f \rightarrow T_f$ being linear, involutive, continuous, and injective as well, we obtain in this way a C^* -algebra embedding $L^\infty(X) \subset B(H)$, as desired.

(2) Regarding the second assertion, this is best viewed in the following way:

$$\begin{aligned} M_N(L^\infty(X)) &= M_N(\mathbb{C}) \otimes L^\infty(X) \\ &\subset M_N(\mathbb{C}) \otimes B(L^2(X)) \\ &= B(\mathbb{C}^N \otimes L^2(X)) \end{aligned}$$

Here we have used (1), and some standard tensor product identifications. \square

Our purpose in what follows is to develop the spectral theory of the C^* -algebras, and in particular that of the random matrix algebras $A = M_N(L^\infty(X))$ that we are interested in, one of our objectives being that of talking about spectral measures, in the normal case, in analogy with what we know about the usual matrices. Let us start with:

THEOREM 15.4. *Given an element $a \in A$ of a C^* -algebra, define its spectrum as:*

$$\sigma(a) = \left\{ \lambda \in \mathbb{C} \mid a - \lambda \notin A^{-1} \right\}$$

The following spectral theory results hold, exactly as in the $A = B(H)$ case:

- (1) *We have $\sigma(ab) \cup \{0\} = \sigma(ba) \cup \{0\}$.*
- (2) *We have $\sigma(f(a)) = f(\sigma(a))$, for any $f \in \mathbb{C}(X)$ having poles outside $\sigma(a)$.*
- (3) *The spectrum $\sigma(a)$ is compact, non-empty, and contained in $D_0(\|a\|)$.*
- (4) *The spectra of unitaries ($u^* = u^{-1}$) and self-adjoints ($a = a^*$) are on \mathbb{T}, \mathbb{R} .*
- (5) *The spectral radius of normal elements ($aa^* = a^*a$) is given by $\rho(a) = \|a\|$.*

In addition, assuming $a \in A \subset B$, the spectra of a with respect to A and to B coincide.

PROOF. Here the assertions (1-5), which are of course formulated a bit informally, are well-known for the full operator algebra $A = B(H)$, and the proof in general is similar:

(1) Assuming that $1 - ab$ is invertible, with inverse c , we have $abc = cab = c - 1$, and it follows that $1 - ba$ is invertible too, with inverse $1 + bca$. Thus $\sigma(ab), \sigma(ba)$ agree on $1 \in \mathbb{C}$, and by linearity, it follows that $\sigma(ab), \sigma(ba)$ agree on any point $\lambda \in \mathbb{C}^*$.

(2) The formula $\sigma(f(a)) = f(\sigma(a))$ is clear for polynomials, $f \in \mathbb{C}[X]$, by factorizing $f - \lambda$, with $\lambda \in \mathbb{C}$. Then, the extension to the rational functions is straightforward, because $P(a)/Q(a) - \lambda$ is invertible precisely when $P(a) - \lambda Q(a)$ is.

(3) By using $1/(1 - b) = 1 + b + b^2 + \dots$ for $\|b\| < 1$ we obtain that $a - \lambda$ is invertible for $|\lambda| > \|a\|$, and so $\sigma(a) \subset D_0(\|a\|)$. It is also clear that $\sigma(a)$ is closed, so what we have is a compact set. Finally, assuming $\sigma(a) = \emptyset$ the function $f(\lambda) = \varphi((a - \lambda)^{-1})$ is well-defined, for any $\varphi \in A^*$, and by Liouville we get $f = 0$, contradiction.

(4) Assuming $u^* = u^{-1}$ we have $\|u\| = 1$, and so $\sigma(u) \subset D_0(1)$. But with $f(z) = z^{-1}$ we obtain via (2) that we have as well $\sigma(u) \subset f(D_0(1))$, and this gives $\sigma(u) \subset \mathbb{T}$. As for the result regarding the self-adjoints, this can be obtained from the result for the unitaries, by using (2) with functions of type $f(z) = (z + it)/(z - it)$, with $t \in \mathbb{R}$.

(5) It is routine to check, by integrating quantities of type $z^n/(z - a)$ over circles centered at the origin, and estimating, that the spectral radius is given by $\rho(a) = \lim \|a^n\|^{1/n}$. But in the self-adjoint case, $a = a^*$, this gives $\rho(a) = \|a\|$, by using exponents of type $n = 2^k$, and then the extension to the general normal case is straightforward.

(6) Regarding now the last assertion, the inclusion $\sigma_B(a) \subset \sigma_A(a)$ is clear. For the converse, assume $a - \lambda \in B^{-1}$, and set $b = (a - \lambda)^*(a - \lambda)$. We have then:

$$\sigma_A(b) - \sigma_B(b) = \left\{ \mu \in \mathbb{C} - \sigma_B(b) \mid (b - \mu)^{-1} \in B - A \right\}$$

Thus this difference is an open subset of \mathbb{C} . On the other hand b being self-adjoint, its two spectra are both real, and so is their difference. Thus the two spectra of b are equal, and in particular b is invertible in A , and so $a - \lambda \in A^{-1}$, as desired. \square

We can now prove a key result, as follows:

THEOREM 15.5 (Gelfand). *If X is a compact space, the algebra $C(X)$ of continuous functions on it $f : X \rightarrow \mathbb{C}$ is a C^* -algebra, with usual norm and involution, namely:*

$$\|f\| = \sup_{x \in X} |f(x)| \quad , \quad f^*(x) = \overline{f(x)}$$

Conversely, any commutative C^ -algebra is of this form, $A = C(X)$, with*

$$X = \left\{ \chi : A \rightarrow \mathbb{C} \text{ , normed algebra character} \right\}$$

with topology making continuous the evaluation maps $ev_a : \chi \rightarrow \chi(a)$.

PROOF. There are several things going on here, the idea being as follows:

(1) The first assertion is clear from definitions. Observe that we have indeed:

$$\|ff^*\| = \sup_{x \in X} |f(x)|^2 = \|f\|^2$$

Observe also that the algebra $C(X)$ is commutative, because $fg = gf$.

(2) Conversely, given a commutative C^* -algebra A , let us define X as in the statement. Then X is compact, and $a \rightarrow ev_a$ is a morphism of algebras, as follows:

$$ev : A \rightarrow C(X)$$

(3) We first prove that ev is involutive. We use the following formula, which is similar to the $z = Re(z) + iIm(z)$ decomposition formula for usual complex numbers:

$$a = \frac{a + a^*}{2} + i \cdot \frac{a - a^*}{2i}$$

Thus it is enough to prove $ev_{a^*} = ev_a^*$ for the self-adjoint elements a . But this is the same as proving that $a = a^*$ implies that ev_a is a real function, which is in turn true, by Theorem 15.4, because $ev_a(\chi) = \chi(a)$ is an element of $\sigma(a)$, contained in \mathbb{R} .

(4) Since A is commutative, each element is normal, so ev is isometric:

$$\|ev_a\| = \rho(a) = \|a\|$$

It remains to prove that ev is surjective. But this follows from the Stone-Weierstrass theorem, because $ev(A)$ is a closed subalgebra of $C(X)$, which separates the points. \square

As a main consequence of the Gelfand theorem, we have:

THEOREM 15.6. *For any normal element $a \in A$ we have an identification as follows:*

$$\langle a \rangle = C(\sigma(a))$$

In addition, given a function $f \in C(\sigma(a))$, we can apply it to a , and we have

$$\sigma(f(a)) = f(\sigma(a))$$

which generalizes the previous rational calculus formula, in the normal case.

PROOF. Since a is normal, the C^* -algebra $\langle a \rangle$ that it generates is commutative, so if we denote by X the space of the characters $\chi : \langle a \rangle \rightarrow \mathbb{C}$, we have:

$$\langle a \rangle = C(X)$$

Now since the map $X \rightarrow \sigma(a)$ given by evaluation at a is bijective, we obtain:

$$\langle a \rangle = C(\sigma(a))$$

Thus, we are dealing here with usual functions, and this gives all the assertions. \square

In order to get now towards noncommutative probability, we first have to develop the theory of positive elements, and linear forms. First, we have the following result:

PROPOSITION 15.7. *For an element $a \in A$, the following are equivalent:*

- (1) *a is positive, in the sense that $\sigma(a) \subset [0, \infty)$.*
- (2) *$a = b^2$, for some $b \in A$ satisfying $b = b^*$.*
- (3) *$a = cc^*$, for some $c \in A$.*

PROOF. This is something very standard, as follows:

(1) \implies (2) Observe first that $\sigma(a) \subset \mathbb{R}$ implies $a = a^*$. Thus the algebra $\langle a \rangle$ is commutative, and by using Theorem 15.6, we can set $b = \sqrt{a}$.

(2) \implies (3) This is trivial, because we can simply set $c = b$.

(2) \implies (1) This is clear too, because we have:

$$\sigma(a) = \sigma(b^2) = \sigma(b)^2 \subset \mathbb{R}^2 = [0, \infty)$$

(3) \implies (1) We proceed by contradiction. By multiplying c by a suitable element of $\langle cc^* \rangle$, we are led to the existence of an element $d \neq 0$ satisfying:

$$-dd^* \geq 0$$

By writing now $d = x + iy$ with $x = x^*, y = y^*$ we have:

$$dd^* + d^*d = 2(x^2 + y^2) \geq 0$$

Thus $d^*d \geq 0$, which is easily seen to contradict the condition $-dd^* \geq 0$. \square

We can talk as well about positive linear forms, as follows:

DEFINITION 15.8. *Consider a linear map $\varphi : A \rightarrow \mathbb{C}$.*

- (1) *φ is called positive when $a \geq 0 \implies \varphi(a) \geq 0$.*
- (2) *φ is called faithful and positive when $a \geq 0, a \neq 0 \implies \varphi(a) > 0$.*

In the commutative case, $A = C(X)$, the positive linear forms appear as follows, with μ being positive, and strictly positive if we want φ to be faithful and positive:

$$\varphi(f) = \int_X f(x) d\mu(x)$$

In general, the positive linear forms can be thought of as being integration functionals with respect to some underlying “positive measures”. We have:

DEFINITION 15.9. *Let A be a C^* -algebra, given with a positive trace $tr : A \rightarrow \mathbb{C}$.*

- (1) *The elements $a \in A$ are called random variables.*
- (2) *The moments of such a variable are the numbers $M_k(a) = tr(a^k)$.*
- (3) *The law of such a variable is the functional $\mu_a : P \rightarrow tr(P(a))$.*

Here the exponent $k = \circ \bullet \bullet \circ \dots$ is by definition a colored integer, and the powers a^k are defined by the following formulae, and multiplicativity:

$$a^\emptyset = 1 \quad , \quad a^\circ = a \quad , \quad a^\bullet = a^*$$

As for the polynomial P , this is a noncommuting $*$ -polynomial in one variable:

$$P \in \mathbb{C} \langle X, X^* \rangle$$

Observe that the law is uniquely determined by the moments, because we have:

$$P(X) = \sum_k \lambda_k X^k \implies \mu_a(P) = \sum_k \lambda_k M_k(a)$$

At the level of the general theory, we have the following key result, extending the various results that we have, regarding the self-adjoint and normal matrices:

THEOREM 15.10. *Let A be a C^* -algebra, with a trace tr , and consider an element $a \in A$ which is normal, in the sense that $aa^* = a^*a$.*

- (1) μ_a is a complex probability measure, satisfying $\text{supp}(\mu_a) \subset \sigma(a)$.
- (2) In the self-adjoint case, $a = a^*$, this measure μ_a is real.
- (3) Assuming that tr is faithful, we have $\text{supp}(\mu_a) = \sigma(a)$.

PROOF. This is something very standard, that we already know for the usual complex matrices, and whose proof in general is quite similar, as follows:

(1) In the normal case, $aa^* = a^*a$, the Gelfand theorem, or rather the subsequent continuous functional calculus theorem, tells us that we have:

$$\langle a \rangle = C(\sigma(a))$$

Thus the functional $f(a) \rightarrow tr(f(a))$ can be regarded as an integration functional on the algebra $C(\sigma(a))$, and by the Riesz theorem, this latter functional must come from a probability measure μ on the spectrum $\sigma(a)$, in the sense that we must have:

$$tr(f(a)) = \int_{\sigma(a)} f(z) d\mu(z)$$

We are therefore led to the conclusions in the statement, with the uniqueness assertion coming from the fact that the elements a^k , taken as usual with respect to colored integer exponents, $k = \circ \bullet \bullet \circ \dots$, generate the whole C^* -algebra $C(\sigma(a))$.

(2) This is something which is clear from definitions.

(3) Once again, this is something which is clear from definitions. □

As a first concrete application now, by getting back to the random matrices, and to the various questions raised in the beginning of this chapter, we have:

THEOREM 15.11. *Given a random matrix $Z \in M_N(L^\infty(X))$ which is normal,*

$$ZZ^* = Z^*Z$$

its law, which is by definition the following abstract functional,

$$\mu : \mathbb{C} \langle X, X^* \rangle \rightarrow \mathbb{C} \quad , \quad P \rightarrow \frac{1}{N} \int_X \text{tr}(P(Z))$$

when restricted to the usual polynomials in two variables,

$$\mu : \mathbb{C}[X, X^*] \rightarrow \mathbb{C} \quad , \quad P \rightarrow \frac{1}{N} \int_X \text{tr}(P(Z))$$

must come from a probability measure on the spectrum $\sigma(Z) \subset \mathbb{C}$, as follows:

$$\mu(P) = \int_{\sigma(Z)} P(x) d\mu(x)$$

We agree to use the symbol μ for all these notions.

PROOF. This follows indeed from what we know from Theorem 15.10, applied to the normal element $a = Z$, belonging to the C^* -algebra $A = M_N(L^\infty(X))$. \square

15b. Gaussian matrices

We have now all the needed ingredients for launching some explicit random matrix computations. Our goal will be that of computing the asymptotic moments, and then the asymptotic laws, with $N \rightarrow \infty$, for the main classes of large random matrices.

Let us begin by specifying the precise classes of matrices that we are interested in. First we have the complex Gaussian matrices, which are constructed as follows:

DEFINITION 15.12. *A complex Gaussian matrix is a random matrix of type*

$$Z \in M_N(L^\infty(X))$$

which has i.i.d. centered complex normal entries.

To be more precise, the assumption in this definition is that all the matrix entries Z_{ij} are independent, and follow the same complex normal law G_t , for a fixed value of $t > 0$. We will see that the above matrices have an interesting, and “central” combinatorics, among all kinds of random matrices, with the study of the other random matrices being usually obtained as a modification of the study of the Gaussian matrices.

As a somewhat surprising remark, using real normal variables in Definition 15.12, instead of the complex ones appearing there, leads nowhere. The correct real versions of the Gaussian matrices are the Wigner random matrices, constructed as follows:

DEFINITION 15.13. *A Wigner matrix is a random matrix of type*

$$Z \in M_N(L^\infty(X))$$

which has i.i.d. centered complex normal entries, up to the constraint $Z = Z^$.*

This definition is something a bit compacted, and to be more precise, a Wigner matrix is by definition a random matrix as follows, with the diagonal entries being real normal variables, $a_i \sim g_t$, for some $t > 0$, the upper diagonal entries being complex normal variables, $b_{ij} \sim G_t$, the lower diagonal entries being the conjugates of the upper diagonal entries, as indicated, and with all the variables a_i, b_{ij} being independent:

$$Z = \begin{pmatrix} a_1 & b_{12} & \dots & \dots & b_{1N} \\ \bar{b}_{12} & a_2 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & a_{N-1} & b_{N-1,N} \\ \bar{b}_{1N} & \dots & \dots & \bar{b}_{N-1,N} & a_N \end{pmatrix}$$

As a comment here, for many concrete applications the Wigner matrices are in fact the central objects in random matrix theory, and in particular, they are often more important than the Gaussian matrices. In fact, these are the random matrices which were first considered and investigated, a long time ago, by Wigner himself.

However, as we will soon discover, the Gaussian matrices are somehow more fundamental than the Wigner matrices, at least from an abstract point of view, and this will be the point of view that we will follow here, with the Gaussian matrices coming first.

Finally, we will be interested as well in the complex Wishart matrices, which are the positive versions of the above random matrices, constructed as follows:

DEFINITION 15.14. *A complex Wishart matrix is a random matrix of type*

$$Z = YY^* \in M_N(L^\infty(X))$$

with Y being a complex Gaussian matrix.

As before with the Gaussian and Wigner matrices, there are many possible comments that can be made here, of technical or historical nature. As a first key fact, using real Gaussian variables instead of complex ones leads to a less interesting combinatorics. Also, these matrices were introduced and studied by Marchenko and Pastur not long after Wigner, and so historically came second. Finally, in what regards their combinatorics and applications, these matrices quite often come first, before both the Gaussian and the Wigner ones, with all this being of course a matter of knowledge and taste.

Summarizing, we have three main types of random matrices, which can be thought of as being “complex”, “real” and “positive”, and that we will study in what follows, in this

precise order, with this order being the one that fits us best here. Let us also mention that there are many other interesting classes of random matrices, which are more specialized, usually appearing as modifications of the above. More on these later.

In order to compute the asymptotic laws of the Gaussian, Wigner and Wishart matrices, we use the moment method. We first have the following result:

THEOREM 15.15. *Given a sequence of Gaussian random matrices*

$$Z_N \in M_N(L^\infty(X))$$

having independent G_t variables as entries, for some fixed $t > 0$, we have

$$M_k \left(\frac{Z_N}{\sqrt{N}} \right) \simeq t^{|k|/2} |\mathcal{NC}_2(k)|$$

for any colored integer $k = \circ \bullet \bullet \circ \dots$, in the $N \rightarrow \infty$ limit.

PROOF. This is something standard, which can be done as follows:

(1) We fix $N \in \mathbb{N}$, and we let $Z = Z_N$. Let us first compute the trace of Z^k . With $k = k_1 \dots k_s$, and with the convention $(ij)^\circ = ij$, $(ij)^\bullet = ji$, we have:

$$\begin{aligned} \text{Tr}(Z^k) &= \text{Tr}(Z^{k_1} \dots Z^{k_s}) \\ &= \sum_{i_1=1}^N \dots \sum_{i_s=1}^N (Z^{k_1})_{i_1 i_2} (Z^{k_2})_{i_2 i_3} \dots (Z^{k_s})_{i_s i_1} \\ &= \sum_{i_1=1}^N \dots \sum_{i_s=1}^N (Z_{(i_1 i_2) k_1})^{k_1} (Z_{(i_2 i_3) k_2})^{k_2} \dots (Z_{(i_s i_1) k_s})^{k_s} \end{aligned}$$

(2) Next, we rescale our variable Z by a \sqrt{N} factor, as in the statement, and we also replace the usual trace by its normalized version, $tr = \text{Tr}/N$. Our formula becomes:

$$tr \left(\left(\frac{Z}{\sqrt{N}} \right)^k \right) = \frac{1}{N^{s/2+1}} \sum_{i_1=1}^N \dots \sum_{i_s=1}^N (Z_{(i_1 i_2) k_1})^{k_1} (Z_{(i_2 i_3) k_2})^{k_2} \dots (Z_{(i_s i_1) k_s})^{k_s}$$

Thus, the moment that we are interested in is given by:

$$M_k \left(\frac{Z}{\sqrt{N}} \right) = \frac{1}{N^{s/2+1}} \sum_{i_1=1}^N \dots \sum_{i_s=1}^N \int_X (Z_{(i_1 i_2) k_1})^{k_1} (Z_{(i_2 i_3) k_2})^{k_2} \dots (Z_{(i_s i_1) k_s})^{k_s}$$

(3) Let us apply now the Wick formula, that we know since chapter 4. We conclude that the moment that we are interested in is given by the following formula:

$$\begin{aligned}
& M_k \left(\frac{Z}{\sqrt{N}} \right) \\
&= \frac{t^{s/2}}{N^{s/2+1}} \sum_{i_1=1}^N \dots \sum_{i_s=1}^N \# \left\{ \pi \in \mathcal{P}_2(k) \mid \pi \leq \ker \left((i_1 i_2)^{k_1}, (i_2 i_3)^{k_2}, \dots, (i_s i_1)^{k_s} \right) \right\} \\
&= t^{s/2} \sum_{\pi \in \mathcal{P}_2(k)} \frac{1}{N^{s/2+1}} \# \left\{ i \in \{1, \dots, N\}^s \mid \pi \leq \ker \left((i_1 i_2)^{k_1}, (i_2 i_3)^{k_2}, \dots, (i_s i_1)^{k_s} \right) \right\}
\end{aligned}$$

(4) Our claim now is that in the $N \rightarrow \infty$ limit the combinatorics of the above sum simplifies, with only the noncrossing partitions contributing to the sum, and with each of them contributing precisely with a 1 factor, so that we will have, as desired:

$$\begin{aligned}
M_k \left(\frac{Z}{\sqrt{N}} \right) &= t^{s/2} \sum_{\pi \in \mathcal{P}_2(k)} \left(\delta_{\pi \in NC_2(k)} + O(N^{-1}) \right) \\
&\simeq t^{s/2} \sum_{\pi \in \mathcal{P}_2(k)} \delta_{\pi \in NC_2(k)} \\
&= t^{s/2} |\mathcal{NC}_2(k)|
\end{aligned}$$

(5) In order to prove this, the first observation is that when k is not uniform, in the sense that it contains a different number of \circ , \bullet symbols, we have $\mathcal{P}_2(k) = \emptyset$, and so:

$$M_k \left(\frac{Z}{\sqrt{N}} \right) = t^{s/2} |\mathcal{NC}_2(k)| = 0$$

(6) Thus, we are left with the case where k is uniform. Let us examine first the case where k consists of an alternating sequence of \circ and \bullet symbols, as follows:

$$k = \underbrace{\circ \bullet \circ \bullet \dots \circ \bullet}_{2p}$$

In this case it is convenient to relabel our multi-index $i = (i_1, \dots, i_s)$, with $s = 2p$, in the form $(j_1, l_1, j_2, l_2, \dots, j_p, l_p)$. With this done, our moment formula becomes:

$$M_k \left(\frac{Z}{\sqrt{N}} \right) = t^p \sum_{\pi \in \mathcal{P}_2(k)} \frac{1}{N^{p+1}} \# \left\{ j, l \in \{1, \dots, N\}^p \mid \pi \leq \ker (j_1 l_1, j_2 l_1, j_2 l_2, \dots, j_1 l_p) \right\}$$

Now observe that, with k being as above, we have an identification $\mathcal{P}_2(k) \simeq S_p$, obtained in the obvious way. With this done too, our moment formula becomes:

$$M_k \left(\frac{Z}{\sqrt{N}} \right) = t^p \sum_{\pi \in S_p} \frac{1}{N^{p+1}} \# \left\{ j, l \in \{1, \dots, N\}^p \mid j_r = j_{\pi(r)+1}, l_r = l_{\pi(r)}, \forall r \right\}$$

(7) We are now ready to do our asymptotic study, and prove the claim in (4). Let indeed $\gamma \in S_p$ be the full cycle, which is by definition the following permutation:

$$\gamma = (1\ 2\ \dots\ p)$$

In terms of γ , the conditions $j_r = j_{\pi(r)+1}$ and $l_r = l_{\pi(r)}$ found above read:

$$\gamma\pi \leq \ker j \quad , \quad \pi \leq \ker l$$

Counting the number of free parameters in our moment formula, we obtain:

$$M_k \left(\frac{Z}{\sqrt{N}} \right) = \frac{t^p}{N^{p+1}} \sum_{\pi \in S_p} N^{|\pi|+|\gamma\pi|} = t^p \sum_{\pi \in S_p} N^{|\pi|+|\gamma\pi|-p-1}$$

(8) The point now is that the last exponent is well-known to be ≤ 0 , with equality precisely when the permutation $\pi \in S_p$ is geodesic, which in practice means that π must come from a noncrossing partition. Thus we obtain, in the $N \rightarrow \infty$ limit, as desired:

$$M_k \left(\frac{Z}{\sqrt{N}} \right) \simeq t^p |\mathcal{NC}_2(k)|$$

This finishes the proof in the case of the exponents k which are alternating, and the case where k is an arbitrary uniform exponent is similar, by permuting everything. \square

The above result is very nice, but the resulting asymptotic measure is still in need to be interpreted. For more on all this, we refer to free probability theory [89].

15c. Wigner and Wishart

Regarding now the Wigner matrices, we have here the following result, coming as a consequence of Theorem 15.15, via some simple algebraic manipulations:

THEOREM 15.16. *Given a sequence of Wigner random matrices*

$$Z_N \in M_N(L^\infty(X))$$

having independent G_t variables as entries, with $t > 0$, up to $Z_N = Z_N^$, we have*

$$M_k \left(\frac{Z_N}{\sqrt{N}} \right) \simeq t^{k/2} |\mathcal{NC}_2(k)|$$

for any integer $k \in \mathbb{N}$, in the $N \rightarrow \infty$ limit.

PROOF. This can be deduced from a direct computation based on the Wick formula, similar to that from the proof of Theorem 15.15, but the best is to deduce this result from Theorem 15.15 itself. Indeed, we know from there that for Gaussian matrices $Y_N \in M_N(L^\infty(X))$ we have the following formula, valid for any colored integer $K = \circ \bullet \bullet \circ \dots$, in the $N \rightarrow \infty$ limit, with \mathcal{NC}_2 standing for noncrossing matching pairings:

$$M_K \left(\frac{Y_N}{\sqrt{N}} \right) \simeq t^{|K|/2} |\mathcal{NC}_2(K)|$$

By doing some combinatorics, we deduce from this that we have the following formula for the moments of the matrices $Re(Y_N)$, with respect to usual exponents, $k \in \mathbb{N}$:

$$\begin{aligned}
 M_k \left(\frac{Re(Y_N)}{\sqrt{N}} \right) &= 2^{-k} \cdot M_k \left(\frac{Y_N}{\sqrt{N}} + \frac{Y_N^*}{\sqrt{N}} \right) \\
 &= 2^{-k} \sum_{|K|=k} M_K \left(\frac{Y_N}{\sqrt{N}} \right) \\
 &\simeq 2^{-k} \sum_{|K|=k} t^{k/2} |\mathcal{NC}_2(K)| \\
 &= 2^{-k} \cdot t^{k/2} \cdot 2^{k/2} |\mathcal{NC}_2(k)| \\
 &= 2^{-k/2} \cdot t^{k/2} |NC_2(k)|
 \end{aligned}$$

Now since the matrices $Z_N = \sqrt{2}Re(Y_N)$ are of Wigner type, this gives the result. \square

Now by putting everything together, we obtain the Wigner theorem, as follows:

THEOREM 15.17. *Given a sequence of Wigner random matrices*

$$Z_N \in M_N(L^\infty(X))$$

which by definition have i.i.d. complex normal entries, up to $Z_N = Z_N^$, we have*

$$Z_N \sim \gamma_t$$

in the $N \rightarrow \infty$ limit, where $\gamma_t = \frac{1}{2\pi t} \sqrt{4t - x^2} dx$ is the Wigner semicircle law.

PROOF. This follows indeed from Theorem 15.16, via some combinatorics, that we know from chapter 13, in order to recover the Wigner law, out of the Catalan numbers. \square

Let us discuss now the Wishart matrices, which are the positive analogues of the Wigner matrices. Quite surprisingly, the computation here leads to the Catalan numbers, but not in the same way as for the Wigner matrices, the result being as follows:

THEOREM 15.18. *Given a sequence of complex Wishart matrices*

$$W_N = Y_N Y_N^* \in M_N(L^\infty(X))$$

with Y_N being $N \times N$ complex Gaussian of parameter $t > 0$, we have

$$M_k \left(\frac{W_N}{N} \right) \simeq t^k C_k$$

for any exponent $k \in \mathbb{N}$, in the $N \rightarrow \infty$ limit.

PROOF. There are several possible proofs for this result, as follows:

(1) A first method is by using the formula that we have in Theorem 15.15, for the Gaussian matrices Y_N . Indeed, we know from there that we have the following formula, valid for any colored integer $K = \circ \bullet \bullet \circ \dots$, in the $N \rightarrow \infty$ limit:

$$M_K \left(\frac{Y_N}{\sqrt{N}} \right) \simeq t^{|K|/2} |\mathcal{NC}_2(K)|$$

With $K = \circ \bullet \bullet \circ \dots$, alternating word of lenght $2k$, with $k \in \mathbb{N}$, this gives:

$$M_k \left(\frac{Y_N Y_N^*}{N} \right) \simeq t^k |\mathcal{NC}_2(K)|$$

Thus, in terms of the Wishart matrix $W_N = Y_N Y_N^*$ we have, for any $k \in \mathbb{N}$:

$$M_k \left(\frac{W_N}{N} \right) \simeq t^k |\mathcal{NC}_2(K)|$$

The point now is that, by doing some combinatorics, we have:

$$|\mathcal{NC}_2(K)| = |\mathcal{NC}_2(2k)| = C_k$$

Thus, we are led to the formula in the statement.

(2) A second method, that we will explain now as well, is by proving the result directly, starting from definitions. The matrix entries of our matrix $W = W_N$ are given by:

$$W_{ij} = \sum_{r=1}^N Y_{ir} \bar{Y}_{jr}$$

Thus, the normalized traces of powers of W are given by the following formula:

$$\begin{aligned} \text{tr}(W^k) &= \frac{1}{N} \sum_{i_1=1}^N \dots \sum_{i_k=1}^N W_{i_1 i_2} W_{i_2 i_3} \dots W_{i_k i_1} \\ &= \frac{1}{N} \sum_{i_1=1}^N \dots \sum_{i_k=1}^N \sum_{r_1=1}^N \dots \sum_{r_k=1}^N Y_{i_1 r_1} \bar{Y}_{i_2 r_1} Y_{i_2 r_2} \bar{Y}_{i_3 r_2} \dots Y_{i_k r_k} \bar{Y}_{i_1 r_k} \end{aligned}$$

By rescaling now W by a $1/N$ factor, as in the statement, we obtain:

$$\text{tr} \left(\left(\frac{W}{N} \right)^k \right) = \frac{1}{N^{k+1}} \sum_{i_1=1}^N \dots \sum_{i_k=1}^N \sum_{r_1=1}^N \dots \sum_{r_k=1}^N Y_{i_1 r_1} \bar{Y}_{i_2 r_1} Y_{i_2 r_2} \bar{Y}_{i_3 r_2} \dots Y_{i_k r_k} \bar{Y}_{i_1 r_k}$$

By using now the Wick rule, we obtain the following formula for the moments, with $K = \circ \bullet \circ \bullet \dots$, alternating word of length $2k$, and with $I = (i_1 r_1, i_2 r_1, \dots, i_k r_k, i_1 r_k)$:

$$\begin{aligned} M_k \left(\frac{W}{N} \right) &= \frac{t^k}{N^{k+1}} \sum_{i_1=1}^N \cdots \sum_{i_k=1}^N \sum_{r_1=1}^N \cdots \sum_{r_k=1}^N \# \left\{ \pi \in \mathcal{P}_2(K) \mid \pi \leq \ker(I) \right\} \\ &= \frac{t^k}{N^{k+1}} \sum_{\pi \in \mathcal{P}_2(K)} \# \left\{ i, r \in \{1, \dots, N\}^k \mid \pi \leq \ker(I) \right\} \end{aligned}$$

In order to compute this quantity, we use the standard bijection $\mathcal{P}_2(K) \simeq S_k$. By identifying the pairings $\pi \in \mathcal{P}_2(K)$ with their counterparts $\pi \in S_k$, we obtain:

$$M_k \left(\frac{W}{N} \right) = \frac{t^k}{N^{k+1}} \sum_{\pi \in S_k} \# \left\{ i, r \in \{1, \dots, N\}^k \mid i_s = i_{\pi(s)+1}, r_s = r_{\pi(s)}, \forall s \right\}$$

Now let $\gamma \in S_k$ be the full cycle, which is by definition the following permutation:

$$\gamma = (1 \ 2 \ \dots \ k)$$

The general factor in the product computed above is then 1 precisely when following two conditions are simultaneously satisfied:

$$\gamma\pi \leq \ker i \quad , \quad \pi \leq \ker r$$

Counting the number of free parameters in our moment formula, we obtain:

$$M_k \left(\frac{W}{N} \right) = t^k \sum_{\pi \in S_k} N^{|\pi| + |\gamma\pi| - k - 1}$$

The point now is that the last exponent is well-known to be ≤ 0 , with equality precisely when the permutation $\pi \in S_k$ is geodesic, which in practice means that π must come from a noncrossing partition. Thus we obtain, in the $N \rightarrow \infty$ limit:

$$M_k \left(\frac{W}{N} \right) \simeq t^k C_k$$

Thus, we are led to the conclusion in the statement. □

We are led in this way to the following result:

THEOREM 15.19. *Given a sequence of complex Wishart matrices*

$$W_N = Y_N Y_N^* \in M_N(L^\infty(X))$$

with Y_N being $N \times N$ complex Gaussian of parameter $t > 0$, we have

$$\frac{W_N}{tN} \sim \frac{1}{2\pi} \sqrt{4x^{-1} - 1} \, dx$$

with $N \rightarrow \infty$, with the limiting measure being the Marchenko-Pastur law π_1 .

PROOF. This follows indeed from Theorem 15.18, via the combinatorics from chapter 13, in order to recover the Marchenko-Pastur law, out of the Catalan numbers. \square

Let us discuss now a generalization of the above results, motivated by a whole array of concrete questions, and bringing into the picture a “true” parameter $t > 0$, which is different from the parameter $t > 0$ used above, which is something quite trivial.

For this purpose, let us go back to the definition of the Wishart matrices. There were as follows, with Y being a $N \times N$ matrix with i.i.d. entries, each following the law G_t :

$$W = YY^*$$

The point now is that, more generally, we can use in this construction a $N \times M$ matrix Y with i.i.d. entries, each following the law G_t , with $M \in \mathbb{N}$ being arbitrary. Thus, we have a new parameter, and by ditching the old parameter $t > 0$, we are led to the following definition, which is the “true” definition of the Wishart matrices:

DEFINITION 15.20. *A complex Wishart matrix is a $N \times N$ matrix of the form*

$$W = YY^*$$

where Y is a $N \times M$ matrix with i.i.d. entries, each following the law G_1 .

In order to see now what is going on, combinatorially, let us compute moments. The result here is substantially more interesting than that for the previous Wishart matrices, with the new relevant numeric parameter being now the number $t = M/N$, as follows:

THEOREM 15.21. *Given a sequence of complex Wishart matrices*

$$W_N = Y_N Y_N^* \in M_N(L^\infty(X))$$

with Y_N being $N \times M$ complex Gaussian of parameter 1, we have

$$M_k \left(\frac{W_N}{N} \right) \simeq \sum_{\pi \in NC(k)} t^{|\pi|}$$

for any exponent $k \in \mathbb{N}$, in the $M = tN \rightarrow \infty$ limit.

PROOF. This is something which is very standard, as follows:

(1) Before starting, let us clarify the relation with our previous Wishart matrix results. In the case $M = N$ we have $t = 1$, and the formula in the statement reads:

$$M_k \left(\frac{W_N}{N} \right) \simeq |NC(k)|$$

Thus, what we have here is the previous Wishart matrix formula, in full generality, at the value $t = 1$ of our old parameter $t > 0$.

(2) Observe also that by rescaling, we can obtain if we want from this the previous Wishart matrix formula, in full generality, at any value $t > 0$ of our old parameter. Thus, things fine, we are indeed generalizing what we did before.

(3) In order to prove now the formula in the statement, we proceed as usual, by using the Wick formula. The matrix entries of our Wishart matrix $W = W_N$ are given by:

$$W_{ij} = \sum_{r=1}^M Y_{ir} \bar{Y}_{jr}$$

Thus, the normalized traces of powers of W are given by the following formula:

$$\begin{aligned} \text{tr}(W^k) &= \frac{1}{N} \sum_{i_1=1}^N \cdots \sum_{i_k=1}^N W_{i_1 i_2} W_{i_2 i_3} \cdots W_{i_k i_1} \\ &= \frac{1}{N} \sum_{i_1=1}^N \cdots \sum_{i_k=1}^N \sum_{r_1=1}^M \cdots \sum_{r_k=1}^M Y_{i_1 r_1} \bar{Y}_{i_2 r_1} Y_{i_2 r_2} \bar{Y}_{i_3 r_2} \cdots Y_{i_k r_k} \bar{Y}_{i_1 r_k} \end{aligned}$$

By rescaling now W by a $1/N$ factor, as in the statement, we obtain:

$$\text{tr} \left(\left(\frac{W}{N} \right)^k \right) = \frac{1}{N^{k+1}} \sum_{i_1=1}^N \cdots \sum_{i_k=1}^N \sum_{r_1=1}^M \cdots \sum_{r_k=1}^M Y_{i_1 r_1} \bar{Y}_{i_2 r_1} Y_{i_2 r_2} \bar{Y}_{i_3 r_2} \cdots Y_{i_k r_k} \bar{Y}_{i_1 r_k}$$

(4) By using now the Wick rule, we obtain the following formula for the moments, with $K = \circ \bullet \circ \bullet \dots$, alternating word of length $2k$, and $I = (i_1 r_1, i_2 r_1, \dots, i_k r_k, i_1 r_k)$:

$$\begin{aligned} M_k \left(\frac{W}{N} \right) &= \frac{1}{N^{k+1}} \sum_{i_1=1}^N \cdots \sum_{i_k=1}^N \sum_{r_1=1}^M \cdots \sum_{r_k=1}^M \# \left\{ \pi \in \mathcal{P}_2(K) \mid \pi \leq \ker I \right\} \\ &= \frac{1}{N^{k+1}} \sum_{\pi \in \mathcal{P}_2(K)} \# \left\{ i \in \{1, \dots, N\}^k, r \in \{1, \dots, M\}^k \mid \pi \leq \ker I \right\} \end{aligned}$$

(5) In order to compute this quantity, we use the standard bijection $\mathcal{P}_2(K) \simeq S_k$. By identifying the pairings $\pi \in \mathcal{P}_2(K)$ with their counterparts $\pi \in S_k$, we obtain:

$$M_k \left(\frac{W}{N} \right) = \frac{1}{N^{k+1}} \sum_{\pi \in S_k} \# \left\{ i \in \{1, \dots, N\}^k, r \in \{1, \dots, M\}^k \mid i_s = i_{\pi(s)+1}, r_s = r_{\pi(s)} \right\}$$

Now let $\gamma \in S_k$ be the full cycle, which is by definition the following permutation:

$$\gamma = (1 \ 2 \ \dots \ k)$$

The general factor in the product computed above is then 1 precisely when following two conditions are simultaneously satisfied:

$$\gamma \pi \leq \ker i \quad , \quad \pi \leq \ker r$$

Counting the number of free parameters in our expectation formula, we obtain:

$$M_k \left(\frac{W}{N} \right) = \frac{1}{N^{k+1}} \sum_{\pi \in S_k} N^{|\gamma\pi|} M^{|\pi|} = \sum_{\pi \in S_k} N^{|\gamma\pi| - k - 1} M^{|\pi|}$$

(6) Now by using the same arguments as in the case $M = N$, from the proof of Theorem 15.18, we conclude that in the $M = tN \rightarrow \infty$ limit the permutations $\pi \in S_k$ which matter are those coming from noncrossing partitions, and so that we have:

$$M_k \left(\frac{W}{N} \right) \simeq \sum_{\pi \in NC(k)} N^{-|\pi|} M^{|\pi|} = \sum_{\pi \in NC(k)} t^{|\pi|}$$

We are therefore led to the conclusion in the statement. \square

In order to recapture now the density out of the moments, we can of course use the Stieltjes inversion formula, but the computations here are a bit opaque. So, inspired from what happens at $t = 1$, let us cheat a bit, and formulate things as follows:

DEFINITION 15.22. *The Marchenko-Pastur law π_t of parameter $t > 0$ is given by:*

$$a \sim \gamma_t \implies a^2 \sim \pi_t$$

That is, π_t the law of the square of a variable following the law γ_t .

This is certainly nice and simple, and we know that at $t = 1$ we obtain indeed the Marchenko-Pastur law π_1 , as constructed above. In general, we have:

PROPOSITION 15.23. *The Marchenko-Pastur law of parameter $t > 0$ is*

$$\pi_t = \max(1 - t, 0) \delta_0 + \frac{\sqrt{4t - (x - 1 - t)^2}}{2\pi x} dx$$

the support being $[0, 4t^2]$, and the moments of this measure are

$$M_k = \sum_{\pi \in NC(k)} t^{|\pi|}$$

exactly as for the asymptotic moments of the complex Wishart matrices.

PROOF. This follows as usual, by doing some computations, either combinatorics, or calculus. To be more precise, we have three formulae for π_t to be connected, namely the one in Definition 15.22, and the two ones from the present statement, and the connections between them can be established exactly as we did before, at $t = 1$. \square

Now back to the complex Wishart matrices that we are interested in, we can now formulate a final result regarding them, as follows:

THEOREM 15.24. *Given a sequence of complex Wishart matrices*

$$W_N = Y_N Y_N^* \in M_N(L^\infty(X))$$

with Y_N being $N \times M$ complex Gaussian of parameter 1, we have

$$\frac{W_N}{N} \sim \max(1-t, 0) \delta_0 + \frac{\sqrt{4t - (x-1-t)^2}}{2\pi x} dx$$

with $M = tN \rightarrow \infty$, with the limiting measure being the Marchenko-Pastur law π_t .

PROOF. This follows indeed from Theorem 15.21 and Proposition 15.23. \square

Many other things can be said, along these lines, and for more on all this, we refer to free probability theory [89], which has answers to nearly all potential questions that can be asked, regarding the various classes of random matrices investigated above.

15d. Block modifications

Our goal now will be that of explaining a surprising result, stating that when suitably block-transposing the entries of a complex Wishart matrix, we obtain as asymptotic distribution a shifted version of Wigner's semicircle law. Let us start with:

DEFINITION 15.25. *The partial transpose of a complex Wishart matrix W of parameters (dn, dm) is the matrix*

$$\tilde{W} = (id \otimes t)W$$

where id is the identity of $M_d(\mathbb{C})$, and t is the transposition of $M_n(\mathbb{C})$.

In more familiar terms of bases and indices, the standard decomposition $\mathbb{C}^{dn} = \mathbb{C}^d \otimes \mathbb{C}^n$ induces an algebra decomposition $M_{dn}(\mathbb{C}) = M_d(\mathbb{C}) \otimes M_n(\mathbb{C})$, and with this convention made, the partial transpose matrix \tilde{W} constructed above has entries as follows:

$$\tilde{W}_{ia,jb} = W_{ib,ja}$$

Our goal in what follows will be that of computing the law of \tilde{W} , first when d, n, m are fixed, and then in the $d \rightarrow \infty$ regime. For this purpose, we will need a number of standard facts regarding the noncrossing partitions. Let us start with:

PROPOSITION 15.26. *For a permutation $\sigma \in S_p$, we have the formula*

$$|\sigma| + \#\sigma = p$$

where $|\sigma|$ is the number of cycles of σ , and $\#\sigma$ is the minimal $k \in \mathbb{N}$ such that σ is a product of k transpositions. Also, the following formula defines a distance on S_p ,

$$(\sigma, \pi) \rightarrow \#(\sigma^{-1}\pi)$$

and the set of permutations $\sigma \in S_p$ which saturate the triangular inequality

$$\#\sigma + \#(\sigma^{-1}\gamma) = \#\gamma = p - 1$$

where $\gamma \in S_p$ is a full cycle, is in bijection with the set $NC(p)$.

PROOF. All this is standard combinatorics, that we will leave here as an exercise. \square

We will need as well the following well-known result:

PROPOSITION 15.27. *The number $||\pi||$ of blocks having even size is given by*

$$1 + ||\pi|| = |\pi\gamma|$$

for every noncrossing partition $\pi \in NC(p)$.

PROOF. We use a recurrence over the number of blocks of π . If π has just one block, its associated geodesic permutation is γ and we have:

$$|\gamma^2| = \begin{cases} 1 & (p \text{ odd}) \\ 2 & (p \text{ even}) \end{cases}$$

For partitions π with more than one block, we can assume without loss of generality that $\pi = \hat{1}_k \sqcup \pi'$, where $\hat{1}_k$ is a contiguous block of size k . Recall that the number of blocks of the permutation $\pi\gamma$ is given by the following formula, where $\rho_{14} \in P_2(2p)$ is the pair partition which pairs an element i with $i + (-1)^{i+1}3$:

$$|\pi\gamma| = |\tilde{\pi} \vee \rho_{14}|$$

If k is an even number, $k = 2r$, consider the following partition, which contains the block $(1\ 4\ 5\ 8 \dots 4r - 3\ 4r)$, along with the blocks coming from elements of the form $4i + 2, 4i + 3$ from $\{1, \dots, 4r\}$ and from π' :

$$\sigma = \widehat{\hat{1}_{2r}} \sqcup \pi' \vee \rho_{14}$$

We can count the blocks of the join of two partitions by drawing them one beneath the other and counting the number of connected components of the curve, without taking into account the possible crossings. We conclude that we have the following formula, where ρ'_{14} is ρ_{14} restricted to the set $\{2k + 1, 2k + 2, \dots, 2p\}$:

$$|\tilde{\pi} \vee \rho_{14}| = 1 + |\tilde{\pi}' \vee \rho'_{14}|$$

If k is odd, $k = 2r + 1$, there is no extra block appearing, so we have:

$$|\tilde{\pi} \vee \rho_{14}| = |\tilde{\pi}' \vee \rho'_{14}|$$

Thus, we are led to the conclusion in the statement. \square

We can now investigate the block-transposed Wishart matrices, and we have:

THEOREM 15.28. *For any $p \geq 1$ we have the formula*

$$\lim_{d \rightarrow \infty} (E \circ tr)(m\tilde{W})^p = \sum_{\pi \in NC(p)} m^{|\pi|} n^{||\pi||}$$

where $|\cdot|$ and $||\cdot||$ are the number of blocks, and the number of blocks of even size.

PROOF. The matrix elements of the partial transpose matrix are given by:

$$\tilde{W}_{ia,jb} = W_{ib,ja} = (dm)^{-1} \sum_{k=1}^d \sum_{c=1}^m G_{ib,kc} \bar{G}_{ja,kc}$$

This gives the following formula:

$$\begin{aligned} tr(\tilde{W}^p) &= (dn)^{-1} (dm)^{-p} \sum_{i_1, \dots, i_p=1}^d \sum_{a_1, \dots, a_p=1}^n \prod_{s=1}^p W_{i_s a_s, i_{s+1} a_{s+1}}^\Gamma \\ &= (dn)^{-1} (dm)^{-p} \sum_{i_1, \dots, i_p=1}^d \sum_{a_1, \dots, a_p=1}^n \prod_{s=1}^p W_{i_s a_{s+1}, i_{s+1} a_s} \\ &= (dn)^{-1} (dm)^{-p} \sum_{i_1, \dots, i_p=1}^d \sum_{a_1, \dots, a_p=1}^n \prod_{s=1}^p \sum_{j_1, \dots, j_p=1}^d \sum_{b_1, \dots, b_p=1}^m G_{i_s a_{s+1}, j_s b_s} \bar{G}_{i_{s+1} a_s, j_s b_s} \end{aligned}$$

The average of the general term can be computed by the Wick rule, namely:

$$E \left(\prod_{s=1}^p G_{i_s a_{s+1}, j_s b_s} \bar{G}_{i_{s+1} a_s, j_s b_s} \right) = \sum_{\pi \in S_p} \prod_{s=1}^p \delta_{i_s, i_{\pi(s)+1}} \delta_{a_{s+1}, a_{\pi(s)}} \delta_{j_s, j_{\pi(s)}} \delta_{b_s, b_{\pi(s)}}$$

Let $\gamma \in S_p$ be the full cycle $\gamma = (1 \ 2 \ \dots \ p)^{-1}$. The general factor in the above product is 1 if and only if the following four conditions are simultaneously satisfied:

$$\gamma^{-1}\pi \leq \ker i \quad , \quad \pi\gamma \leq \ker a \quad , \quad \pi \leq \ker j \quad , \quad \pi \leq \ker b$$

Counting the number of free parameters in the above equation, we obtain:

$$\begin{aligned} (E \circ tr)(\tilde{W}^p) &= (dn)^{-1} (dm)^{-p} \sum_{\pi \in S_p} d^{|\pi| + |\gamma^{-1}\pi|} m^{|\pi|} n^{|\pi\gamma|} \\ &= \sum_{\pi \in S_p} d^{|\pi| + |\gamma^{-1}\pi| - p - 1} m^{|\pi| - p} n^{|\pi\gamma| - 1} \end{aligned}$$

The exponent of d in the last expression on the right is:

$$\begin{aligned} N(\pi) &= |\pi| + |\gamma^{-1}\pi| - p - 1 \\ &= p - 1 - (\#\pi + \#(\gamma^{-1}\pi)) \\ &= p - 1 - (\#\pi + \#(\pi^{-1}\gamma)) \end{aligned}$$

As explained in the beginning of this section, this quantity is known to be ≤ 0 , with equality iff π is geodesic, hence associated to a noncrossing partition. Thus:

$$(E \circ tr)(\tilde{W}^p) = (1 + O(d^{-1})) m^{-p} n^{-1} \sum_{\pi \in NC(p)} m^{|\pi|} n^{|\pi\gamma|}$$

Together with $|\pi\gamma| = \|\pi\| + 1$, this gives the result. \square

We would like now to find an equation for the moment generating function of the asymptotic law of $m\tilde{W}$. This moment generating function is defined by:

$$F(z) = \lim_{d \rightarrow \infty} (E \circ tr) \left(\frac{1}{1 - zm\tilde{W}} \right)$$

We have the following result, regarding this moment generating function:

THEOREM 15.29. *The moment generating function of $m\tilde{W}$ satisfies the equation*

$$(F - 1)(1 - z^2 F^2) = mzF(1 + nzF)$$

in the $d \rightarrow \infty$ limit.

PROOF. We use the formula in Theorem 15.28. If we denote by $N(p, b, e)$ the number of partitions in $NC(p)$ having b blocks and e even blocks, we have:

$$\begin{aligned} F &= 1 + \sum_{p=1}^{\infty} \sum_{\pi \in NC(p)} z^p m^{|\pi|} n^{||\pi||} \\ &= 1 + \sum_{p=1}^{\infty} \sum_{b=0}^{\infty} \sum_{e=0}^{\infty} z^p m^b n^e N(p, b, e) \end{aligned}$$

Let us try to find a recurrence formula for the numbers $N(p, b, e)$. If we look at the block containing 1, this block must have $r \geq 0$ other legs, and we get:

$$\begin{aligned} N(p, b, e) &= \sum_{r \in 2\mathbb{N}} \sum_{p=\sum p_i+r+1} \sum_{b=\sum b_i+1} \sum_{e=\sum e_i} N(p_1, b_1, e_1) \dots N(p_{r+1}, b_{r+1}, e_{r+1}) \\ &+ \sum_{r \in 2\mathbb{N}+1} \sum_{p=\sum p_i+r+1} \sum_{b=\sum b_i+1} \sum_{e=\sum e_i+1} N(p_1, b_1, e_1) \dots N(p_{r+1}, b_{r+1}, e_{r+1}) \end{aligned}$$

Here p_1, \dots, p_{r+1} are the number of points between the legs of the block containing 1, so that we have $p = (p_1 + \dots + p_{r+1}) + r + 1$, and the whole sum is split over two cases, r even or odd, because the parity of r affects the number of even blocks of our partition. Now by multiplying everything by a $z^p m^b n^e$ factor, and by carefully distributing the various powers of z, m, b on the right, we obtain the following formula:

$$\begin{aligned} z^p m^b n^e N(p, b, e) &= m \sum_{r \in 2\mathbb{N}} z^{r+1} \sum_{p=\sum p_i+r+1} \sum_{b=\sum b_i+1} \sum_{e=\sum e_i} \prod_{i=1}^{r+1} z^{p_i} m^{b_i} n^{e_i} N(p_i, b_i, e_i) \\ &+ mn \sum_{r \in 2\mathbb{N}+1} z^{r+1} \sum_{p=\sum p_i+r+1} \sum_{b=\sum b_i+1} \sum_{e=\sum e_i+1} \prod_{i=1}^{r+1} z^{p_i} m^{b_i} n^{e_i} N(p_i, b_i, e_i) \end{aligned}$$

Let us sum now all these equalities, over all $p \geq 1$ and over all $b, e \geq 0$. According to the definition of F , at left we obtain $F - 1$. As for the two sums appearing on the right,

that is, at right of the two z^{r+1} factors, when summing them over all $p \geq 1$ and over all $b, e \geq 0$, we obtain in both cases F^{r+1} . So, we have the following formula:

$$\begin{aligned} F - 1 &= m \sum_{r \in 2\mathbb{N}} (zF)^{r+1} + mn \sum_{r \in 2\mathbb{N}+1} (zF)^{r+1} \\ &= m \frac{zF}{1 - z^2 F^2} + mn \frac{z^2 F^2}{1 - z^2 F^2} \\ &= mzF \frac{1 + nzF}{1 - z^2 F^2} \end{aligned}$$

But this gives the formula in the statement, and we are done. \square

We can reformulate Theorem 15.29 as follows:

THEOREM 15.30. *The Cauchy transform of $m\tilde{W}$ satisfies the equation*

$$(\xi G - 1)(1 - G^2) = mG(1 + nG)$$

in the $d \rightarrow \infty$ limit. Moreover, this equation simply reads

$$R = \frac{m}{2} \left(\frac{n+1}{1-z} - \frac{n-1}{1+z} \right)$$

with the substitutions $G \rightarrow z$ and $\xi \rightarrow R + z^{-1}$.

PROOF. We have two assertions to be proved, the idea being as follows:

(1) Consider the equation of F , found in Theorem 15.29, namely:

$$(F - 1)(1 - z^2 F^2) = mzF(1 + nzF)$$

With $z \rightarrow \xi^{-1}$ and $F \rightarrow \xi G$, so that $zF \rightarrow G$, we obtain, as desired:

$$(\xi G - 1)(1 - G^2) = mG(1 + nG)$$

(2) Thus, we have our equation for the Cauchy transform, and with this in hand, we can try to go ahead, and use somehow the Stieltjes inversion formula, in order to reach to a formula for the density. This is certainly possible, but our claim is that we can do better, by performing first some clever manipulations on the Cauchy transform.

(3) To be more precise, with $\xi \rightarrow K$ and $G \rightarrow z$, this equation becomes:

$$(zK - 1)(1 - z^2) = mz(1 + nz)$$

The point now is that with $K \rightarrow R + z^{-1}$ this latter equation becomes:

$$zR(1 - z^2) = mz(1 + nz)$$

But the solution of this latter equation is trivial to compute, given by:

$$R = m \frac{1 + nz}{1 - z^2} = \frac{m}{2} \left(\frac{n+1}{1-z} - \frac{n-1}{1+z} \right)$$

Thus, we are led to the conclusion in the statement. \square

All the above suggests the following definition:

DEFINITION 15.31. *Given a real probability measure μ , define its R -transform by:*

$$G_\mu(\xi) = \int_{\mathbb{R}} \frac{d\mu(t)}{\xi - t} \implies G_\mu \left(R_\mu(\xi) + \frac{1}{\xi} \right) = \xi$$

That is, the R -transform is the inverse of the Cauchy transform, up to a ξ^{-1} factor.

Getting back now to our questions, we would like to find the probability measure having as R -transform the function in Theorem 15.30. But here, we can only expect to find some kind of modification of the Marchenko-Pastur law, so as a first piece of work, let us just compute the R -transform of the Marchenko-Pastur law. We have here:

PROPOSITION 15.32. *The R -transform of the Marchenko-Pastur law π_t is*

$$R_{\pi_t}(\xi) = \frac{t}{1 - \xi}$$

for any $t > 0$.

PROOF. This can be done in two steps, as follows:

(1) At $t = 1$, we know that the moments of π_1 are the Catalan numbers, $M_k = C_k$, and we obtain that the Cauchy transform is given by the following formula:

$$G(\xi) = \frac{1}{2} - \frac{1}{2} \sqrt{1 - 4\xi^{-1}}$$

Now with $R(\xi) = \frac{1}{1-\xi}$ being the function in the statement, at $t = 1$, we have:

$$\begin{aligned} G \left(R(\xi) + \frac{1}{\xi} \right) &= G \left(\frac{1}{1-\xi} + \frac{1}{\xi} \right) \\ &= G \left(\frac{1}{\xi - \xi^2} \right) \\ &= \frac{1}{2} - \frac{1}{2} \sqrt{1 - 4\xi + 4\xi^2} \\ &= \frac{1}{2} - \frac{1}{2} (1 - 2\xi) \\ &= \xi \end{aligned}$$

Thus, the function $R(\xi) = \frac{1}{1-\xi}$ is indeed the R -transform of π_1 , in the above sense.

(2) In the general case, $t > 0$, the proof is similar, by using the moment formula for π_t , that we know from the above, and we will leave this as an exercise. \square

All this is very nice, and we can now further build on Theorem 15.30, as follows:

THEOREM 15.33. *The R -transform of $m\tilde{W}$ is given by*

$$R = R_{\pi_s} - R_{\pi_t}$$

in the $d \rightarrow \infty$ limit, where $s = m(n+1)/2$ and $t = m(n-1)/2$.

PROOF. We know from Theorem 15.30 that the R -transform of $m\tilde{W}$ is given by:

$$R = \frac{m}{2} \left(\frac{n+1}{1-z} - \frac{n-1}{1+z} \right)$$

By using now the formula in Proposition 15.32, this gives the result. \square

We can now formulate a final result, due to Aubrun, as follows:

THEOREM 15.34. *For a block-transposed Wishart matrix $\tilde{W} = (id \otimes t)W$ we have, in the $n = \beta m \rightarrow \infty$ limit, with $\beta > 0$ fixed, the formula*

$$\frac{\tilde{W}}{d} \sim \gamma_\beta^1$$

with γ_β^1 being the shifted version of the semicircle law γ_β , with support centered at 1.

PROOF. This follows from Theorem 15.33. Indeed, in the $n = \beta m \rightarrow \infty$ limit, with $\beta > 0$ fixed, we are led to the following formula for the Stieltjes transform:

$$f(x) = \frac{\sqrt{4\beta - (1-x)^2}}{2\beta\pi}$$

But this is the density of the shifted semicircle law having support as follows:

$$S = [1 - 2\sqrt{\beta}, 1 + 2\sqrt{\beta}]$$

Thus, we are led to the conclusion in the statement. \square

15e. Exercises

This was a quite exciting chapter, and as exercises on this, we have:

EXERCISE 15.35. *Compute the spectral measures of various matrices, of your choice.*

EXERCISE 15.36. *Learn more about C^* -algebras, including the GNS theorem.*

EXERCISE 15.37. *Importantly, learn as well about the von Neumann algebras.*

EXERCISE 15.38. *Learn as well some other approaches to the spectral measures.*

EXERCISE 15.39. *Is the asymptotic law of Gaussian matrices a “circular law”?*

EXERCISE 15.40. *Is the Wigner semicircle law some sort of “free Gaussian law”?*

EXERCISE 15.41. *Is the Marchenko-Pastur law some sort of “free Poisson law”?*

EXERCISE 15.42. *Read more, from Aubrun and others, about block modifications.*

As bonus exercise, learn some free probability theory, from [89].

CHAPTER 16

Free integration

16a. Free tori

Welcome to freeness. We will be interested here in developing free geometry and analysis, with the hope that all this might be related to physics, at very small scales, quarks and below. The idea being very simple, based on old findings of Heisenberg and others, if it is true indeed that the more you zoom down, the more commutativity disappears, then, logically, if you zoom hard enough, things will become free.

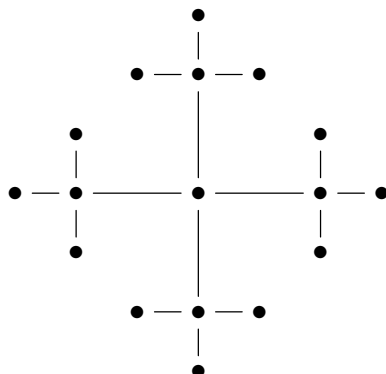
So, what is free? The simplest free object in mathematics is the free group F_N :

DEFINITION 16.1. *The free group F_N is the infinite group*

$$F_N = \langle g_1, \dots, g_N \mid \emptyset \rangle$$

generated by N variables g_1, \dots, g_N , with no relations between them.

This might look a bit abstract, but no worries, F_N has some interesting mathematics, coming right away, if you have some knowledge in discrete groups, and know how to look for interesting questions. For instance if you want to draw the Cayley graph of F_N , whose vertices are the elements of F_N , with edges $h - k$ drawn when $h = g_i^{\pm 1}k$ for some i , you will end up with an interesting picture, which at $N = 2$ looks like this:



And this type of graph certainly has interesting mathematics. One good question for instance is that of computing the number of length $2k$ loops based at the root. Another question, which is in fact equivalent, via moments, is that of computing the Kesten

measure of F_N , which is that of the following variable in the group algebra of F_N :

$$\chi = g_1 + \dots + g_N$$

All this looks very good, we most likely have here our first object of free geometry. In order now to formally understand this, let us recall the following formula, with $\mathbb{T}_N = \mathbb{T}^N$ being the usual torus, and with \mathbb{Z}^N being the free abelian group:

$$\mathbb{T}_N = \widehat{\mathbb{Z}^N}$$

Thus, getting back now to our free group F_N , which is the free analogue of \mathbb{Z}^N , it is in fact its dual $\widehat{F_N}$ which is a free manifold, and more specifically the free analogue of \mathbb{T}^N . Which is a nice finding, so let us formulate our conclusions as follows:

DEFINITION 16.2. *The free torus \mathbb{T}_N^+ is the dual of the free group F_N ,*

$$\mathbb{T}_N^+ = \widehat{F_N}$$

in analogy with the fact that the usual torus $\mathbb{T}_N = \mathbb{T}^N$ appears as

$$\mathbb{T}_N = \widehat{\mathbb{Z}^N}$$

with on the right the group \mathbb{Z}^N being the free abelian group.

It is of course possible to formulate things more precisely, and we will be back to this in a moment, but before that, isn't this a bit too abstract? But the point here is that no, at the level of questions to be solved, these remain the same, as for instance the computation of the Kesten measure, which is now a "function" on the free torus:

$$\chi \in C(\mathbb{T}_N^+)$$

In fact, this function is the main character of \mathbb{T}_N^+ , regarded as a compact quantum group, and so our Kesten problem suddenly becomes something very conceptual, namely the computation of the law of the main character of \mathbb{T}_N^+ . Which is very nice.

Before getting into details regarding all this, recall that \mathbb{R}^N is as interesting as \mathbb{C}^N . So, let us formulate as well the real version of Definition 16.2, as follows:

DEFINITION 16.3. *The free real torus, or free cube, T_N^+ is the dual*

$$T_N^+ = \widehat{L_N}$$

of the group $L_N = F_N / \langle g_i^2 = 1 \rangle$, in analogy with the fact that the usual cube is

$$T_N = \widehat{\mathbb{Z}_2^N}$$

with on the right the group \mathbb{Z}_2^N being the free real abelian group.

Here the “real” at the end stands for the fact that the generators must satisfy the real reflection condition $g^2 = 1$. As for the fact that “real torus = cube”, as stated, this needs some thinking, and in the hope that, after such thinking, you will agree with me that there is indeed a standard torus inside \mathbb{R}^N , and that is the unit cube.

Summarizing, all this sounds good, we have a beginning of free geometry, both real and complex, worth developing, by knowing at least what the torus of each theory is. In practice now, at the level of details, in order to talk about $\mathbb{T}_N^+ = \widehat{F}_N$ and $T_N^+ = \widehat{L}_N$ we need an extension of the usual Pontrjagin duality theory for the abelian groups, and this is best done via operator algebras, and the related notion of compact quantum group.

In order to understand all this, let us start with operator algebras. As explained in chapter 15, in relation with our random matrix needs there, the starting point is:

DEFINITION 16.4. *A C^* -algebra is a complex algebra A , having:*

- (1) *A norm $a \rightarrow \|a\|$, making it a Banach algebra.*
- (2) *An involution $a \rightarrow a^*$, satisfying $\|aa^*\| = \|a\|^2$.*

As basic examples, we have $B(H)$ itself, as well as any norm closed $*$ -subalgebra $A \subset B(H)$. It is possible to prove that any C^* -algebra appears in this way, but we will not need in what follows this deep result, called GNS theorem after Gelfand, Naimark, Segal. So, let us simply agree that, by definition, the C^* -algebras A are some sort of “generalized operator algebras”, and their elements $a \in A$ can be thought of as being some kind of “generalized operators”, on some Hilbert space which is not present.

In practice, this vague idea is all that we need. Indeed, by taking some inspiration from linear algebra, we can emulate spectral theory in our setting, as follows:

PROPOSITION 16.5. *Given $a \in A$, define its spectrum as being the set*

$$\sigma(a) = \left\{ \lambda \in \mathbb{C} \mid a - \lambda \notin A^{-1} \right\}$$

and its spectral radius $\rho(a)$ as the radius of the smallest centered disk containing $\sigma(a)$.

- (1) *The spectrum of a norm one element is in the unit disk.*
- (2) *The spectrum of a unitary element ($a^* = a^{-1}$) is on the unit circle.*
- (3) *The spectrum of a self-adjoint element ($a = a^*$) consists of real numbers.*
- (4) *The spectral radius of a normal element ($aa^* = a^*a$) is equal to its norm.*

PROOF. This is something that we know well from chapter 12 in the case $A \subset B(H)$, and from chapter 15 in general, and we refer to the material there for details. \square

Generally speaking, Proposition 16.5 is all you need for doing some further operator algebras. As a main application, that we already met in chapter 15, we have:

THEOREM 16.6 (Gelfand). *If X is a compact space, the algebra $C(X)$ of continuous functions $f : X \rightarrow \mathbb{C}$ is a commutative C^* -algebra, with structure as follows:*

- (1) *The norm is the usual sup norm, $\|f\| = \sup_{x \in X} |f(x)|$.*
- (2) *The involution is the usual involution, $f^*(x) = \overline{f(x)}$.*

Conversely, any commutative C^ -algebra is of the form $C(X)$, with its “spectrum” $X = \text{Spec}(A)$ appearing as the space of characters $\chi : A \rightarrow \mathbb{C}$.*

PROOF. Given a commutative C^* -algebra A , we can define indeed X to be the set of characters $\chi : A \rightarrow \mathbb{C}$, with the topology making continuous all the evaluation maps $ev_a : \chi \rightarrow \chi(a)$. Then X is a compact space, and $a \rightarrow ev_a$ is a morphism of algebras:

$$ev : A \rightarrow C(X)$$

We first prove that ev is involutive. We use the following formula:

$$a = \frac{a + a^*}{2} - i \cdot \frac{i(a - a^*)}{2}$$

Thus it is enough to prove the equality $ev_{a^*} = ev_a^*$ for self-adjoint elements a . But this is the same as proving that $a = a^*$ implies that ev_a is a real function, which is in turn true, because $ev_a(\chi) = \chi(a)$ is an element of $\sigma(a)$, contained in \mathbb{R} . So, claim proved. Also, since A is commutative, each element is normal, so ev is isometric:

$$\|ev_a\| = \rho(a) = \|a\|$$

It remains to prove that ev is surjective. But this follows from the Stone-Weierstrass theorem, because $ev(A)$ is a closed subalgebra of $C(X)$, which separates the points. \square

The Gelfand theorem suggests formulating the following definition:

DEFINITION 16.7. *Given a C^* -algebra A , not necessarily commutative, we write*

$$A = C(X)$$

and call the abstract object X a “compact quantum space”.

This might look quite revolutionary, but in practice, this definition changes nothing to what we have been doing so far, namely studying the C^* -algebras. So, we will keep studying the C^* -algebras, but by using the above fancy quantum space terminology. For instance whenever we have a morphism $\Phi : A \rightarrow B$, we will write $A = C(X)$, $B = C(Y)$, and rather speak of the corresponding morphism $\phi : Y \rightarrow X$. And so on.

Let us also mention that, at a closer look, there are in fact some problems with the above definition, coming from amenability and others. More on this later.

Now that we have our notion of quantum spaces, good time to get back towards Definitions 16.2 and 16.3. In order to understand what that free tori are, we will need:

THEOREM 16.8. *Let Γ be a discrete group, and consider the complex group algebra $\mathbb{C}[\Gamma]$, with involution given by the fact that all group elements are unitaries, $g^* = g^{-1}$.*

- (1) *The maximal C^* -seminorm on $\mathbb{C}[\Gamma]$ is a C^* -norm, and the closure of $\mathbb{C}[\Gamma]$ with respect to this norm is a C^* -algebra, denoted $C^*(\Gamma)$.*
- (2) *When Γ is abelian, we have an isomorphism $C^*(\Gamma) \simeq C(G)$, where $G = \widehat{\Gamma}$ is its Pontrjagin dual, formed by the characters $\chi : \Gamma \rightarrow \mathbb{T}$.*

PROOF. All this is very standard, the idea being as follows:

(1) In order to prove the result, we must find a $*$ -algebra embedding $\mathbb{C}[\Gamma] \subset B(H)$, with H being a Hilbert space. For this purpose, consider the space $H = l^2(\Gamma)$, having $\{h\}_{h \in \Gamma}$ as orthonormal basis. Our claim is that we have an embedding, as follows:

$$\pi : \mathbb{C}[\Gamma] \subset B(H) \quad , \quad \pi(g)(h) = gh$$

Indeed, since $\pi(g)$ maps the basis $\{h\}_{h \in \Gamma}$ into itself, this operator is well-defined, bounded, and is an isometry. It is also clear from the formula $\pi(g)(h) = gh$ that $g \rightarrow \pi(g)$ is a morphism of algebras, and since this morphism maps the unitaries $g \in \Gamma$ into isometries, this is a morphism of $*$ -algebras. Finally, the faithfulness of π is clear.

(2) Since Γ is abelian, the corresponding group algebra $A = C^*(\Gamma)$ is commutative. Thus, we can apply the Gelfand theorem, and we obtain $A = C(X)$, with:

$$X = \text{Spec}(A)$$

But the spectrum $X = \text{Spec}(A)$, consisting of the characters $\chi : C^*(\Gamma) \rightarrow \mathbb{C}$, can be identified with the Pontrjagin dual $G = \widehat{\Gamma}$, and this gives the result. \square

The above result suggests the following definition:

DEFINITION 16.9. *Given a discrete group Γ , the compact quantum space G given by*

$$C(G) = C^*(\Gamma)$$

is called abstract dual of Γ , and is denoted $G = \widehat{\Gamma}$.

Good news, this definition is exactly what we need, in order to understand the meaning of Definitions 16.2 and 16.3. To be more precise, we have the following result:

PROPOSITION 16.10. *The basic tori are all group duals, as follows,*

$$\begin{array}{ccc} T_N^+ & \longrightarrow & \mathbb{T}_N^+ \\ \uparrow & & \uparrow \\ T_N & \longrightarrow & \mathbb{T}_N \end{array} = \begin{array}{ccc} \widehat{L}_N & \longrightarrow & \widehat{F}_N \\ \uparrow & & \uparrow \\ \mathbb{Z}_2^N & \longrightarrow & \mathbb{T}^N \end{array}$$

where $F_N = \mathbb{Z}^{*N}$ is the free group on N generators, and $L_N = \mathbb{Z}_2^{*N}$ is its real version.

PROOF. The basic tori appear indeed as group duals, and together with the Fourier transform identifications from Theorem 16.8 (2), this gives the result. \square

Moving ahead, now that we have our formalism, we can start developing free geometry. As a first objective, we would like to better understand the relation between the classical and free tori. In order to discuss this, let us introduce the following notion:

DEFINITION 16.11. *Given a compact quantum space X , its classical version is the usual compact space $X_{class} \subset X$ obtained by dividing $C(X)$ by its commutator ideal:*

$$C(X_{class}) = C(X)/I \quad , \quad I = \langle [a, b] \rangle$$

In this situation, we also say that X appears as a “liberation” of X .

In other words, the space X_{class} appears as the Gelfand spectrum of the commutative C^* -algebra $C(X)/I$. Observe in particular that X_{class} is indeed a classical space.

In relation now with our tori, we have the following result:

THEOREM 16.12. *We have inclusions between the various tori, as follows,*

$$\begin{array}{ccc} T_N^+ & \longrightarrow & \mathbb{T}_N^+ \\ \uparrow & & \uparrow \\ T_N & \longrightarrow & \mathbb{T}_N \end{array}$$

and the free tori on top appear as liberations of the tori on the bottom.

PROOF. This is indeed clear from definitions, because commutativity of a group algebra means precisely that the group in question is abelian. \square

16b. Free spheres

In order to extend now the free geometries that we have, real and complex, let us begin with the spheres. We have the following notions:

DEFINITION 16.13. *We have free real and complex spheres, defined via*

$$C(S_{\mathbb{R},+}^{N-1}) = C^* \left(x_1, \dots, x_N \mid x_i = x_i^*, \sum_i x_i^2 = 1 \right)$$

$$C(S_{\mathbb{C},+}^{N-1}) = C^* \left(x_1, \dots, x_N \mid \sum_i x_i x_i^* = \sum_i x_i^* x_i = 1 \right)$$

where the symbol C^ stands for universal enveloping C^* -algebra.*

Here the fact that these algebras are indeed well-defined comes from the following estimate, which shows that the biggest C^* -norms on these $*$ -algebras are bounded:

$$\|x_i\|^2 = \|x_i x_i^*\| \leq \left\| \sum_i x_i x_i^* \right\| = 1$$

As a first result now, regarding the above free spheres, we have:

THEOREM 16.14. *We have embeddings of compact quantum spaces, as follows,*

$$\begin{array}{ccc} S_{\mathbb{R},+}^{N-1} & \longrightarrow & S_{\mathbb{C},+}^{N-1} \\ \uparrow & & \uparrow \\ S_{\mathbb{R}}^{N-1} & \longrightarrow & S_{\mathbb{C}}^{N-1} \end{array}$$

and the spaces on top appear as liberations of the spaces on the bottom.

PROOF. The first assertion, regarding the inclusions, comes from the fact that at the level of the associated C^* -algebras, we have surjective maps, as follows:

$$\begin{array}{ccc} C(S_{\mathbb{R},+}^{N-1}) & \longleftarrow & C(S_{\mathbb{C},+}^{N-1}) \\ \downarrow & & \downarrow \\ C(S_{\mathbb{R}}^{N-1}) & \longleftarrow & C(S_{\mathbb{C}}^{N-1}) \end{array}$$

For the second assertion, we must establish the following isomorphisms, where the symbol C_{comm}^* stands for “universal commutative C^* -algebra generated by”:

$$\begin{aligned} C(S_{\mathbb{R}}^{N-1}) &= C_{comm}^* \left(x_1, \dots, x_N \mid x_i = x_i^*, \sum_i x_i^2 = 1 \right) \\ C(S_{\mathbb{C}}^{N-1}) &= C_{comm}^* \left(x_1, \dots, x_N \mid \sum_i x_i x_i^* = \sum_i x_i^* x_i = 1 \right) \end{aligned}$$

It is enough to establish the second isomorphism. So, consider the second universal commutative C^* -algebra A constructed above. Since the standard coordinates on $S_{\mathbb{C}}^{N-1}$ satisfy the defining relations for A , we have a quotient map of as follows:

$$A \rightarrow C(S_{\mathbb{C}}^{N-1})$$

Conversely, let us write $A = C(S)$, by using the Gelfand theorem. Then x_1, \dots, x_N become in this way true coordinates, providing us with an embedding as follows:

$$S \subset \mathbb{C}^N$$

Also, the quadratic relations become $\sum_i |x_i|^2 = 1$, so we have $S \subset S_{\mathbb{C}}^{N-1}$. Thus, we have a quotient map $C(S_{\mathbb{C}}^{N-1}) \rightarrow A$, as desired, and this gives all the results. \square

By using the free spheres constructed above, we can now formulate:

DEFINITION 16.15. *A real algebraic manifold $X \subset S_{\mathbb{C},+}^{N-1}$ is a closed quantum subspace defined, at the level of the corresponding C^* -algebra, by a formula of type*

$$C(X) = C(S_{\mathbb{C},+}^{N-1}) / \langle f_i(x_1, \dots, x_N) = 0 \rangle$$

for certain family of noncommutative polynomials, as follows:

$$f_i \in \mathbb{C} \langle x_1, \dots, x_N \rangle$$

We denote by $\mathcal{C}(X)$ the $*$ -subalgebra of $C(X)$ generated by the coordinates x_1, \dots, x_N .

As a basic example here, we have the free real sphere $S_{\mathbb{R},+}^{N-1}$. The classical spheres $S_{\mathbb{C}}^{N-1}$, $S_{\mathbb{R}}^{N-1}$, and their real submanifolds, are covered as well by this formalism. At the level of the general theory, we have the following version of the Gelfand theorem:

THEOREM 16.16. *If $X \subset S_{\mathbb{C},+}^{N-1}$ is an algebraic manifold, as above, we have*

$$X_{class} = \left\{ x \in S_{\mathbb{C}}^{N-1} \mid f_i(x_1, \dots, x_N) = 0 \right\}$$

and X appears as a liberation of X_{class} .

PROOF. This is something that we already met, in the context of the free spheres. In general, the proof is similar, by using the Gelfand theorem. Indeed, if we denote by X'_{class} the manifold constructed in the statement, then we have a quotient map of C^* -algebras as follows, mapping standard coordinates to standard coordinates:

$$C(X_{class}) \rightarrow C(X'_{class})$$

Conversely now, from $X \subset S_{\mathbb{C},+}^{N-1}$ we obtain $X_{class} \subset S_{\mathbb{C}}^{N-1}$. Now since the relations defining X'_{class} are satisfied by X_{class} , we obtain an inclusion $X_{class} \subset X'_{class}$. Thus, at the level of algebras of continuous functions, we have a quotient map of C^* -algebras as follows, mapping standard coordinates to standard coordinates:

$$C(X'_{class}) \rightarrow C(X_{class})$$

Thus, we have constructed a pair of inverse morphisms, and we are done. \square

Finally, once again at the level of the general theory, we have:

DEFINITION 16.17. *We agree to identify two real algebraic submanifolds $X, Y \subset S_{\mathbb{C},+}^{N-1}$ when we have a $*$ -algebra isomorphism between $*$ -algebras of coordinates*

$$f : \mathcal{C}(Y) \rightarrow \mathcal{C}(X)$$

mapping standard coordinates to standard coordinates.

We will see later the reasons for making this convention, coming from amenability. Now back to the tori, as constructed before, we can see that these are examples of algebraic manifolds, in the sense of Definition 16.15. In fact, we have the following result:

THEOREM 16.18. *The four main quantum spheres produce the main quantum tori*

$$\begin{array}{ccc}
 S_{\mathbb{R},+}^{N-1} & \longrightarrow & S_{\mathbb{C},+}^{N-1} \\
 \uparrow & & \uparrow \\
 S_{\mathbb{R}}^{N-1} & \longrightarrow & S_{\mathbb{C}}^{N-1}
 \end{array}
 \quad \longrightarrow \quad
 \begin{array}{ccc}
 T_N^+ & \longrightarrow & \mathbb{T}_N^+ \\
 \uparrow & & \uparrow \\
 T_N & \longrightarrow & \mathbb{T}_N
 \end{array}$$

via the formula $T = S \cap \mathbb{T}_N^+$, with the intersection being taken inside $S_{\mathbb{C},+}^{N-1}$.

PROOF. This comes from the above results, the situation being as follows:

(1) Free complex case. Here the formula in the statement reads:

$$\mathbb{T}_N^+ = S_{\mathbb{C},+}^{N-1} \cap \mathbb{T}_N^+$$

But this is something trivial, because we have $\mathbb{T}_N^+ \subset S_{\mathbb{C},+}^{N-1}$.

(2) Free real case. Here the formula in the statement reads:

$$T_N^+ = S_{\mathbb{R},+}^{N-1} \cap \mathbb{T}_N^+$$

But this is clear as well, the real version of \mathbb{T}_N^+ being T_N^+ .

(3) Classical complex case. Here the formula in the statement reads:

$$\mathbb{T}_N = S_{\mathbb{C}}^{N-1} \cap \mathbb{T}_N^+$$

But this is clear as well, the classical version of \mathbb{T}_N^+ being \mathbb{T}_N .

(4) Classical real case. Here the formula in the statement reads:

$$T_N = S_{\mathbb{R}}^{N-1} \cap \mathbb{T}_N^+$$

But this follows by intersecting the formulae from the proof of (2) and (3). \square

16c. Free rotations

In order to better understand the structure of the free spheres $S_{\mathbb{R},+}^{N-1}, S_{\mathbb{C},+}^{N-1}$, we need to talk about free rotations. Following Woronowicz, let us start with:

DEFINITION 16.19. *A Woronowicz algebra is a C^* -algebra A , given with a unitary matrix $u \in M_N(A)$ whose coefficients generate A , such that the formulae*

$$\Delta(u_{ij}) = \sum_k u_{ik} \otimes u_{kj} \quad , \quad \varepsilon(u_{ij}) = \delta_{ij} \quad , \quad S(u_{ij}) = u_{ji}^*$$

define morphisms of C^* -algebras $\Delta : A \rightarrow A \otimes A$, $\varepsilon : A \rightarrow \mathbb{C}$, $S : A \rightarrow A^{opp}$.

We say that A is cocommutative when $\Sigma\Delta = \Delta$, where $\Sigma(a \otimes b) = b \otimes a$ is the flip. We have the following result, which justifies the terminology and axioms:

THEOREM 16.20. *The following are Woronowicz algebras:*

(1) $C(G)$, with $G \subset U_N$ compact Lie group. Here the structural maps are:

$$\Delta(\varphi) = (g, h) \rightarrow \varphi(gh) \quad , \quad \varepsilon(\varphi) = \varphi(1) \quad , \quad S(\varphi) = g \rightarrow \varphi(g^{-1})$$

(2) $C^*(\Gamma)$, with $F_N \rightarrow \Gamma$ finitely generated group. Here the structural maps are:

$$\Delta(g) = g \otimes g \quad , \quad \varepsilon(g) = 1 \quad , \quad S(g) = g^{-1}$$

Moreover, we obtain in this way all the commutative/cocommutative algebras.

PROOF. This is something very standard, the idea being as follows:

(1) Given $G \subset U_N$, we can set $A = C(G)$, which is a Woronowicz algebra, together with the matrix $u = (u_{ij})$ formed by coordinates of G , given by:

$$g = \begin{pmatrix} u_{11}(g) & \dots & u_{1N}(g) \\ \vdots & & \vdots \\ u_{N1}(g) & \dots & u_{NN}(g) \end{pmatrix}$$

Conversely, if (A, u) is a commutative Woronowicz algebra, by using the Gelfand theorem we can write $A = C(X)$, with X being a certain compact space. The coordinates u_{ij} give then an embedding $X \subset M_N(\mathbb{C})$, and since the matrix $u = (u_{ij})$ is unitary we actually obtain an embedding $X \subset U_N$, and finally by using the maps Δ, ε, S we conclude that our compact subspace $X \subset U_N$ is in fact a compact Lie group, as desired.

(2) Consider a finitely generated group $F_N \rightarrow \Gamma$. We can set $A = C^*(\Gamma)$, which is by definition the completion of the complex group algebra $\mathbb{C}[\Gamma]$, with involution given by $g^* = g^{-1}$, for any $g \in \Gamma$, with respect to the biggest C^* -norm, and we obtain a Woronowicz algebra, together with the diagonal matrix formed by the generators of Γ :

$$u = \begin{pmatrix} g_1 & & 0 \\ & \ddots & \\ 0 & & g_N \end{pmatrix}$$

Conversely, if (A, u) is a cocommutative Woronowicz algebra, the Peter-Weyl theory of Woronowicz, to be explained below, shows that the irreducible corepresentations of A are all 1-dimensional, and form a group Γ , and so we have $A = C^*(\Gamma)$, as desired. \square

The above result makes it quite clear that what we have in Definition 16.19 is some sort of joint definition for the compact and discrete quantum groups. In order to further comment on this, let us go back to the formulae in Definition 16.19, namely:

$$\Delta(u_{ij}) = \sum_k u_{ik} \otimes u_{kj} \quad , \quad \varepsilon(u_{ij}) = \delta_{ij} \quad , \quad S(u_{ij}) = u_{ji}^*$$

The morphisms Δ, ε, S are called comultiplication, counit and antipode, and they have the following properties, which are something very familiar in abstract algebra:

THEOREM 16.21. *Let (A, u) be a Woronowicz algebra.*

(1) Δ, ε satisfy the usual axioms for a comultiplication and a counit, namely:

$$\begin{aligned} (\Delta \otimes id)\Delta &= (id \otimes \Delta)\Delta \\ (\varepsilon \otimes id)\Delta &= (id \otimes \varepsilon)\Delta = id \end{aligned}$$

(2) S satisfies the antipode axiom, on the $*$ -subalgebra generated by entries of u :

$$m(S \otimes id)\Delta = m(id \otimes S)\Delta = \varepsilon(\cdot)1$$

(3) In addition, the square of the antipode is the identity, $S^2 = id$.

PROOF. Observe first that the result holds indeed in the case where the algebra A is commutative. Indeed, by using Theorem 16.20 we can write:

$$\Delta = m^t, \quad \varepsilon = u^t, \quad S = i^t$$

The above 3 conditions come then by transposition from the basic 3 group theory conditions satisfied by m, u, i , which are as follows, with $\delta(g) = (g, g)$:

$$\begin{aligned} m(m \times id) &= m(id \times m) \\ m(id \times u) &= m(u \times id) = id \\ m(id \times i)\delta &= m(i \times id)\delta = 1 \end{aligned}$$

Observe that $S^2 = id$ is satisfied as well, coming from $i^2 = id$, which is a consequence of the group axioms. Finally, observe that the result holds also, trivially, when A is cocommutative, again by Theorem 16.20. In general now, the proof goes as follows:

(1) We have indeed the following computation:

$$(\Delta \otimes id)\Delta(u_{ij}) = \sum_l \Delta(u_{il}) \otimes u_{lj} = \sum_{kl} u_{ik} \otimes u_{kl} \otimes u_{lj}$$

On the other hand, we have as well the following computation:

$$(id \otimes \Delta)\Delta(u_{ij}) = \sum_k u_{ik} \otimes \Delta(u_{kj}) = \sum_{kl} u_{ik} \otimes u_{kl} \otimes u_{lj}$$

The proof for the counit axiom is quite similar. We first have:

$$(id \otimes \varepsilon)\Delta(u_{ij}) = \sum_k u_{ik} \otimes \varepsilon(u_{kj}) = u_{ij}$$

On the other hand, we have as well the following computation:

$$(\varepsilon \otimes id)\Delta(u_{ij}) = \sum_k \varepsilon(u_{ik}) \otimes u_{kj} = u_{ij}$$

(2) By using the fact that the matrix $u = (u_{ij})$ is unitary, we obtain:

$$\begin{aligned} m(id \otimes S)\Delta(u_{ij}) &= \sum_k u_{ik}S(u_{kj}) \\ &= \sum_k u_{ik}u_{jk}^* \\ &= (uu^*)_{ij} \\ &= \delta_{ij} \end{aligned}$$

Similarly, we have the following computation, which gives the antipode axiom:

$$\begin{aligned} m(S \otimes id)\Delta(u_{ij}) &= \sum_k S(u_{ik})u_{kj} \\ &= \sum_k u_{ki}^*u_{kj} \\ &= (u^*u)_{ij} \\ &= \delta_{ij} \end{aligned}$$

(3) Finally, the formula $S^2 = id$ holds as well on the generators, and we are done. \square

Summarizing, the Woronowicz algebras appear to have nice properties. In view of Theorem 16.20 and Theorem 16.21, we can formulate the following definition:

DEFINITION 16.22. *Given a Woronowicz algebra A , we formally write*

$$A = C(G) = C^*(\Gamma)$$

and call G compact quantum group, and Γ discrete quantum group.

In relation with this, there are actually some analytic subtleties, coming from amenability, so our objects must be divided by a certain equivalence relation, for everything to work fine. To be more precise, we agree to write $(A, u) = (B, v)$ when there is a $*$ -algebra isomorphism as follows, mapping standard coordinates to standard coordinates:

$$\langle u_{ij} \rangle \simeq \langle v_{ij} \rangle \quad , \quad u_{ij} \rightarrow v_{ij}$$

Moving ahead now, let us call now corepresentation of A any unitary matrix $v \in M_n(A)$ satisfying the same conditions as those satisfied by u , namely:

$$\Delta(v_{ij}) = \sum_k v_{ik} \otimes v_{kj} \quad , \quad \varepsilon(v_{ij}) = \delta_{ij} \quad , \quad S(v_{ij}) = v_{ji}^*$$

These corepresentations can be thought of as corresponding representations of the underlying compact quantum group G . Following Woronowicz, we have:

THEOREM 16.23. *Any Woronowicz algebra has a unique Haar integration functional,*

$$\left(\int_G \otimes id\right) \Delta = \left(id \otimes \int_G\right) \Delta = \int_G (.)1$$

which can be constructed by starting with any faithful positive form $\varphi \in A^$, and setting*

$$\int_G = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n \varphi^{*k}$$

*where $\phi * \psi = (\phi \otimes \psi)\Delta$. Moreover, for any corepresentation $v \in M_n(\mathbb{C}) \otimes A$ we have*

$$\left(id \otimes \int_G\right) v = P$$

where P is the orthogonal projection onto $Fix(v) = \{\xi \in \mathbb{C}^n | v\xi = \xi\}$.

PROOF. This can be done in 3 steps, as follows:

(1) Given $\varphi \in A^*$, our claim is that the following limit converges, for any $a \in A$:

$$\int_\varphi a = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n \varphi^{*k}(a)$$

Indeed, by linearity we can assume that a is the coefficient of corepresentation, $a = (\tau \otimes id)v$. But in this case, an elementary computation shows that we have the following formula, where P_φ is the orthogonal projection onto the 1-eigenspace of $(id \otimes \varphi)v$:

$$\left(id \otimes \int_\varphi\right) v = P_\varphi$$

(2) Since $v\xi = \xi$ implies $[(id \otimes \varphi)v]\xi = \xi$, we have $P_\varphi \geq P$, where P is the orthogonal projection onto the space $Fix(v) = \{\xi \in \mathbb{C}^n | v\xi = \xi\}$. The point now is that when $\varphi \in A^*$ is faithful, by using a positivity trick, one can prove that we have $P_\varphi = P$. Thus our linear form \int_φ is independent of φ , and is given on coefficients $a = (\tau \otimes id)v$ by:

$$\left(id \otimes \int_\varphi\right) v = P$$

(3) With the above formula in hand, the left and right invariance of $\int_G = \int_\varphi$ is clear on coefficients, and so in general, and this gives all the assertions. \square

Consider now the dense $*$ -subalgebra $\mathcal{A} \subset A$ generated by the coefficients of u , that we met after Definition 16.22, and endow it with the following scalar product:

$$\langle a, b \rangle = \int_G ab^*$$

Still following Woronowicz, we have the following key result:

THEOREM 16.24. *We have the following Peter-Weyl type results:*

- (1) *Any corepresentation decomposes as a sum of irreducible corepresentations.*
- (2) *Each irreducible corepresentation appears inside a certain $u^{\otimes k}$.*
- (3) $\mathcal{A} = \bigoplus_{v \in \text{Irr}(A)} M_{\dim(v)}(\mathbb{C})$, *the summands being pairwise orthogonal.*
- (4) *The characters of irreducible corepresentations form an orthonormal system.*

PROOF. All these results are very standard, the idea being as follows:

(1) Given $v \in M_n(A)$, its intertwiner algebra $\text{End}(v) = \{T \in M_n(\mathbb{C}) | Tv = vT\}$ is a finite dimensional C^* -algebra, and so decomposes as $\text{End}(v) = M_{n_1}(\mathbb{C}) \oplus \dots \oplus M_{n_r}(\mathbb{C})$. But this gives a decomposition of type $v = v_1 + \dots + v_r$, as desired.

(2) Consider indeed the Peter-Weyl corepresentations, $u^{\otimes k}$ with k colored integer, defined by $u^{\otimes \emptyset} = 1$, $u^{\otimes \circ} = u$, $u^{\otimes \bullet} = \bar{u}$ and multiplicativity. The coefficients of these corepresentations span the dense algebra \mathcal{A} , and by using (1), this gives the result.

(3) Here the direct sum decomposition, which is technically a $*$ -coalgebra isomorphism, follows from (2). As for the second assertion, this follows from the fact that $(\text{id} \otimes \int_G)v$ is the orthogonal projection P_v onto the space $\text{Fix}(v)$, for any corepresentation v .

(4) Let us define indeed the character of $v \in M_n(A)$ to be the matrix trace, $\chi_v = \text{Tr}(v)$. Since this character is a coefficient of v , the orthogonality assertion follows from (3). As for the norm 1 claim, this follows once again from $(\text{id} \otimes \int_G)v = P_v$. \square

Good news, we can now talk about free rotations. Following Wang, we have:

THEOREM 16.25. *The following universal algebras are Woronowicz algebras,*

$$C(O_N^+) = C^* \left((u_{ij})_{i,j=1,\dots,N} \middle| u = \bar{u}, u^t = u^{-1} \right)$$

$$C(U_N^+) = C^* \left((u_{ij})_{i,j=1,\dots,N} \middle| u^* = u^{-1}, u^t = \bar{u}^{-1} \right)$$

so the underlying spaces O_N^+, U_N^+ are compact quantum groups.

PROOF. This follows from the elementary fact that if a matrix $u = (u_{ij})$ is orthogonal or biunitary, then so must be the following matrices:

$$u_{ij}^\Delta = \sum_k u_{ik} \otimes u_{kj} \quad , \quad u_{ij}^\varepsilon = \delta_{ij} \quad , \quad u_{ij}^S = u_{ji}^*$$

Thus, we can indeed define morphisms Δ, ε, S as in Definition 16.19, by using the universal properties of $C(O_N^+)$, $C(U_N^+)$, and this gives the result. \square

In order to discuss now the relation with the spheres, which can only come via some sort of “isometric actions”, let us start with the following standard fact:

PROPOSITION 16.26. *Given a closed subset $X \subset S_{\mathbb{C}}^{N-1}$, the formula*

$$G(X) = \left\{ U \in U_N \mid U(X) = X \right\}$$

defines a compact group of unitary matrices, or isometries, called affine isometry group of X . For the spheres $S_{\mathbb{R}}^{N-1}, S_{\mathbb{C}}^{N-1}$ we obtain in this way the groups O_N, U_N .

PROOF. The fact that $G(X)$ as defined above is indeed a group is clear, its compactness is clear as well, and finally the last assertion is clear as well. In fact, all this works for any closed subset $X \subset \mathbb{C}^N$, but we are not interested here in such general spaces. \square

Observe that in the case of the real and complex spheres, the affine isometry group $G(X)$ leaves invariant the Riemannian metric, because this metric is equivalent to the one inherited from \mathbb{C}^N , which is preserved by our isometries $U \in U_N$.

Thus, we could have constructed as well $G(X)$ as being the group of metric isometries of X , with of course some extra care in relation with the complex structure, as for the complex sphere $X = S_{\mathbb{C}}^{N-1}$ to produce $G(X) = U_N$ instead of $G(X) = O_{2N}$. But, such things won't really work for the free spheres, and so are to be avoided.

The point now is that we have the following quantum analogue of Proposition 16.26, which is a perfect analogue, save for the fact that X is now assumed to be algebraic, for some technical reasons, which allows us to talk about quantum isometry groups:

THEOREM 16.27. *Given an algebraic manifold $X \subset S_{\mathbb{C},+}^{N-1}$, the category of the closed subgroups $G \subset U_N^+$ acting affinely on X , in the sense that the formula*

$$\Phi(x_i) = \sum_j x_j \otimes u_{ji}$$

defines a morphism of C^ -algebras $\Phi : C(X) \rightarrow C(X) \otimes C(G)$, has a universal object, denoted $G^+(X)$, and called affine quantum isometry group of X .*

PROOF. Assume indeed that our manifold $X \subset S_{\mathbb{C},+}^{N-1}$ comes as follows:

$$C(X) = C(S_{\mathbb{C},+}^{N-1}) / \left\langle f_{\alpha}(x_1, \dots, x_N) = 0 \right\rangle$$

In order to prove the result, consider the following variables:

$$X_i = \sum_j x_j \otimes u_{ji} \in C(X) \otimes C(U_N^+)$$

Our claim is that the quantum group in the statement $G = G^+(X)$ appears as:

$$C(G) = C(U_N^+) / \left\langle f_{\alpha}(X_1, \dots, X_N) = 0 \right\rangle$$

In order to prove this, pick one of the defining polynomials, and write it as follows:

$$f_\alpha(x_1, \dots, x_N) = \sum_r \sum_{i_1^r \dots i_{s_r}^r} \lambda_r \cdot x_{i_1^r} \dots x_{i_{s_r}^r}$$

With $X_i = \sum_j x_j \otimes u_{ji}$ as above, we have the following formula:

$$f_\alpha(X_1, \dots, X_N) = \sum_r \sum_{i_1^r \dots i_{s_r}^r} \lambda_r \sum_{j_1^r \dots j_{s_r}^r} x_{j_1^r} \dots x_{j_{s_r}^r} \otimes u_{j_1^r i_1^r} \dots u_{j_{s_r}^r i_{s_r}^r}$$

Since the variables on the right span a certain finite dimensional space, the relations $f_\alpha(X_1, \dots, X_N) = 0$ correspond to certain relations between the variables u_{ij} . Thus, we have indeed a closed subspace $G \subset U_N^+$, with a universal map, as follows:

$$\Phi : C(X) \rightarrow C(X) \otimes C(G)$$

In order to show now that G is a quantum group, consider the following elements:

$$u_{ij}^\Delta = \sum_k u_{ik} \otimes u_{kj} \quad , \quad u_{ij}^\varepsilon = \delta_{ij} \quad , \quad u_{ij}^S = u_{ji}^*$$

Consider as well the following elements, with $\gamma \in \{\Delta, \varepsilon, S\}$:

$$X_i^\gamma = \sum_j x_j \otimes u_{ji}^\gamma$$

From the relations $f_\alpha(X_1, \dots, X_N) = 0$ we deduce that we have:

$$f_\alpha(X_1^\gamma, \dots, X_N^\gamma) = (id \otimes \gamma) f_\alpha(X_1, \dots, X_N) = 0$$

Thus we can map $u_{ij} \rightarrow u_{ij}^\gamma$ for any $\gamma \in \{\Delta, \varepsilon, S\}$, and we are done. \square

We can now formulate a result about spheres and rotations, as follows:

THEOREM 16.28. *The quantum isometry groups of the basic spheres are*

$$\begin{array}{ccc} S_{\mathbb{R},+}^{N-1} & \longrightarrow & S_{\mathbb{C},+}^{N-1} \\ \uparrow & & \uparrow \\ S_{\mathbb{R}}^{N-1} & \longrightarrow & S_{\mathbb{C}}^{N-1} \end{array} \quad \rightarrow \quad \begin{array}{ccc} O_N^+ & \longrightarrow & U_N^+ \\ \uparrow & & \uparrow \\ O_N & \longrightarrow & U_N \end{array}$$

modulo identifying, as usual, the various C^ -algebraic completions.*

PROOF. We have 4 results to be proved, the idea being as follows:

$S_{\mathbb{C},+}^{N-1}$. Let us first construct an action $U_N^+ \curvearrowright S_{\mathbb{C},+}^{N-1}$. We must prove here that the variables $X_i = \sum_j x_j \otimes u_{ji}$ satisfy the defining relations for $S_{\mathbb{C},+}^{N-1}$, namely:

$$\sum_i x_i x_i^* = \sum_i x_i^* x_i = 1$$

By using the biunitarity of u , we have the following computation:

$$\sum_i X_i X_i^* = \sum_{ijk} x_j x_k^* \otimes u_{ji} u_{ki}^* = \sum_j x_j x_j^* \otimes 1 = 1 \otimes 1$$

Once again by using the biunitarity of u , we have as well:

$$\sum_i X_i^* X_i = \sum_{ijk} x_j^* x_k \otimes u_{ji}^* u_{ki} = \sum_j x_j^* x_j \otimes 1 = 1 \otimes 1$$

Thus we have an action $U_N^+ \curvearrowright S_{\mathbb{C},+}^{N-1}$, which gives $G^+(S_{\mathbb{C},+}^{N-1}) = U_N^+$, as desired.

$S_{\mathbb{R},+}^{N-1}$. Let us first construct an action $O_N^+ \curvearrowright S_{\mathbb{R},+}^{N-1}$. We already know that the variables $X_i = \sum_j x_j \otimes u_{ji}$ satisfy the defining relations for $S_{\mathbb{C},+}^{N-1}$, so we just have to check that these variables are self-adjoint. But this is clear from $u = \bar{u}$, as follows:

$$X_i^* = \sum_j x_j^* \otimes u_{ji}^* = \sum_j x_j \otimes u_{ji} = X_i$$

Conversely, assume that we have an action $G \curvearrowright S_{\mathbb{R},+}^{N-1}$, with $G \subset U_N^+$. The variables $X_i = \sum_j x_j \otimes u_{ji}$ must be then self-adjoint, and the above computation shows that we must have $u = \bar{u}$. Thus our quantum group must satisfy $G \subset O_N^+$, as desired.

$S_{\mathbb{C}}^{N-1}$. The fact that we have an action $U_N \curvearrowright S_{\mathbb{C}}^{N-1}$ is clear. Conversely, assume that we have an action $G \curvearrowright S_{\mathbb{C}}^{N-1}$, with $G \subset U_N^+$. We must prove that this implies $G \subset U_N$, and we will use a standard trick of Bhowmick-Goswami. We have:

$$\Phi(x_i) = \sum_j x_j \otimes u_{ji}$$

By multiplying this formula with itself we obtain:

$$\Phi(x_i x_k) = \sum_{jl} x_j x_l \otimes u_{ji} u_{lk}$$

$$\Phi(x_k x_i) = \sum_{jl} x_l x_j \otimes u_{lk} u_{ji}$$

Since the variables x_i commute, these formulae can be written as:

$$\Phi(x_i x_k) = \sum_{j < l} x_j x_l \otimes (u_{ji} u_{lk} + u_{li} u_{jk}) + \sum_j x_j^2 \otimes u_{ji} u_{jk}$$

$$\Phi(x_i x_k) = \sum_{j < l} x_j x_l \otimes (u_{lk} u_{ji} + u_{jk} u_{li}) + \sum_j x_j^2 \otimes u_{jk} u_{ji}$$

Since the tensors at left are linearly independent, we must have:

$$u_{ji} u_{lk} + u_{li} u_{jk} = u_{lk} u_{ji} + u_{jk} u_{li}$$

By applying the antipode to this formula, then applying the involution, and then relabelling the indices, we successively obtain:

$$\begin{aligned} u_{kl}^* u_{ij}^* + u_{kj}^* u_{il}^* &= u_{ij}^* u_{kl}^* + u_{il}^* u_{kj}^* \\ u_{ij} u_{kl} + u_{il} u_{kj} &= u_{kl} u_{ij} + u_{kj} u_{il} \\ u_{ji} u_{lk} + u_{jk} u_{li} &= u_{lk} u_{ji} + u_{li} u_{jk} \end{aligned}$$

Now by comparing with the original formula, we obtain from this:

$$u_{li} u_{jk} = u_{jk} u_{li}$$

In order to finish, it remains to prove that the coordinates u_{ij} commute as well with their adjoints. For this purpose, we use a similar method. We have:

$$\begin{aligned} \Phi(x_i x_k^*) &= \sum_{jl} x_j x_l^* \otimes u_{ji} u_{lk}^* \\ \Phi(x_k^* x_i) &= \sum_{jl} x_l^* x_j \otimes u_{lk}^* u_{ji} \end{aligned}$$

Since the variables on the left are equal, we deduce from this that we have:

$$\sum_{jl} x_j x_l^* \otimes u_{ji} u_{lk}^* = \sum_{jl} x_j x_l^* \otimes u_{lk}^* u_{ji}$$

Thus we have $u_{ji} u_{lk}^* = u_{lk}^* u_{ji}$, and so $G \subset U_N$, as claimed.

$S_{\mathbb{R}}^{N-1}$. The fact that we have an action $O_N \curvearrowright S_{\mathbb{R}}^{N-1}$ is clear. In what regards the converse, this follows by combining the results that we already have, as follows:

$$\begin{aligned} G \curvearrowright S_{\mathbb{R}}^{N-1} &\implies G \curvearrowright S_{\mathbb{R},+}^{N-1}, S_{\mathbb{C}}^{N-1} \\ &\implies G \subset O_N^+, U_N \\ &\implies G \subset O_N^+ \cap U_N = O_N \end{aligned}$$

Thus, we conclude that we have $G^+(S_{\mathbb{R}}^{N-1}) = O_N$, as desired. \square

16d. Fine structure

Let us discuss now the correspondence $U \rightarrow S$. In the classical case the situation is very simple, because the sphere $S = S^{N-1}$ appears by rotating the point $x = (1, 0, \dots, 0)$ by the isometries in $U = U_N$. Moreover, the stabilizer of this action is the subgroup $U_{N-1} \subset U_N$ acting on the last $N-1$ coordinates, and so the sphere $S = S^{N-1}$ appears from the corresponding rotation group $U = U_N$ as an homogeneous space, as follows:

$$S^{N-1} = U_N / U_{N-1}$$

In functional analytic terms, all this becomes even simpler, the correspondence $U \rightarrow S$ being obtained, at the level of algebras of functions, as follows:

$$C(S^{N-1}) \subset C(U_N) \quad , \quad x_i \rightarrow u_{1i}$$

In general now, the straightforward homogeneous space interpretation of S as above fails. However, we can have some theory going by using the functional analytic viewpoint, with an embedding $x_i \rightarrow u_{1i}$ as above. Let us start with the following result:

PROPOSITION 16.29. *For the basic spheres, we have a diagram as follows,*

$$\begin{array}{ccc} C(S) & \xrightarrow{\Phi} & C(S) \otimes C(U) \\ \downarrow \alpha & & \downarrow \alpha \otimes id \\ C(U) & \xrightarrow{\Delta} & C(U) \otimes C(U) \end{array}$$

where on top $\Phi(x_i) = \sum_j x_j \otimes u_{ji}$, and on the left $\alpha(x_i) = u_{1i}$.

PROOF. The diagram in the statement commutes indeed on the standard coordinates, the corresponding arrows being as follows, on these coordinates:

$$\begin{array}{ccc} x_i & \longrightarrow & \sum_j x_j \otimes u_{ji} \\ \downarrow & & \downarrow \\ u_{1i} & \longrightarrow & \sum_j u_{1j} \otimes u_{ji} \end{array}$$

Thus by linearity and multiplicativity, the whole the diagram commutes. \square

As a consequence of the above result, we can now formulate:

PROPOSITION 16.30. *We have a quotient map and an inclusion as follows,*

$$U \rightarrow S_U \subset S$$

with S_U being the first row space of U , given by

$$C(S_U) = \langle u_{1i} \rangle \subset C(U)$$

at the level of the corresponding algebras of functions.

PROOF. At the algebra level, we have an inclusion and a quotient map as follows:

$$C(S) \rightarrow C(S_U) \subset C(U)$$

Thus, we obtain the result, by transposing. \square

The above result is all that we need, for getting started with our study, and we will prove in what follows that the inclusion $S_U \subset S$ constructed above is an isomorphism. This will produce the correspondence $U \rightarrow S$ that we are currently looking for.

In order to do so, we will use the uniform integration over S , which can be introduced, in analogy with what happens in the classical case, in the following way:

DEFINITION 16.31. *We endow each of the algebras $C(S)$ with its integration functional*

$$\int_S : C(S) \rightarrow C(U) \rightarrow \mathbb{C}$$

obtained by composing the morphism $x_i \rightarrow u_{1i}$ with the Haar integration of $C(U)$.

With this done, in order now to efficiently integrate over our various spheres S , and in the lack of some trick like spherical coordinates, we need to know how to efficiently integrate over the corresponding quantum isometry groups U .

Now regarding this latter question, we already have some good experience with it, in the classical case, from chapter 14. And the point is that the material there extends without much difficulties in the free case, and we have the following result:

THEOREM 16.32. *Assuming that a compact quantum group $G \subset U_N^+$ is easy, coming from a category of partitions $D \subset P$, we have the Weingarten formula*

$$\int_G u_{i_1 j_1}^{e_1} \dots u_{i_k j_k}^{e_k} = \sum_{\pi, \sigma \in D(k)} \delta_\pi(i) \delta_\sigma(j) W_{kN}(\pi, \sigma)$$

*for any indices $i_r, j_r \in \{1, \dots, N\}$ and any exponents $e_r \in \{\emptyset, *\}$, where δ are the usual Kronecker type symbols, and where*

$$W_{kN} = G_{kN}^{-1}$$

is the inverse of the matrix $G_{kN}(\pi, \sigma) = N^{|\pi \vee \sigma|}$.

PROOF. Let us arrange indeed all the integrals to be computed, at a fixed value of the exponent $k = (e_1 \dots e_k)$, into a single matrix, of size $N^k \times N^k$, as follows:

$$P_{i_1 \dots i_k, j_1 \dots j_k} = \int_G u_{i_1 j_1}^{e_1} \dots u_{i_k j_k}^{e_k}$$

According to the construction of the Haar measure of Woronowicz, explained in the above, this matrix P is the orthogonal projection onto the following space:

$$Fix(u^{\otimes k}) = span \left(\xi_\pi \middle| \pi \in D(k) \right)$$

In order to compute this projection, consider the following linear map:

$$E(x) = \sum_{\pi \in D(k)} \langle x, \xi_\pi \rangle \xi_\pi$$

Consider as well the inverse W of the restriction of E to the following space:

$$span \left(T_\pi \middle| \pi \in D(k) \right)$$

By a standard linear algebra computation, it follows that we have:

$$P = WE$$

But the restriction of E is the linear map corresponding to G_{kN} , so W is the linear map corresponding to W_{kN} , and this gives the result. \square

With this in hand, we can now integrate over the spheres S , as follows:

THEOREM 16.33. *The integration over the basic spheres is given by*

$$\int_S x_{i_1}^{e_1} \dots x_{i_k}^{e_k} = \sum_{\pi} \sum_{\sigma \leq \ker i} W_{kN}(\pi, \sigma)$$

with $\pi, \sigma \in D(k)$, where $W_{kN} = G_{kN}^{-1}$ is the inverse of $G_{kN}(\pi, \sigma) = N^{|\pi \vee \sigma|}$.

PROOF. According to our conventions, the integration over S is a particular case of the integration over U , via $x_i = u_{1i}$. By using now Theorem 16.32, we obtain:

$$\begin{aligned} \int_S x_{i_1}^{e_1} \dots x_{i_k}^{e_k} &= \int_U u_{1i_1}^{e_1} \dots u_{1i_k}^{e_k} \\ &= \sum_{\pi, \sigma \in D(k)} \delta_{\pi}(1) \delta_{\sigma}(i) W_{kN}(\pi, \sigma) \\ &= \sum_{\pi, \sigma \in D(k)} \delta_{\sigma}(i) W_{kN}(\pi, \sigma) \end{aligned}$$

Thus, we are led to the formula in the statement. \square

Again with some inspiration from the classical case, we have the following key result:

THEOREM 16.34. *The integration functional of S has the ergodicity property*

$$\left(id \otimes \int_U \right) \Phi(x) = \int_S x$$

where $\Phi : C(S) \rightarrow C(S) \otimes C(U)$ is the universal affine coaction map.

PROOF. In the real case, $x_i = x_i^*$, it is enough to check the equality in the statement on an arbitrary product of coordinates, $x_{i_1} \dots x_{i_k}$. The left term is as follows:

$$\begin{aligned} \left(id \otimes \int_U \right) \Phi(x_{i_1} \dots x_{i_k}) &= \sum_{j_1 \dots j_k} x_{j_1} \dots x_{j_k} \int_U u_{j_1 i_1} \dots u_{j_k i_k} \\ &= \sum_{j_1 \dots j_k} \sum_{\pi, \sigma \in D(k)} \delta_{\pi}(j) \delta_{\sigma}(i) W_{kN}(\pi, \sigma) x_{j_1} \dots x_{j_k} \\ &= \sum_{\pi, \sigma \in D(k)} \delta_{\sigma}(i) W_{kN}(\pi, \sigma) \sum_{j_1 \dots j_k} \delta_{\pi}(j) x_{j_1} \dots x_{j_k} \end{aligned}$$

Let us look now at the last sum on the right. The situation is as follows:

– In the free case we have to sum quantities of type $x_{j_1} \dots x_{j_k}$, over all choices of multi-indices $j = (j_1, \dots, j_k)$ which fit into our given noncrossing pairing π , and just by using the condition $\sum_i x_i^2 = 1$, we conclude that the sum is 1.

– The same happens in the classical case. Indeed, our pairing π can now be crossing, but we can use the commutation relations $x_i x_j = x_j x_i$, and the sum is again 1.

Thus the sum on the right is 1, in all cases, and we obtain:

$$\left(id \otimes \int_U \right) \Phi(x_{i_1} \dots x_{i_k}) = \sum_{\pi, \sigma \in D(k)} \delta_\sigma(i) W_{kN}(\pi, \sigma)$$

On the other hand, another application of the Weingarten formula gives:

$$\begin{aligned} \int_S x_{i_1} \dots x_{i_k} &= \int_U u_{1i_1} \dots u_{1i_k} \\ &= \sum_{\pi, \sigma \in D(k)} \delta_\pi(1) \delta_\sigma(i) W_{kN}(\pi, \sigma) \\ &= \sum_{\pi, \sigma \in D(k)} \delta_\sigma(i) W_{kN}(\pi, \sigma) \end{aligned}$$

Thus, we are done with the proof of the result, in the real case. In the complex case the proof is similar, by adding exponents everywhere. \square

We can now deduce a useful characterization of the integration, as follows:

THEOREM 16.35. *There is a unique positive unital trace $tr : C(S) \rightarrow \mathbb{C}$ satisfying*

$$(tr \otimes id)\Phi(x) = tr(x)1$$

where Φ is the coaction map of the corresponding quantum isometry group,

$$\Phi : C(S) \rightarrow C(S) \otimes C(U)$$

and this is the canonical integration, as constructed in Definition 16.31.

PROOF. First of all, it follows from the Haar integral invariance condition for U that the canonical integration has indeed the invariance property in the statement, namely:

$$(tr \otimes id)\Phi(x) = tr(x)1$$

In order to prove now the uniqueness, let tr be as in the statement. We have:

$$\begin{aligned} tr \left(id \otimes \int_U \right) \Phi(x) &= \int_U (tr \otimes id)\Phi(x) \\ &= \int_U (tr(x)1) \\ &= tr(x) \end{aligned}$$

On the other hand, according to Theorem 16.34, we have as well:

$$\mathrm{tr} \left(\mathrm{id} \otimes \int_U \right) \Phi(x) = \mathrm{tr} \left(\int_S x \right) = \int_S x$$

We therefore conclude that tr equals the standard integration, as claimed. \square

Getting back now to our axiomatization questions, we have:

THEOREM 16.36. *The operation $S \rightarrow S_U$ produces a correspondence as follows,*

$$\begin{array}{ccc} S_{\mathbb{R},+}^{N-1} & \longrightarrow & S_{\mathbb{C},+}^{N-1} \\ \uparrow & & \uparrow \\ S_{\mathbb{R}}^{N-1} & \longrightarrow & S_{\mathbb{C}}^{N-1} \end{array} \quad \longrightarrow \quad \begin{array}{ccc} O_N^+ & \longrightarrow & U_N^+ \\ \uparrow & & \uparrow \\ O_N & \longrightarrow & U_N \end{array}$$

between basic unitary groups and the basic noncommutative spheres.

PROOF. We use the ergodicity formula from Theorem 16.35, namely:

$$\left(\mathrm{id} \otimes \int_U \right) \Phi = \int_S$$

We know that \int_U is faithful on $\mathcal{C}(U)$, and that we have:

$$(\mathrm{id} \otimes \varepsilon)\Phi = \mathrm{id}$$

The coaction map Φ follows to be faithful as well. Thus for any $x \in \mathcal{C}(S)$ we have:

$$\int_S x x^* = 0 \implies x = 0$$

Thus \int_S is faithful on $\mathcal{C}(S)$. But this shows that we have:

$$S = S_U$$

Thus, we are led to the conclusion in the statement. \square

Very nice all this, all the above is quite encouraging, we definitely have a beginning of free geometry theory here. It is possible to go much further, along these lines, and certain mathematicians have done indeed so. However, before explaining all this, we should perhaps doublecheck with the cats that all this is indeed relevant to quantum physics.

Unfortunately the cats are gone, so I will leave it to you, reader, clarifying such questions. And in case you believe it is so, do not hesitate to have a look at the free geometry, analysis, probability and so on literature, that might help you, in your quest.

16e. Exercises

Congratulations for having read this book, and no exercises for this final chapter. But, of course, with the hope to hear from you sometimes soon, with some heavy integration theorems of your own. The one who knows how to integrate wins.

Bibliography

- [1] V.I. Arnold, Ordinary differential equations, Springer (1973).
- [2] V.I. Arnold, Mathematical methods of classical mechanics, Springer (1974).
- [3] V.I. Arnold, Catastrophe theory, Springer (1974).
- [4] V.I. Arnold, Lectures on partial differential equations, Springer (1997).
- [5] V.I. Arnold and B.A. Khesin, Topological methods in hydrodynamics, Springer (1998).
- [6] M.F. Atiyah, K-theory, CRC Press (1964).
- [7] M.F. Atiyah, The geometry and physics of knots, Cambridge Univ. Press (1990).
- [8] T. Banica, Calculus and applications (2024).
- [9] T. Banica, Principles of operator algebras (2024).
- [10] T. Banica, Introduction to modern physics (2025).
- [11] R.J. Baxter, Exactly solved models in statistical mechanics, Academic Press (1982).
- [12] I. Bengtsson and K. Życzkowski, Geometry of quantum states, Cambridge Univ. Press (2006).
- [13] N. Berline, E. Getzler and M. Vergne, Heat kernels and Dirac operators, Springer (2004).
- [14] S.J. Blundell and K.M. Blundell, Concepts in thermal physics, Oxford Univ. Press (2006).
- [15] S.M. Carroll, Spacetime and geometry, Cambridge Univ. Press (2004).
- [16] A.R. Choudhuri, Astrophysics for physicists, Cambridge Univ. Press (2012).
- [17] A. Connes, Noncommutative geometry, Academic Press (1994).
- [18] J.B. Conway, A course in functional analysis, Springer (1985).
- [19] W.N. Cottingham and D.A. Greenwood, An introduction to the standard model of particle physics, Cambridge Univ. Press (2012).
- [20] P.A. Davidson, Introduction to magnetohydrodynamics, Cambridge Univ. Press (2001).
- [21] P.A.M. Dirac, Principles of quantum mechanics, Oxford Univ. Press (1930).
- [22] M.P. do Carmo, Differential geometry of curves and surfaces, Dover (1976).
- [23] M.P. do Carmo, Riemannian geometry, Birkhäuser (1992).
- [24] S. Dodelson, Modern cosmology, Academic Press (2003).
- [25] R. Durrett, Probability: theory and examples, Cambridge Univ. Press (1990).
- [26] A. Einstein, Relativity: the special and the general theory, Dover (1916).

- [27] L.C. Evans, Partial differential equations, AMS (1998).
- [28] W. Feller, An introduction to probability theory and its applications, Wiley (1950).
- [29] E. Fermi, Thermodynamics, Dover (1937).
- [30] R.P. Feynman, R.B. Leighton and M. Sands, The Feynman lectures on physics I: mainly mechanics, radiation and heat, Caltech (1963).
- [31] R.P. Feynman, R.B. Leighton and M. Sands, The Feynman lectures on physics II: mainly electromagnetism and matter, Caltech (1964).
- [32] R.P. Feynman, R.B. Leighton and M. Sands, The Feynman lectures on physics III: quantum mechanics, Caltech (1966).
- [33] R.P. Feynman and A.R. Hibbs, Quantum mechanics and path integrals, Dover (1965).
- [34] P. Flajolet and R. Sedgewick, Analytic combinatorics, Cambridge Univ. Press (2009).
- [35] A.P. French, Special relativity, Taylor and Francis (1968).
- [36] W. Fulton, Algebraic topology, Springer (1995).
- [37] W. Fulton and J. Harris, Representation theory, Springer (1991).
- [38] C. Godsil and G. Royle, Algebraic graph theory, Springer (2001).
- [39] H. Goldstein, C. Saffo and J. Poole, Classical mechanics, Addison-Wesley (1980).
- [40] M.B. Green, J.H. Schwarz and E. Witten, Superstring theory, Cambridge Univ. Press (2012).
- [41] D.J. Griffiths, Introduction to electrodynamics, Cambridge Univ. Press (2017).
- [42] D.J. Griffiths and D.F. Schroeter, Introduction to quantum mechanics, Cambridge Univ. Press (2018).
- [43] D.J. Griffiths, Introduction to elementary particles, Wiley (2020).
- [44] G.H. Hardy and E.M. Wright, An introduction to the theory of numbers, Oxford Univ. Press (1938).
- [45] J. Harris, Algebraic geometry, Springer (1992).
- [46] R. Hartshorne, Algebraic geometry, Springer (1977).
- [47] A. Hatcher, Algebraic topology, Cambridge Univ. Press (2002).
- [48] H. Hofer and E. Zehnder, Symplectic invariants and Hamiltonian dynamics, Birkhäuser (1994).
- [49] L. Hörmander, The analysis of linear partial differential operators, Springer (1983).
- [50] R.A. Horn and C.R. Johnson, Matrix analysis, Cambridge Univ. Press (1985).
- [51] K. Huang, Introduction to statistical physics, CRC Press (2001).
- [52] K. Huang, Quantum field theory, Wiley (1998).
- [53] K. Huang, Quarks, leptons and gauge fields, World Scientific (1982).
- [54] J.E. Humphreys, Introduction to Lie algebras and representation theory, Springer (1972).
- [55] K. Ireland and M. Rosen, A classical introduction to modern number theory, Springer (1982).

- [56] V.F.R. Jones, Subfactors and knots, AMS (1991).
- [57] L.P. Kadanoff, Statistical physics: statics, dynamics and renormalization, World Scientific (2000).
- [58] T. Kibble and F.H. Berkshire, Classical mechanics, Imperial College Press (1966).
- [59] M. Kumar, Quantum: Einstein, Bohr, and the great debate about the nature of reality, Norton (2009).
- [60] T. Lancaster and K.M. Blundell, Quantum field theory for the gifted amateur, Oxford Univ. Press (2014).
- [61] L.D. Landau and E.M. Lifshitz, Mechanics, Pergamon Press (1960).
- [62] L.D. Landau and E.M. Lifshitz, The classical theory of fields, Addison-Wesley (1951).
- [63] L.D. Landau and E.M. Lifshitz, Quantum mechanics: non-relativistic theory, Pergamon Press (1959).
- [64] S. Lang, Algebra, Addison-Wesley (1993).
- [65] P. Lax, Linear algebra and its applications, Wiley (2007).
- [66] P. Lax, Functional analysis, Wiley (2002).
- [67] P. Lax and M.S. Terrell, Calculus with applications, Springer (2013).
- [68] P. Lax and M.S. Terrell, Multivariable calculus with applications, Springer (2018).
- [69] D. McDuff and D. Salamon, Introduction to symplectic topology, Oxford Univ. Press (2017).
- [70] M.L. Mehta, Random matrices, Elsevier (2004).
- [71] M.A. Nielsen and I.L. Chuang, Quantum computation and quantum information, Cambridge Univ. Press (2000).
- [72] R.K. Pathria and P.D. Beale, Statistical mechanics, Elsevier (1972).
- [73] A. Peres, Quantum theory: concepts and methods, Kluwer (1993).
- [74] P. Petersen, Linear algebra, Springer (2012).
- [75] W. Rudin, Principles of mathematical analysis, McGraw-Hill (1964).
- [76] W. Rudin, Real and complex analysis, McGraw-Hill (1966).
- [77] W. Rudin, Functional analysis, McGraw-Hill (1973).
- [78] W. Rudin, Fourier analysis on groups, Dover (1974).
- [79] B. Ryden, Introduction to cosmology, Cambridge Univ. Press (2002).
- [80] D.V. Schroeder, An introduction to thermal physics, Oxford Univ. Press (1999).
- [81] J. Schwinger and B.H. Englert, Quantum mechanics: symbolism of atomic measurements, Springer (2001).
- [82] J.P. Serre, A course in arithmetic, Springer (1973).
- [83] J.P. Serre, Linear representations of finite groups, Springer (1977).
- [84] I.R. Shafarevich, Basic algebraic geometry, Springer (1974).

- [85] R. Shankar, Fundamentals of physics, Yale Univ. Press (2014).
- [86] A.M. Steane, Thermodynamics, Oxford Univ. Press (2016).
- [87] S. Sternberg, Dynamical systems, Dover (2010).
- [88] J.R. Taylor, Classical mechanics, Univ. Science Books (2003).
- [89] D.V. Voiculescu, K.J. Dykema and A. Nica, Free random variables, AMS (1992).
- [90] J. von Neumann, Mathematical foundations of quantum mechanics, Princeton Univ. Press (1955).
- [91] J. von Neumann and O. Morgenstern, Theory of games and economic behavior, Princeton Univ. Press (1944).
- [92] J. Watrous, The theory of quantum information, Cambridge Univ. Press (2018).
- [93] S. Weinberg, Foundations of modern physics, Cambridge Univ. Press (2011).
- [94] S. Weinberg, Lectures on quantum mechanics, Cambridge Univ. Press (2012).
- [95] S. Weinberg, Lectures on astrophysics, Cambridge Univ. Press (2019).
- [96] S. Weinberg, Cosmology, Oxford Univ. Press (2008).
- [97] H. Weyl, The theory of groups and quantum mechanics, Princeton Univ. Press (1931).
- [98] H. Weyl, The classical groups: their invariants and representations, Princeton Univ. Press (1939).
- [99] H. Weyl, Space, time, matter, Princeton Univ. Press (1918).
- [100] B. Zwiebach, A first course in string theory, Cambridge Univ. Press (2004).

Index

- abelian group, 57
- absolutely continuous, 174
- abstract algebra, 345
- ADE graph, 319
- adjacency matrix, 299
- adjoint operator, 227, 243, 244, 281, 295
- algebra character, 347
- algebraic curve, 177
- alternating group, 59
- analytic function, 255
- arcsine law, 310
- area of sphere, 158
- arithmetic group, 59
- asymptotic characters, 339
- asymptotic independence, 173
- asymptotic law, 173
- average of function, 27

- Banach algebra, 345
- Bell numbers, 53
- Bernoulli laws, 51
- Bessel law, 72, 73, 340
- bidual, 220
- Borel set, 119
- boundary of set, 111
- bounded operator, 242
- Brauer theorem, 335

- Cardano formula, 22
- Catalan numbers, 302, 304, 356
- category of partitions, 332
- Cauchy formula, 255, 285
- Cauchy transform, 307
- Cauchy-Schwarz, 205, 232
- CCLT, 98
- celestial mechanics, 178

- central binomial coefficients, 301
- Central Limit Theorem, 172
- Central Limit theorem, 81, 91
- Cesàro limit, 327
- change of variables, 160
- character, 61, 321
- chart, 182
- Chebycheff polynomials, 240
- circular measure, 316, 319
- classical cumulant, 77
- classical mechanics, 178
- closed and bounded, 115
- closed subalgebra, 346
- closure of set, 111
- CLT, 81, 91, 172
- coefficients of representations, 325
- colored integers, 98
- colored moments, 98, 349
- colored partitions, 73
- colored powers, 323
- commutative algebra, 347
- commuting normal operators, 295
- commuting self-adjoint operators, 293
- compact group, 321
- compact set, 115
- compact space, 347
- compact support, 250
- complex Bessel law, 340
- complex Bessel laws, 69
- Complex Central Limit Theorem, 98
- Complex CLT, 98
- complex conjugate, 256
- complex Gaussian law, 97
- complex hyperspherical laws, 174
- complex normal law, 97, 98

- complex number, 20
- complex reflection group, 60, 73, 336
- complex sphere, 174
- complex variables, 98
- compound Poisson law, 68
- Compound Poisson Limit Theorem, 69
- concave function, 204
- conic, 178
- conjugate representation, 322
- connected set, 117
- conservative force, 187
- continuous calculus, 289, 295, 348
- continuous function, 117
- convex function, 204
- convolution, 45, 170, 249, 251
- convolution exponential, 50
- convolution semigroup, 49, 92, 97, 171
- coordinate axes, 60
- coordinates, 182
- CPLT, 69
- crossed product, 71
- cubic root, 22
- cumulant, 74, 77
- cumulant-generating function, 74
- cyclic group, 58

- Dedekind cut, 15
- degree 2 curve, 177
- degree 3 equation, 22
- degree 4 equation, 22
- degree 5 polynomial, 25
- derangements, 62
- diagonalization, 292–295
- differential equation, 178
- differential manifold, 182
- dihedral group, 59
- Dirac mass, 38
- discrete convolution, 45
- discrete integration, 38
- discrete law, 38
- discrete measure, 38
- discrete probability, 38
- discriminant, 22
- distance, 108
- distribution, 252, 349
- double factorial, 88, 156
- double factorials, 156, 166

- dual Banach space, 217
- Dyck paths, 302

- easy group, 335, 341, 343
- eigenvalue, 275, 276
- eigenvector, 275
- End space, 323
- equal almost everywhere, 214
- equation of motion, 178
- exchange of hat, 261

- field, 12
- finite abelian group, 58
- finite field, 13
- finite group, 59
- finite non-abelian group, 59
- Fix space, 323
- fixed points, 62
- Fourier inversion, 265
- Fourier transform, 47, 50, 68, 170, 171, 259
- friction, 187
- Frobenius isomorphism, 329
- full reflection group, 60
- function space, 214
- functional calculus, 346

- Gauss integral, 163
- Gaussian matrix, 351, 353
- Gaussian variable, 81, 91
- Gelfand theorem, 347
- general linear, 59
- generalized Bessel laws, 69
- generalized cumulant, 77
- Gram determinant, 339
- Gram matrix, 337, 338, 340, 341
- Gram-Schmidt, 237
- gravity, 178
- group, 57
- group representation, 61

- Hölder inequality, 205, 206, 213
- Haar integration, 327
- Haar measure, 327
- Hankel determinant, 308
- harmonic function, 256, 257
- Hermite polynomials, 241
- Hilbert space, 232, 242
- holomorphic calculus, 285, 295

- holomorphic function, 255
- Hom space, 323
- Hooke law, 266
- hypercube, 60, 71
- hyperoctahedral group, 60, 71, 336
- hyperspherical law, 173
- independence, 45, 47, 169, 170
- infinite matrix, 242
- infinitely differentiable, 255
- interior of set, 111
- intertwiners, 323
- invertible operator, 275, 276
- involution of algebra, 243
- isometry, 245, 246
- Jacobi polynomials, 240
- Jacobian, 160
- Jensen inequality, 204
- joint moments, 102
- Kepler laws, 178
- Kronecker symbols, 333
- Laguerre polynomials, 241
- Laplace operator, 256, 257, 266, 271
- latitude, 163
- lattice, 76
- lattice model, 266
- law, 349
- Legendre polynomials, 240
- Lie group, 59
- linear operator, 242
- Liouville theorem, 256, 257
- longitude, 163
- Möbius function, 76, 338
- Möbius inversion, 76, 338
- main value formula, 255
- maps associated to partitions, 333
- Marchenko-Pastur law, 309, 361
- matching pairings, 101, 335
- matrix determinant, 160
- maximum modulus, 257
- maximum principle, 255
- mean value formula, 257
- measurable calculus, 291, 295
- measurable set, 118
- metric space, 108
- middle binomial coefficients, 311
- Minkowski inequality, 207, 213, 232
- modified arcsine law, 311
- moment-cumulant formula, 77
- moments, 45, 349
- moments of characters, 339
- Monte Carlo integration, 27
- N-gon, 59
- Netwon law, 178
- Newton law, 266
- Newton theorem, 178
- non-abelian group, 59
- norm closed, 345
- norm of operator, 242
- normal element, 350
- normal law, 30, 340
- normal operator, 230, 274, 288, 289, 291, 294, 346
- normal variable, 81, 91
- normed algebra, 345
- normed space, 208, 212, 214
- number of blocks, 65
- operator algebra, 345
- order of element, 58
- order on partitions, 337
- oriented cycle, 60
- orthogonal group, 59, 335
- orthogonal polynomials, 238
- orthonormal basis, 237
- p-norm, 208, 212, 214
- partial derivatives, 158
- partial isometry, 247
- permutation group, 59
- permutations, 59
- Peter-Weyl, 324, 325, 329, 330
- Peter-Weyl representations, 322, 331
- PLT, 51
- Poisson law, 31, 47, 62
- Poisson Limit Theorem, 51
- polar coordinates, 162, 178
- polynomial calculus, 279, 289, 295
- polynomial integrals, 65
- positive element, 348
- potential, 189
- probability paradox, 30

- probability zero, 30
- product of representations, 322
- projection, 227, 246
- quantum isometry group, 384
- random matrix algebra, 345
- random number, 27
- random permutations, 62
- random variable, 349
- random walk, 299
- Randon-Nikodym, 175
- rational calculus, 281, 295
- rational number, 12
- real Bessel law, 340
- real number, 15
- reflection group, 60, 73, 336
- reflexivity, 220
- regularization, 251
- remainder modulo N , 58
- representation, 61, 321
- Riemann integration, 26
- Riemann-Lebesgue property, 264
- Rodrigues formula, 240
- root of unity, 20
- roots, 25
- roots of unity, 58
- rotation group, 174
- scalar product, 227, 232
- self-adjoint operator, 273, 282, 292, 346
- semicircle law, 309
- separable space, 237
- sequence of cumulants, 74
- sequence of moments, 74
- shift, 245
- smooth manifold, 182
- smooth representation, 325
- special linear, 59
- special orthogonal group, 59
- special unitary group, 59
- spectral measure, 350
- spectral radius, 286, 346
- spectrum, 275, 277, 346
- spectrum of algebra, 347
- spectrum of products, 278
- spherical coordinates, 163, 164, 271
- spherical integral, 166, 168
- square root, 14
- square-summable, 233
- star operation, 243
- Stieltjes inversion, 307
- sum of representations, 322
- support, 250
- symmetric group, 59, 335
- symmetry group, 59
- symplectic group, 59
- Tannakian category, 331
- Tannakian duality, 332, 334
- Taylor coefficient, 74
- tensor category, 323, 331
- trigonometric integral, 156, 165
- truncated character, 65, 72, 343
- truncated characters, 342
- uniform group, 342, 343
- unitary, 227, 229, 246, 282
- unitary group, 59, 174, 335
- unitary operator, 346
- unitary representation, 61
- variance, 91
- volume inflation, 160
- volume of sphere, 155–157
- Wallis formula, 156, 165
- wave equation, 266
- Weierstrass basis, 238
- Weingarten formula, 340, 341
- Weingarten matrix, 340
- Wick formula, 102, 103
- Wigner law, 309
- Wigner matrix, 351
- Wishart matrix, 352, 356, 358, 361
- work, 187
- wreath product, 60, 71